IMPLEMENTATION OF DTS AUDIO IN MEDIA FILES BASED ON ISO/IEC 14496

Document No.: 9302J81100
Revision: G
Version: 1.4
Effective Date: February 2014

DTS, Inc.
5220 Las Virgenes Road
Calabasas, CA 91302
USA

www.dts.com

CONFIDENTIAL

No Warranty

THE FOLLOWING TERMS SHALL APPLY TO ANY USE OF ANY HARDWARE, SOFTWARE AND METHODS ASSOCIATED WITH THIS DOCUMENT AND, AS APPLICABLE, ANY RELATED DOCUMENTATION (THE "PRODUCT"). THE PRODUCT IS PROVIDED "AS IS" AND WITHOUT WARRANTY OF ANY KIND. DTS EXPRESSLY DISCLAIMS ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. DTS DOES NOT WARRANT THAT THE PRODUCT WILL MEET USER'S REQUIREMENTS, OR THAT DEFECTS IN THE PRODUCT WILL BE CORRECTED. DTS DOES NOT WARRANT THAT THE OPERATION OF ANY HARDWARE OR SOFTWARE ASSOCIATED WITH THIS DOCUMENT WILL BE UNINTERRUPTED OR ERROR-FREE. UNDER NO CIRCUMSTANCES SHALL DTS OR THE DIRECTORS, OFFICERS, EMPLOYEES, OR AGENTS OF DTS, BE LIABLE TO ANY PARTY FOR INCIDENTAL, INDIRECT, SPECIAL, OR CONSEQUENTIAL DAMAGES ARISING OUT OF THE USE, MISUSE, OR INABILITY TO USE THE PRODUCT.

COPYRIGHT

Contents

1 Introduction 6
1.1 Document conventions 6
1.2 References 6
1.3 Terms, Definitions, and Acronyms 6
2 DTS Elementary Streams 8
2.1 Introduction 8
2.2 Design Rules 8
2.2.1 Track Header Box 8
2.2.2 Sync Sample Box 8
2.2.3 Handler Reference Box 8
2.2.4 Sound Media Header Box 9
2.2.5 Sample Description Box 9
2.2.6 AudioSampleEntry Box for DTS Formats 9
2.2.7 DTS_SampleEntry 9
2.2.8 DTSSpecificBox 10
2.2.9 Semantics 10
2.2.10 Reserved Boxes 13
2.3 Storage of DTS elementary streams 15
2.4 Restrictions on DTS Formats 15
Annex A Implementation Guidelines (informative) 16
Annex B Alternative Carrier Codec Guidelines for DTS Neural Surround (informative) 17
B.1 Introduction 17
B.2 Signaling 17
B.2.1 Handler Reference Box 17

Tables

Table 2-1 - Defined Audio Formats 9
Table 2-2 – StreamConstruction 11
Table 2-3 - Core Layout 12
Table 2-4 - Representation Type 12
Table 2-5 – Channel Layout 13
Table 2-6 - ChannelLayoutExt 15

1 Introduction

1.1 Document conventions

Certain stylistic conventions have been adopted for use throughout this document as a matter of convenience and readability. Some of those conventions are described here.

References are indicated by a bracketed number, e.g. [1].

Hexadecimal numerical values are indicated as '*n*h'.

Binary numerical values are indicated as '*n*b'.

Character strings as parameters are enclosed in a single quotation, such as 'abcd'.

Variable names or parameter names which are defined as such in this or referenced documents, when used in the context of a paragraph, are highlighted in bold case such as **FSIZE**, or may be distinguished by a quote or double quote and italics, depending on context.

All bit fields, indicated using the bit(x) notation, applying enumerated sequence as parameter values are most significant bit first. If any of the possible enumerated values are not defined in this document, they shall be considered reserved for future use.

Parameters, data fields, or enumerated values marked or implied as 'reserved' may be defined in future or other specifications. A device built applying only this specification will not read such parameters, data fields, or enumerated values. . Any data structures that are created using only this specification shall set reserved fields to 0.

1.2 References

[1] ISO/IEC 14496-12, "Information technology - Coding of audio-visual objects - Part 12: ISO Base Media File Format"

[2] ISO/IEC 14496-14:2003, "Information technology - Coding of audio-visual objects - Part 14: MP4 file format"

[3] ETSI TS 102 114 v1.4.1 (2012-09), "DTS Coherent Acoustics; Core and Extensions with Additional Profiles"

[4] "DTS-HD PBR API Library Description", (*DTS Document No.: 9302J19200)*

1.3 Conformance

As used in this document, "shall" denotes mandatory provisions of the standard. "Should" denotes a provision that is recommended but not mandatory. "May" denotes a feature whose presence does not preclude compliance and implementation of which is optional. "Optional" denotes items that may or may not be present in a compliant device.

1.4 Terms, Definitions, and Acronyms

The following terms, definitions and abbreviations shall apply to all clauses in this document.

AU – access unit

audio stream – A sequence of synchronized audio frames.

audio frame – A component of an audio stream that corresponds to a certain number of PCM audio samples. Usually also an AU.

base layer – In DTS layered audio, the base layer is the foundation of the layered hierarchy. The base layer may be a fully compliant and decodable DTS bitstream.

CBR - Constant Bit Rate.

core substream: A DTS audio stream, or a component of a DTS audio stream, conforming to [3] and always begins with the 32-bit Sync word of 0x7FFE8001.

dependent base layer – A base layer that is not a decodable DTS bitstream. At least one enhancement layer is required to reconstruct a DTS bitstream using the dependent base layer.

duration – The time represented by one decoded audio frame, may be represented in audio samples per channel at a specific audio sampling frequency or in seconds.

enhancement layer – An audio track used in DTS layered audio that contains binary information used to modify another DTS audio track. Multiple enhancement layers are common to make successive improvements in a hierarchical manner.

execution layer – An enhancement layer whose commands are used in the reconstruction hierarchy for a given sample, normally the highest available layer in the layer hierarchy. All enhancement layers using the 'dts+' codec name are capable of serving as execution layers.

extension – For DTS bitstreams, a component of an audio frame, may or may not exist in sequence with other extension components or a core component.

extension substream: A DTS audio stream, or a component of a DTS audio stream, conforming to [3] and always begins with the 32-bit sync word of 0x64582025.

LFE: Low Frequency Effects or subwoofer channel.

VBR - Variable Bit Rate.

XLL – The DTS-HD® lossless audio coding extension, a logical element within the DTS elementary stream containing compressed audio data that will decode into bit exact representation of the original signal.

2 DTS Elementary Streams

2.1 Introduction

This chapter describes the DTS audio track in relation to the ISO file and the constraints on the DTS audio formats.

In general, the system layer definition described in 14496-1 is used to embed the audio. All DTS audio formats comply with the conventions described herein.

2.2 Design Rules

In this section, operational rules for boxes defined in ISO Base Media File Format [1] and MP4 File Format [2] as well as definitions of private extensions to those ISO file format standards are described.

An ISO media file may contain one or more audio tracks. The tracks are composed in conformity to ISO base media file format and described in [1], for an audio track structure. The sub-sections that follow describe the construction of an audio track containing DTS audio.

2.2.1 Track Header Box

The syntax and values for the Track Box and its sub-boxes shall conform to section 8.5 of ISO base media file format [1], and the following fields of each box shall be set to the following specified values. There are some “template” fields declared to use; see [1].

flags = 000007h, except for the case where the track belongs to an alternate group;
layer = 0;
volume = 0100h;
matrix = {00010000h,0,0,0, 00010000h,0,0,0, 40000000h};
width = 0;
height = 0;

2.2.1.1 Track Reference Box

For DTS enhancement layers, (codingname = ‘dts+’ in Table 2-1 below), use of the track reference box (‘tref’) is required. The reference_type in ‘tref’ shall be ‘lyra’, indicating layered audio. The contents of ‘tref’ shall otherwise comply with [1].

2.2.2 Sync Sample Box

As all DTS audio access units are random access points (sync samples), the Sync Sample Box shall not be present in the track time structure of any DTS audio track within an ISO media file.

2.2.3 Handler Reference Box

The syntax and values for the Handler Reference Box shall conform to section 8.4.3 of [1], and the fields of this box shall be set to the following specified values.

handler_type = ‘soun’

2.2.4 Sound Media Header Box

The syntax and values for the Sound Media Header box shall conform to section 8.4.53 of [1], and the fields of this box shall be set to the following specified values.

balance = 0;

2.2.5 Sample Description Box

For the DTS audio formats supported by the ISO media file format, a specific SampleEntry box that is derived from the AudioSampleEntry box defined in [1] is used. The dts-specific SampleEntry box is identified by a unique codingname value (see Table 2-1). It specifies the audio format used to encode the audio track and describes the configuration of the audio elementary stream.

Table 2-1 - Defined Audio Formats

codingname	Description
dtsc	DTS formats prior to DTS-HD
dtsh	DTS-HD audio formats
dtsl	DTS-HD Lossless formats
dtse	DTS Low Bit Rate (LBR) formats
dts-	DTS dependent base layer
dts+	DTS enhancement layer

2.2.6 AudioSampleEntry Box for DTS Formats

The syntax and values of the AudioSampleEntry Box shall conform to DTSSampleEntry.

The configuration of the DTS elementary stream is described in the DTSSpecificBox (‘ddts’), within DTSSampleEntry. The syntax and semantics of the DTSSpecificBox are defined in 2.2.8.

2.2.7 DTS_SampleEntry

DTSSampleEntry extends the AudioSampleEntry box defined in [1]:

```
class DTSSampleEntry (codingname) extends AudioSampleEntry (codingname) {
    DTSSpecificBox()            // ‘ddts’ box
}
```

For DTSSampleEntry(), the following values inherited from AudioSampleEntry are set as follows:

codingname shall be set to one of the values listed in Table 2-1. For DTS bitstreams compliant with [3], the coding name shall be in accordance with Table 2-2. For bitstream data being used in a subsequent reconstruction process, ‘dts+’ shall be used for enhancement layers and ‘dts-‘ shall be used for dependent base layers.

channelcount shall be set to the number of decodable output channels in basic playback, as described in the ‘ddts’ configuration box.

samplesize shall be set to 16,

samplerate shall be set according to DTSSamplingFrequency (see section 2.2.9) of either:

48000 for original sampling frequencies of 24000Hz, 48000Hz, 96000Hz or 192000Hz;

44100 for original sampling frequencies of 22050Hz, 44100Hz, 88200Hz or 176400Hz;

32000 for original sampling frequencies of 16000Hz, 32000Hz, 64000Hz or 128000Hz.

If codingname is 'dts-' then all entries in DTSSampleEntry() following codingname shall be ignored and only the DTSSampleEntry data of the *execution layer* shall be deemed valid.

If codingname is 'dts+', then all parameters in DTSSampleEntry() shall represent the condition of the sample following reconstruction assuming the present layer is the execution layer, with the exception of codingname, and two parameters found in the DTSSpecificBox, maxBitrate and avgBitrate. See section 2.2.8 and 2.2.10.1 for further details.

2.2.8 DTSSpecificBox

The syntax and semantics of the DTSSpecificBox are shown below.

```
class DTSSpecificBox extends Box ('ddts') {
    unsigned int(32)    DTSSampling Frequency;
    unsigned int(32)    maxBitrate;
    unsigned int(32)    avgBitrate;
    unsigned char       pcmSampleDepth;         // value is 16 or 24 bits
    bit(2)              FrameDuration;          // 0 = 512, 1 = 1024, 2 = 2048, 3 = 4096
    bit(5)              StreamConstruction;     // Table 2-2
    bit(1)              CoreLFEPresent;         // 0 = none; 1 = LFE exists
    bit(6)              CoreLayout;             // Table 2-3
    bit(14)             CoreSize;
    bit(1)              StereoDownmix           // 0 = none; 1 = embedded downmix present
    bit(3)              RepresentationType;     // Table 2-4
    bit(16)             ChannelLayout;          // Table 2-5
    bit(1)              MultiAssetFlag          // 0 = single asset, 1 = multiple asset
    bit(1)              LBRDurationMod          // 0 = ignore, 1 = Special LBR duration modifier
    bit(1)              ReservedBoxPresent      // 0 = no ReservedBox, 1 = Reserved present
    bit(5)              Reserved                // Shall be set to 0
    if ( ReservedBoxPresent) {
        box ReservedBox[];
    }
};
```

2.2.9 Semantics

DTSSamplingFrequency – The maximum sampling frequency stored in the compressed audio stream.

pcmSampleDepth – The bit depth of the rendered audio. For DTS formats this is usually 24-bits.

maxBitrate – The peak bit rate, in bits per second, of the audio elementary stream over the duration of the track. If the stream is a constant bit rate, this parameter shall have the same value as avgBitrate. If the maximum bit rate is unknown, this parameter shall be set to 0.

avgBitrate – The average bit rate, in bits per second, of the audio elementary stream over the duration of the track.

FrameDuration – This code represents the number of audio samples represented in a complete audio access unit at DTSSampling Frequency.

StreamConstructon – Provides complete information on the existence and of location of extensions in any synchronized frame. See Table 2-2. For any stream type not listed in Table 2-2, this parameter SHALL be set to 0 and the coding name SHALL default to 'dtsh'.

Table 2-2 – StreamConstruction

StreamConstruction	Core Substream				Extension Substream						codingname
	Core	XCH	X96	XXCH	Core	XXCH	X96	XBR	XLL	LBR	
1	✓										dtsc
2	✓	✓									dtsc
3	✓			✓							dtsh
4	✓		✓								dtsc
5	✓					✓					dtsh
6	✓							✓			dtsh
7	✓	✓						✓			dtsh
8	✓			✓				✓			dtsh
9	✓					✓		✓			dtsh
10	✓						✓				dtsh
11	✓	✓					✓				dtsh
12	✓			✓			✓				dtsh
13	✓					✓	✓				dtsh
14	✓								✓		dtsh
15	✓	✓							✓		dtsh
16	✓		✓						✓		dtsh
17									✓		dtsl
18										✓	dtse
19					✓						dtsh
20					✓	✓					dtsh
21					✓				✓		dtsh

CoreLFEPresent – Indicates the presence of an LFE channel in the core. If no core substream exists, this value shall be ignored.

CoreLayout – This parameter represents the channel layout of the core within the core substream and shall be set according to Table 2-3. If no core substream exists, this parameter shall be ignored and ChannelLayout or RepresentationType shall be used to determine channel configuration.

Table 2-3 - Core Layout

Core Layout	Description
0	Mono (1/0)
2	Stereo (2/0)
4	LT,RT (2/0)
5	L, C, R (3/0)
6	L, R, S (2/1)
7	L, C, R, S (3/1)
8	L, R, LS, RS (2/2)
9	L, C, R, LS, RS (3/2)
31	use ChannelLayout

For streams where StreamConstruction is undefined (i.e. StreamConstruction = 0), or a DTS core component only exists in the extension substream (e.g. StreamConstruction = 19, 20 or 21 in Table 2-2), CoreLayout SHALL be set to 31.

All undefined values for CoreLayout are reserved for future use.

CoreSize – The size of a core substream AU in bytes. If no core substream exists, CoreSize is 0, and parameters CoreLayout and CoreLFEPresent shall be ignored.

StereoDownmix – Indicates the presence of an embedded stereo downmix in the stream. This parameter is not valid for stereo or mono streams.

RepresentationType – This indicates special properties of the audio presentation, as indicated in Table 2-4. This parameter is only valid when all flags in ChannelLayout are set to 0. If ChannelLayout $\neq 0$, this value shall be ignored.

Table 2-4 - Representation Type

RepresentationType	Description	Number of Channels
0	Audio asset designated for mixing with another audio asset	
2	Lt/Rt Encoded for matrix surround decoding	2
3	Audio processed for headphone playback	2
1 and 4 through 7	Reserved	

ChannelLayout – Provides complete information on channels coded in the audio stream including core and extensions. The binary masks of the channels present, as shown in Table 2-5, are added together to create ChannelLayout.

Table 2-5 – Channel Layout

Bit Masks	Loudspeaker Location Description	Number of Channels
0000h	Consult RepresentationType	see Table 2-4
0001h	Center in front of listener	1
0002h	Left/Right in front	2
0004h	Left/Right surround on side in rear	2
0008h	Low frequency effects subwoofer	1
0010h	Center surround in rear	1
0020h	Left/Right height in front	2
0040h	Left/Right surround in rear	2
0080h	Center Height in front	1
0100h	Over the listener's head	1
0200h	Between left/right and center in front	2
0400h	Left/Right on side in front	2
0800h	Left/Right surround on side	2
1000h	Second low frequency effects subwoofer	1
2000h	Left/Right height on side	2
4000h	Center height in rear	1
8000h	Left/Right height in rear	2

MultiAssetFlag – This flag shall be set if the stream contains more than one asset. This also implies that a DTS extension substream is present. Multiple asset streams shall use the 'dtsh' coding type. When multiple assets exist, the remaining parameters in the DTSSpecificBox only reflect the coding parameters of the first asset.

LBRDurationMod – This flag indicates a special case of the LBR coding bandwidth, resulting in 1/3 or 2/3 band limiting. The result of this is the LBR frame duration is 50% larger than indicated in FrameDuration. For example, when this flag is set to 1, the FrameDuration is 6144 samples instead of 4096 samples.

ReservedBoxPresent – This flag indicates that at least 1 box occupies the ReservedBox placeholder.

2.2.10 Reserved Boxes

ReservedBox serves as a placeholder for expansion. One or more additional private boxes to 'ddts' may overload ReservedBox. If any such boxes are present, ReservedBoxPresent shall be set to 1.

Playback devices not equipped to support these extensions shall be able to depend on the defined parameters in 'ddts' for basic playback capability.

The following boxes are defined for use as reserved boxes in 'ddts'. If ReservedBoxPresent is set to 1, then capable playback clients should continue parsing 'ddts' for expansion boxes until *ddts.size* is accounted for. Any expansion boxes not recognized by a client shall be skipped over.

2.2.10.1 LayeredReconstruction box

DTS enhancement layers may be reunited with the defined base and other enhancement layers creating a new DTS-HD audio stream that complies with [3]. In such cases it is convenient to have certain parameters available to form the correct DTSSampleEntry.

If DTSSampleEntry has a 'dts+' codec designation, this track contains a DTS enhancement layer and a LayerReconstruction box shall be present as one ReservedBox. The LayerReconstruction box is defined as follows:

```
class LayerReconstruction extends FullBox ('lyrb', version = 0, 0) {
    unsigned int(32)    codingname;
    unsigned int(32)    maxBitrate;
    unsigned int(32)    avgBitrate;
}
```

The LayerReconsruction box is only valid for enhancement layers when they are also the execution layer.

2.2.10.1.1 Semantics

codingname – This shall be one of 'dtse', 'dtsc', 'dtsh', or 'dtsl' 4 character codes corresponding to the construction parameters shown in Table 2-2 of the reassembled bitstream.

maxBitrate – This value represents the peak bit rate, in bits per second, of the reconstructed audio elementary stream for the duration of the track. If the stream is a constant bit rate, this parameter shall have the same value as *avgBitrate*. If the maximum bit rate is unknown, this parameter shall be set to 0.

avgBitrate – This value represents the average bit rate, in bits per second, of the reconstructed audio elementary stream for the duration of the track.

2.2.10.2 ExtendedParameters box

The ExtendedParameters box contains additional metadata in an extensible manner.

```
class ExtendedParameters extends FullBox ('epba', version = 0, 0) {
    bit(1)          ChannelLayoutExtFlag;       // 0 = no ChannelLayoutExt, 1 = ChannelLayoutExt present
    bit(7)          ReservedFlags               // shall be set to 0;
    if (ChannelLayoutExtFlag == 1) {
        bit(16)     ChannelLayoutExt;
    }
}
```

2.2.10.2.1 Semantics

ChannelLayoutExtFlag – If ChannelLayoutExtFlag is set to 1, then ChannelLayoutExt is present.

ReservedFlags – These flags are reserved for future definition and for the purposes of this specification shall be set to 0.

ChannelLayoutExt – extension to ChannelLayout, according to Table 2-6.

Table 2-6 - ChannelLayoutExt

Bit Masks	*Loudspeaker Location Description*	*Number of Channels*
0x0001	Center below in front	1
0x0002	Left/Right below in front	2
0x0004 – 0x8000	Reserved	

2.3 Storage of DTS elementary streams

Storage of DTS elementary streams within an ISO media file shall be according to this section.

- If codingname equals ‘dtse’, ‘dtsc’, ‘dtsh’, or ‘dtsl’, an audio sample shall consist of a single DTS audio frame, as defined in [3].
- If codingname equals ‘dts+’ or ‘dts-‘, the final construction of the audio frame following layer reconstruction shall comply with that defined in [3].

2.4 Restrictions on DTS Formats

This section describes the restrictions that apply to the DTS formats encapsulated in an ISO media file.

The following conditions shall not change in any DTS audio stream:

- Duration of Synchronized Frame
- Sampling Frequency
- Audio Channel Arrangement
- Low Frequency Effects flag
- Extension assignment

Annex A Implementation Guidelines (informative)

The information needed to derive the elements of the DTS Sample Entry box and boxes contained within it may be extracted from the respective elementary stream. DTS has tools available to implementers that will analyze DTS elementary streams and extract the information necessary to populate these parameters. DTS document number *9302J19200* [4], describes the function calls and return structures. To obtain this tool and additional documentation, please direct all document requests to DTS Business Support at BusinessSupport@dts.com .

Annex B Alternative Carrier Codec Guidelines for DTS Neural Surround™ (informative)

B.1 Introduction

DTS Neural Surround™ constitutes a method of encoding and decoding additional audio channels into a host audio stream. A common application example would be to encode 5.1 channel surround sound into a stereo audio program.

DTS Neural Surround coding is an active intensity phase encoding methodology, and does not carry embedded metadata signaling. A system unaware of this processing will operate without impediment. The nature of this encoding permits the carriage independent of audio encoding algorithm, provided sufficient quality is maintained.

A common system application example would be to use DTS Neural Surround to encode 5.1 channel surround sound for a stereo MPEG 1 layer 2 host audio, enabling the delivery of surround sound audio in early generations of digital media distribution systems designed for stereo audio distribution.

B.2 Signaling

B.2.1 Handler Reference Box

The syntax and values for the Handler Reference Box shall conform to section 8.4.3 of [1].

The value for the string “name” is a null-terminated string in UTF-8 characters which gives a human-readable name for the track type (for debugging and inspection purposes).

DTS recommends that in order for an audio track handler to recognize the presence of DTS Neural, one (and only one) of the following strings shall exist at the start of the “name” string field:

"DTSNeural522"
"DTSNeural722"
"DTSNeural725"

The appropriate string is chosen based on the type of Neural downmix performed prior to encoding into the “host” codec specified for this audio track object.