Ultra HD Forum Draft:
Ultra HD Forum Phase B Guidelines

April 07, 2018
Revision: 1.0

Ultra HD Forum
8377 Fremont Blvd., Suite 117,
Fremont, CA 94538
UNITED STATES

Notice

The Ultra HD Forum Guidelines are intended to serve the public interest by providing recommendations and procedures that promote uniformity of product, interchangeability and ultimately the long-term reliability of audio/video service transmission. This document shall not in any way preclude any member or nonmember of the Ultra HD Forum from manufacturing or selling products not conforming to such documents, nor shall the existence of such guidelines preclude their voluntary use by those other than Ultra HD Forum members, whether used domestically or internationally.

The Ultra HD Forum assumes no obligations or liability whatsoever to any party who may adopt the guidelines. Such adopting party assumes all risks associated with adoption of these guidelines, and accepts full responsibility for any damage and/or claims arising from the adoption of such guidelines.

Attention is called to the possibility that implementation of the recommendations and procedures described in these guidelines may require the use of subject matter covered by patent rights. By publication of these guidelines, no position is taken with respect to the existence or validity of any patent rights in connection therewith. Ultra HD Forum shall not be responsible for identifying patents for which a license may be required or for conducting inquiries into the legal validity or scope of those patents that are brought to its attention.

Patent holders who believe that they hold patents which are essential to the implementation of the recommendations and procedures described in these guidelines have been requested to provide information about those patents and any related licensing terms and conditions.

Revision History

Version	Date
Revision 1.0, released to the public	April 07, 2018

Table of Contents

Index of Tables and Figures

1. Purpose and Scope

Welcome to the first version of the Ultra HD Forum UHD Phase B Guidelines. While the UHD Phase A Guidelines were focused on UHD technologies that were commercially deployed as early as 2016, this Phase B Guidelines document focuses on the next generation of UHD technologies. Some of the Phase B technologies are already commercially deployed, while others are being actively tested and nearing commercial deployment

The Phase B technologies were carefully selected to help service operators plan for next generation UHD services. In August 2017, the Ultra HD Forum conducted a Service Operator Survey with the goal of learning about up-and-coming UHD technologies that have captured the interest of service operators. The survey results served as a guide to the Ultra HD Forum in drafting this document.

This version of the UHD Phase B Guidelines is a preliminary look at these important UHD technologies. The goal of this version is to introduce and de-mystify the technologies and provide information to operators that are considering incorporating one or more of these advanced features into their UHD services.

While this version focuses on descriptions and foundational information, future versions of this document will address incorporation of these technologies into an end-to-end workflow, as is the hallmark of the Ultra HD Forum. The Forum will consider the full ecosystem from production to distribution to rendering, with different distribution paths considered, such as internet streaming (aka, OTT), terrestrial broadcast (aka, OTA), and cable, satellite and IPTV delivery (aka, MVPD).

This document builds on the Ultra HD Forum UHD Phase A Guidelines. Readers will not find significant overlap between the Phase A and Phase B documents. As such, readers are encouraged to also familiarize themselves with the Phase A Guidelines [1].

2. References

This section contains references used in this text, which are an essential component of these guidelines. At the time of publication, the editions indicated were valid. All standards are subject to revision, and parties are encouraged to investigate the applicability of the most recent editions of the materials listed in this section.

2.1 Reference List

[1] Ultra HD Forum: "Phase A Guidelines - Revision 1.4," August 25, 2017, https://ultrahdforum.org/wp-content/uploads/Ultra-HD-Forum-Guidelines-v1.4-final-for-release.pdf

[2] ATSC: A/85:2013, "Techniques for Establishing and Maintaining Audio Loudness for Digital Television" , March 12, 2013, https://www.atsc.org/wp-content/uploads/2015/03/Techniques-for-establishing-and-maintaining-audio-loudness.pdf

[3] ATSC: A/300:2017, "ATSC 3.0 System", October 19, 2017, https://www.atsc.org/atsc-30-standard/a3002017-atsc-3-0-system/

[4] ATSC: A/322:2017, "Physical Layer Protocol", June 6, 2017, https://www.atsc.org/wp-content/uploads/2016/10/A322-2017a-Physical-Layer-Protocol.pdf

[5] ATSC: A/341:2018, "Video-HEVC with Amendments No. 1 and No. 2", March 9, 2018, https://www.atsc.org/wp-content/uploads/2017/05/A341-2018-Video-HEVC-1.pdf

[6] ATSC: A/342-1:2017, "Audio Common Elements", January 24, 2017, https://www.atsc.org/wp-content/uploads/2017/01/A342-1-2017-Audio-Part-1-5.pdf

[7] ATSC: A/342-2:2017, "AC-4 System", February 23, 2017, https://www.atsc.org/wp-content/uploads/2017/02/A342-2-2017-AC-4-System-5.pdf

[8] ATSC: A/342-3:2017, "MPEG-H System", March 3, 2017, https://www.atsc.org/wp-content/uploads/2017/03/A342-3-2017-MPEG-H-System-2.pdf

[9] CableLabs OC-TR-IP-MULTI-ARCH-C01-161026:2016, "IP Multicast Adaptive Bit Rate Architecture Technical Report", November 26, 2016, https://apps.cablelabs.com/specification/ip-multicast-adaptive-bit-rate-architecture-technical-report/

[10] CTA 861-G, "A DTV Profile for Uncompressed High Speed Digital Interfaces", November 2016, http://www.techstreet.com/standards/cta-861-g?product_id=1934129

[11] DASH-IF: "Guidelines for Implementation: DASH-IF Interoperability Points for ATSC 3.0, Version 1.0," DASH Interoperability Forum", January 31, 2016, http://dashif.org/wp-content/uploads/2017/02/DASH-IF-IOP-for-ATSC3-0-v1.0.pdf

[12] DVB: A168:2017, "MPEG-DASH Profile for Transport of ISO BMFF Based DVB Services over IP Based Networks", November 2017, https://www.dvb.org/resources/public/standards/a168_dvb_mpeg-dash_nov_2017.pdf

[13] EBU Tech 3364, "Audio Definition Model Metadata Specification Ver. 1.0", January 2014, https://tech.ebu.ch/docs/tech/tech3364.pdf

[14] EBU R 128, "Loudness Normalisation and Permitted Maximum Level of Audio Signals", June 2014, https://tech.ebu.ch/docs/r/r128.pdf

[15] ETSI TS 101 154 v2.3.1 (2017-02), "Digital Video Broadcasting (DVB); Specification for the use of Video and Audio Coding in Broadcasting Application based on the MPEG-2 Transport Stream", February 14, 2017, http://www.etsi.org/deliver/etsi_ts/101100_101199/101154/02.03.01_60/ts_101154v020301p.pdf

[16] ETSI TS 102 366 Annex H, "Digital Audio Compression (AC-3, Enhanced AC-3) Standard", August 20, 2008, http://www.etsi.org/deliver/etsi_ts/102300_102399/102366/01.02.01_60/ts_102366v010201p.pdf

[17] ETSI TS 103 190-2 (2015-09), "Digital Audio Compression (AC-4) Standard Part2: Immersive and personalized audio", September 25, 2015, http://www.etsi.org/deliver/etsi_ts/103100_103199/10319002/01.01.01_60/ts_10319002v010101p.pdf

[18] ETSI TS 103 433-1 v1.2.1 (2017-08), "High-Performance Single Layer Directly Standard Dynamic Range (SDR) Compatible High Dynamic Range (HDR) System for use in Consumer Electronics devices (SL-HDR1)", August, 11, 2017, http://www.etsi.org/deliver/etsi_ts/103400_103499/10343301/01.02.01_60/ts_10343301v010201p.pdf

[19] ETSI GS CCM 001 (2017-02), "Compound Content Management v1.1.1", February 8, 2017, http://www.etsi.org/deliver/etsi_gs/CCM/001_099/001/01.01.01_60/gs_ccm001v010101p.pdf

[20] APPLE "HLS Authoring Specification for Apple Devices", April 9, 2017, https://developer.apple.com/library/content/documentation/General/Reference/HLSAuthoringSpec/Requirements.html

[21] ISO/IEC: 14496-12, "Information technology—Coding of audio-visual objects—Part 12: ISO base media file format", December 2015, https://www.iso.org/standard/68960.html

[22] ISO/IEC: 23008-2, "Information technology -- High efficiency coding and media delivery in heterogeneous environments -- Part 2: High efficiency video coding, " May 2015, https://www.iso.org/standard/67660.html [1]

[23] ISO/IEC: 23008-3, "Information technology – High efficiency coding and media delivery in heterogeneous environments – Part 3: 3D audio", October 2015, https://www.iso.org/standard/63878.html; including ISO/IEC: 23008-3:2015/Amd 2:2016, "MPEG-H 3D Audio File Format Support ", September 2016, https://www.iso.org/standard/68592.html and ISO/IEC: 23008-3:2015/Amd 3:2017, "MPEG-H 3D Audio Phase 2", January 2017, https://www.iso.org/standard/69561.html

[24] ITU-R BS.1770-4, "Algorithms to measure audio programme loudness and true-peak audio level", November 2015, https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.1770-4-201510-I!!PDF-E.pdf

[25] ITU-R BS.1771, "Requirements for loudness and true-peak indicating meters", January 2012, https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.1771-1-201201-I!!PDF-E.pdf

[26] ITU-R BS.2076-1, "Audio Definition Model", June 2017, https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.2076-1-201706-I!!PDF-E.pdf

[1] Also published by ITU as ITU-T Recommendation H.265: 2015.

[27] ITU-R BS.2088-0, "Long-form file format for the international exchange of audio programme materials with metadata", October 2015, https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.2088-0-201510-I!!PDF-E.pdf

[28] ITU-R BR.1352-3, "File format for the exchange of audio program materials with metadata on information technology media", January 11, 2008, https://www.itu.int/dms_pubrec/itu-r/rec/br/R-REC-BR.1352-3-200712-W!!PDF-E.pdf

[29] ITU-R BT.1886, "Reference electro-optical transfer function for flat panel displays used in HDTV studio production", March 2011, https://www.itu.int/dms_pubrec/itu-r/rec/bt/R-REC-BT.1886-0-201103-I!!PDF-E.pdf

[30] ITU-R BT.2100-1, "Image parameter values for high dynamic range television for use in production and international programme exchange", June 2017, https://www.itu.int/dms_pubrec/itu-r/rec/bt/R-REC-BT.2100-1-201706-I!!PDF-E.pdf

[31] Report ITU-R BT.2390-3, "High dynamic range television for production and international programme exchange", October 2017, https://www.itu.int/dms_pub/itu-r/opb/rep/R-REP-BT.2390-3-2017-PDF-E.pdf

[32] SCTE: 135:2013, "Data-Over-Cable Service Interface Specification version 3.0", March 14, 2013, http://www.scte.org/documents/pdf/Standards/ANSI_SCTE 135-1 2013.pdf

[33] SCTE 242-3:2017, "Next Generation Audio Coding Constraints for Cable Systems: Part 3 –MPEG-H Audio Coding Constraints", September 25, 2017, http://www.scte.org/SCTEDocs/Standards/SCTE 242-3 2017.pdf

[34] SMPTE ST 337, "Format for Non-PCM Audio and Data in an AES3 Serial Digital Audio Interface" , May 6, 2015, http://ieeexplore.ieee.org/servlet/opac?punumber=7291909

[35] SMPTE ST 425-3:2015, "Image Format and Ancillary Data Mapping for the Dual Link 3Gb/s Serial Interface", June 21, 2015, http://ieeexplore.ieee.org/servlet/opac?punumber=7290046 ; and ST 425-5, "Image Format and Ancillary Data Mapping for the Quad Link 3Gb/s Serial Interface", June 21, 2015, http://ieeexplore.ieee.org/servlet/opac?punumber=7291843

[36] SMPTE ST 2022-2:2007, "Unidirectional Transport of Constant Bit Rate MPEG-2 Transport Streams on IP Networks", May 24, 2007, http://ieeexplore.ieee.org/servlet/opac?punumber=7291738

[37] SMPTE ST 2022-6:2012, "Transport of High Bit Rate Media Signals over IP Networks (HBRMT)", October 9, 2012, http://ieeexplore.ieee.org/servlet/opac?punumber=7289941

[38] SMPTE ST 2081-10:2018, "2160-line and 1080-line Source Image and Ancillary Data Mapping for 6G-SDI", March 12, 2018, http://ieeexplore.ieee.org/servlet/opac?punumber=8320053

[39] SMPTE ST 2082-10:2018, "2160-line and 1080-line Source Image and Ancillary Data Mapping for 12G-SDI", March 12, 2018, http://ieeexplore.ieee.org/servlet/opac?punumber=8320050

[40] SMPTE ST 2084:2014: "High Dynamic Range Electro-Optical Transfer Function of Mastering Reference Displays", August 29, 2014, http://ieeexplore.ieee.org/servlet/opac?punumber=7291450

[41] SMPTE ST 2086:2014, "Mastering Display Color Volume Metadata Supporting High Luminance And Wide Color Gamut Images", October 30, 2014, http://ieeexplore.ieee.org/servlet/opac?punumber=7291705

[42] SMPTE ST 2094-1:2016, "Dynamic Metadata for Color Volume Transform – Core Components", June 13, 2016, http://ieeexplore.ieee.org/servlet/opac?punumber=7513359

[43] SMPTE ST 2094-10:2016, "Dynamic Metadata for Color Volume Transform – Application #1", June 13, 2016, http://ieeexplore.ieee.org/servlet/opac?punumber=7513368

[44] SMPTE ST 2110-10:2017, "Professional Media Over IP Networks : System Timing and Definitions", November 27, 2017, http://ieeexplore.ieee.org/servlet/opac?punumber=8165972

[45] TTA: TTAK-KO-07.0127R1:2016, "Transmission and Reception for Terrestrial UHDTV Broadcasting Service", December 27, 2016, http://www.tta.or.kr/English/new/standardization/eng_ttastddesc.jsp?stdno=TTAK.KO-07.0127

3. Acronyms and Abbreviations

The below acronyms and abbreviations are used in this document. Definitions of the various terms are located in the relevant section of the document. For example, Next Generation Audio definitions are located in Section 7.1.

AVR	Audio/Video Receiver
BL	Base Layer
CAE	Content Aware Encoding or Content Adaptive Encoding
CBR	Constant Bitrate
CVBR	Capped Variable Bitrate
DE	Dialog Enhancement
DRC	Dynamic Range Control.
DRM	Digital Rights Management
EL	Enhancement Layer
HEVC	High Efficiency Video Coding [22]
HFR	High Frame Rate
HOA	Higher Order Ambisonics
ISO	International Standards Organization
ISOBMFF	ISO Base Media File Format [21]
LFE	Low Frequency Effects (Channel)
NGA	Next Generation Audio
PCM	Pulse-code Modulation
SEI	Supplemental Enhancement Information
SFR	Standard Frame Rate
SHVC	Scalable High-Efficiency Video Coding (see Annex H of [22])
VBR	Variable Bitrate
VDS	Video Description Service

4. Phase B Introduction

This version of the UHD Phase B Guidelines is a preliminary look at the next generation of UHD technologies. The goal of this version is to introduce and de-mystify the technologies and provide information to operators that are considering incorporating one or more of these advanced features into their UHD services. In order to document UHD technologies that are or will soon be relevant to the UHD ecosystem, the Ultra HD Forum considered the following Phase B selection criteria:

1. proven to be functional in an end-to-end workflow, either via early deployment or via interop testing to members' satisfaction,

AND

2. service providers have demonstrated interest in the technology.

4.1 Phase B Technologies

Phase B technologies are the next generation of UHD technologies that are expected to soon emerge, or are already emerging, in UHD service offerings. These technologies can be used in conjunction with the UHD Phase A technologies described in Phase A Guidelines [1].

- Dynamic HDR metadata systems, including Dolby Vision™[2] and SL-HDR1
- Dual layer HDR technologies
- High frame rate
- Next generation audio including Dolby® AC-4 and MPEG-H audio
- Content-aware encoding
- AVS2 codec[3]

[2] Dolby, Dolby Atmos, Dolby Digital and Dolby Vision are trademarks of Dolby Laboratories.

[3] The Ultra HD Forum is evaluating the AVS2 technology pending availability of an English language version of the specification. For this version of the Phase B Guidelines, it is included in an Annex for readers' convenience.

5. High Dynamic Range

5.1 Dolby Vision

Dolby Vision is an ecosystem solution to create, distribute and render HDR content with the ability to preserve artistic intent across a wide variety of distribution systems and consumer rendering environments. Dolby Vision began as a purely proprietary system, first introduced for OTT delivery. In order to make it suitable for use in Broadcasting the individual elements of the system have been incorporated into Standards issued by bodies such as SMPTE, ITU-R, ETSI, and ATSC, so that now Broadcast Standards can deliver the Dolby Vision experience.

Dolby Vision incorporates a number of key technologies, which are described and referenced in this document, including an optimized EOTF or Perceptual Quantizer, ("PQ"), increased bit depth (10 bit or 12 bit), wide color gamut, an improved color opponent encoding (IC_TC_P), re-shaping to optimize low-bit rate encoding, metadata for mastering display color volume parameters, and dynamic display mapping metadata.

Key technologies that have been incorporated into Standards:

- PQ EOTF and increased bit depth: SMPTE ST 2084 [40], Recommendation ITU-R BT.2100 [30]
- Wide color gamut: Recommendation ITU-R BT.2100 [30]
- IC_TC_P: Recommendation ITU-R BT.2100 [30]
- Mastering display metadata: SMPTE ST 2086 [41]
- Dynamic metadata: SMPTE 2094-10 [43]

5.1.1 Dolby Vision Encoding/Decoding Overview

Figure 1 illustrates a functional block diagram of the encoding system. HDR content in PQ is presented to the encoder. The video can undergo content analysis to create the display management data at the encoder (typically for Live encoding) or the data can be received from an upstream source (typically for prerecorded content in a file based workflow).

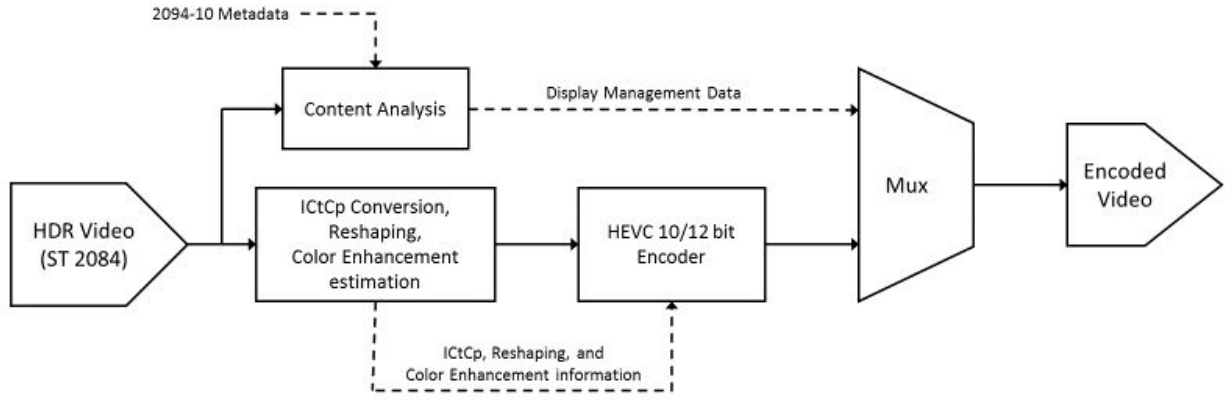

Figure 1 Encoder functional block diagram

If not natively in IC_TC_P space, it may be advantageous to convert the HDR video into IC_TC_P space. The video may be analyzed for reshaping and color enhancement information. If re-shaping is being employed to improve efficiency of delivery and apparent bit-depth, the pixel values are re-shaped (mapped by a re-shaping curve) so as to provide higher compression efficiency as compared to standard HEVC compression performance. The resulting reshaped HDR signal is then applied to the HEVC encoder and compressed. Simultaneously, the various signaling elements are then set and multiplexed with the static and dynamic display management metadata data, and are inserted into the stream (using the SEI message mechanism). This metadata enables improved rendering on displays that employ the Dolby Vision display mapping technology.

Figure 2 illustrates the functional block diagram of the decoder. It is important to note that the system in no way alters the HEVC decoder: An off-the-shelf, unmodified HEVC decoder is used, thereby preserving the investment made by hardware vendors and owners.

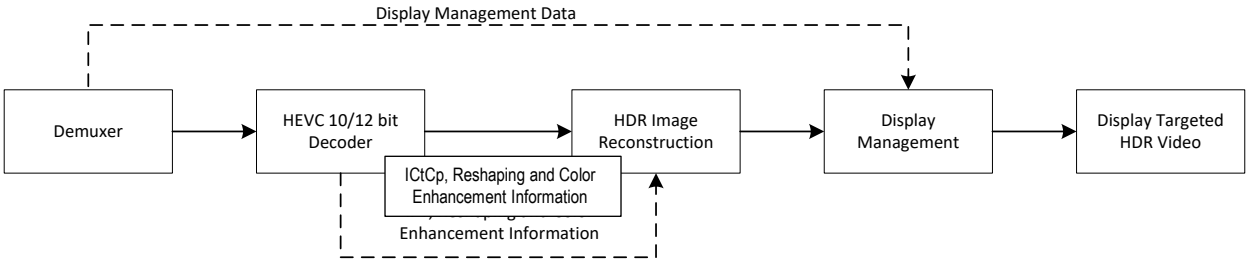

Figure 2 Decoder function block diagram

The HDR bitstream is demuxed in order to separate the various elements in the stream. The HDR video bitstream along with the signaling is passed to the standard HEVC decoder where the bitstream is decoded into the sequence of baseband images. If re-shaping was employed in encoding, the images are then un-shaped using the reshaping function back to the original luminance and chrominance range.

The display management data is separated during the demultiplexing step and sent to the display management block. In the case of a display that has the full capabilities of the HDR

mastering display in luminance range and color gamut, the reconstructed video can be displayed directly. In the case of a display that is a subset of the performance, some sort of display management is necessary. The display management block may be located in the terminal device such as in a television or mobile device or the data may be passed through a convertor or Set-Top Box to the final display device where the function would exist.

5.1.2 Dolby Vision Color Volume Mapping (Display Management)

Dolby Vision is designed to be scalable to support display of any arbitrary color volume within the BT.2100 standard [30], onto a display device of any color volume capability. The key is analysis of content on a scene-by-scene basis and the generation of metadata, which defines parameters of the source content; this metadata is then used to guide downstream color volume mapping based on the color volume of the target device. SMPTE ST 2094-10 [43] is the standardized mechanism to carry this metadata.

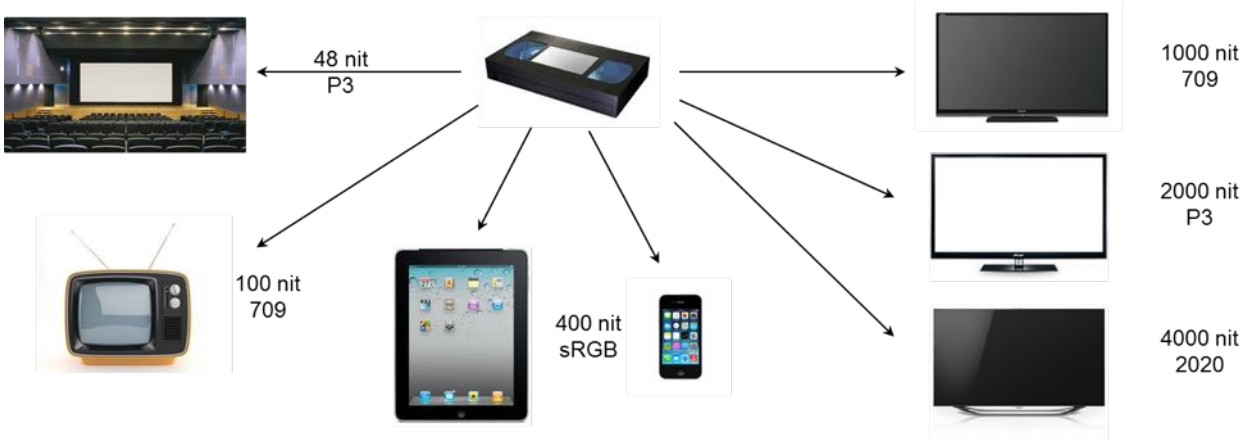

Figure 3 Example display device color volumes

While Dolby Vision works with the $Y'C'_BC'_R$ color representation model, in light of the limitations of $Y'C'_BC'_R$, especially at higher dynamic range, Dolby Vision also supports the use of IC_TC_P color representation model as defined in BT.2100 [30]. IC_TC_P isolates intensity from the color difference channels and may be a superior space in which to perform color volume mapping.

5.1.3 Dolby Vision in Broadcast

In a production facility, the general look and feel of the programming is established in the master control suite. Figure 4 shows a pictorial diagram of a typical broadcast production system. While each device in live production generally contains a monitoring display, only the main display located at the switcher is shown for simplicity. The programming look and feel is subject to the capabilities of the display used for creative approval – starting at the camera control unit and extending to the master control monitor.

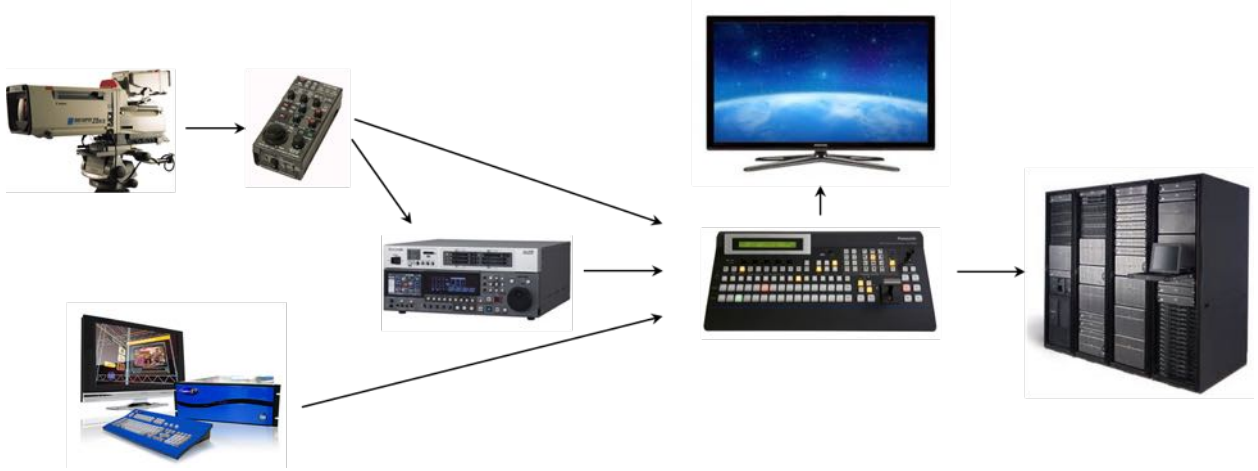

Figure 4 Example broadcast production facility components

Figure 5 shows a block diagram of the workflow in an HDR Broadcast facility using BT.2100 [30] PQ workflow. What is important to note is that in the transition phase from SDR to HDR, there will typically be a hybrid environment of both SDR and HDR devices and potentially a need to support both HDR and SDR outputs simultaneously. This is illustrated in the block diagram. In addition, because existing broadcast plants do not generally support metadata distribution today, the solution is to generate the ST 2094-10 [43] metadata in real time in just prior to, or inside of, the emission encoder as shown (block labelled "HPU" in green in Figure 5). In the case of generation at the encoder, the display management metadata can be inserted directly into the bitstream using standardized SEI messages by the HPU. Each payload of the display management metadata message is about 500 bits. It may be sent once per scene, per GOP, or per frame.

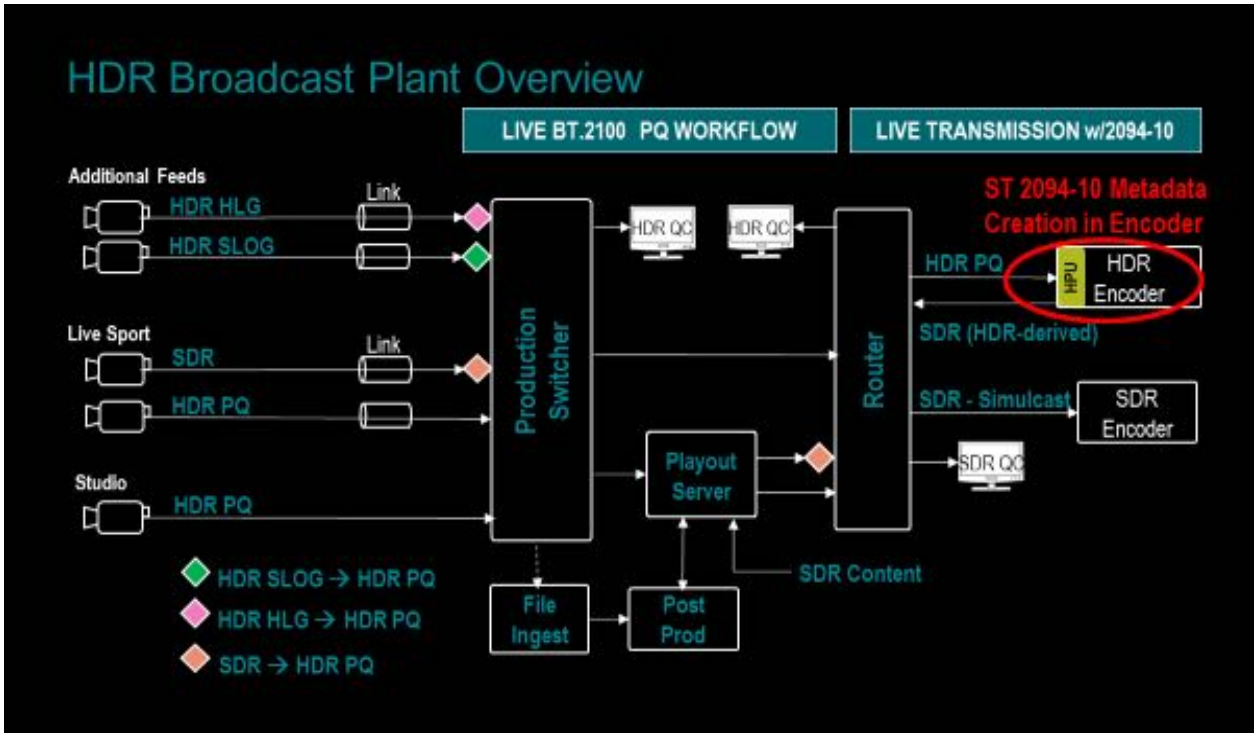

Figure 5 HDR broadcast production facility with BT.2100 PQ workflow-transition phase

Work is currently underway in SMPTE to standardize the carriage of HDR metadata via ANC packets in both SDI and IP interfaces. Once completed, this standard will allow the ST 2094-10 [43] dynamic metadata to be passed via SDI and IP links and interfaces through the broadcast plant to the encoder. This can be seen in Figure 6 where the metadata (shown in green blocks) would go from the camera or post production suite to the switcher/router (or an ancillary device) and then to the encoder. Using this method allows human control of the display mapping quality and consistency and would be useful for post-produced content such as commercials to preserve the intended look and feel as originally produced in the color suite while for live content, metadata could be generated in real time and passed via SDI/IP to the encoder, or generated in the encoder itself as mentioned in transition phase above.

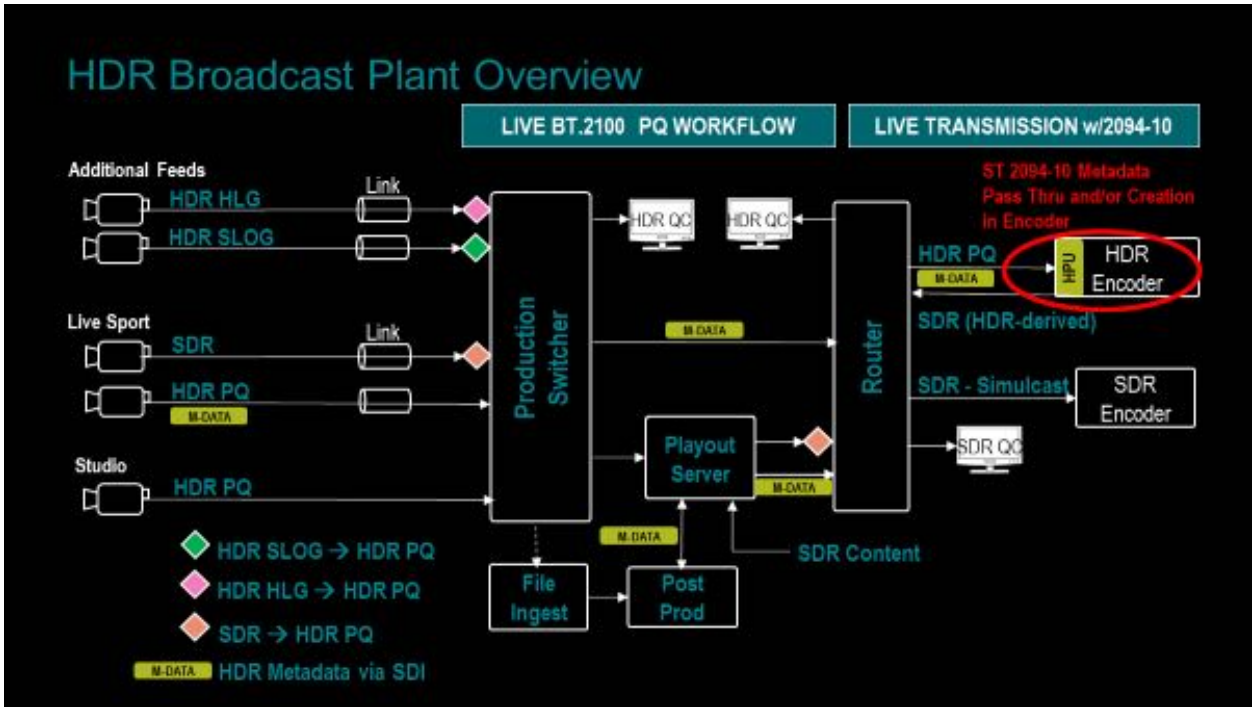

Figure 6 HDR broadcast production facility with BT.2100 PQ workflow- SDI metadata

5.2 Dual Layer HDR

Scalable High-Efficiency Video Coding (SHVC) is specified in Annex H of the HEVC specification [22]. Of particular interest is the ability of SHVC to decompose an image signal into two layers having different spatial resolutions: A Base Layer (BL), containing a lower resolution image, and an Enhancement Layer (EL), which contributes higher resolution details. When the enhancement layer is combined with the BL image, a higher resolution image is reconstituted. SHVC is commonly shown to support resolution scaling of 1.5x or 2x, so for example a BL might provide a 540p image, which may be combined with a 1080p EL. While SHVC allows an AVC-coded BL with an HEVC-coded EL, encoding the BL at the same quality using HEVC consumes less bandwidth.

The BL parameters are selected for use over a lower bitrate channel. The BL container, or the channel carrying it, should provide error resiliency. Such a BL is well suited for use when an OTT channel suffers from bandwidth constraints or network congestion, or when an OTA receiver is mobile or is located inside of a building without an external antenna.

The EL targets devices with more reliable access and higher bandwidth, e.g., a stationary OTA receiver, particularly one with a fixed, external antenna or one having access to a fast broadband connection for receiving a hybrid service (ATSC 3.0 supports a hybrid mode service delivery, see [3] section 5.1.6, wherein one or more program elements may be transported over a broadband path, as might be used for an EL). The EL may be delivered over a less resilient channel, since if lost, the image decoded from the BL is likely to remain available. The ability to tradeoff capacity and robustness is a significant feature of the physical layer protocols in ATSC 3.0, as discussed in Section 4.1 of [4] and in more detail elsewhere in that document.

To support fast channel changes, the BL may be encoded with a short GOP (e.g., 1/2 second), allowing fast picture acquisition, whereas the EL may be encoded with a long GOP (e.g., 2-4 seconds), to improve coding efficiency.

While SHVC permits configurations, where the color gamuts and/or transfer functions of the base and ELs are different, acquisition or loss of the EL in such configurations may result in an undesirable change to image appearance, compromising the viewing experience. Caution is warranted if the selection of the color gamut and transfer function is not the same for both the base and ELs.

Thus, though SHVC supports many differences between the image characteristics of the BL and EL, including variation in color space, transfer function, bit depth, and frame rate, for Phase B, only differences in spatial resolution and quality are supported. In addition, while SHVC permits use of multiple ELs, only a single EL is used in Phase B.

The combined BL and ELs should provide UHD Phase A content, i.e., HDR plus WCG at a resolution of at least 1080p, unless receipt of the EL is interrupted. The BL by itself is a lower resolution image, which alone might not qualify as UHD Phase A content. For example, for reception on a mobile device, a 540p BL may be selected, with a 1080p EL. Both layers may be provided in HDR plus WCG, but here, the EL is necessary to obtain sufficient resolution to qualify as UHD Phase A content.

As an alternative, the base and ELs may be provided in an SDR format, which with metadata (see [18]) provided in either one of the two layers is decodable as HDR plus WCG, yet allows non-HDR devices to provide a picture with either just the BL, or both the base and ELs.

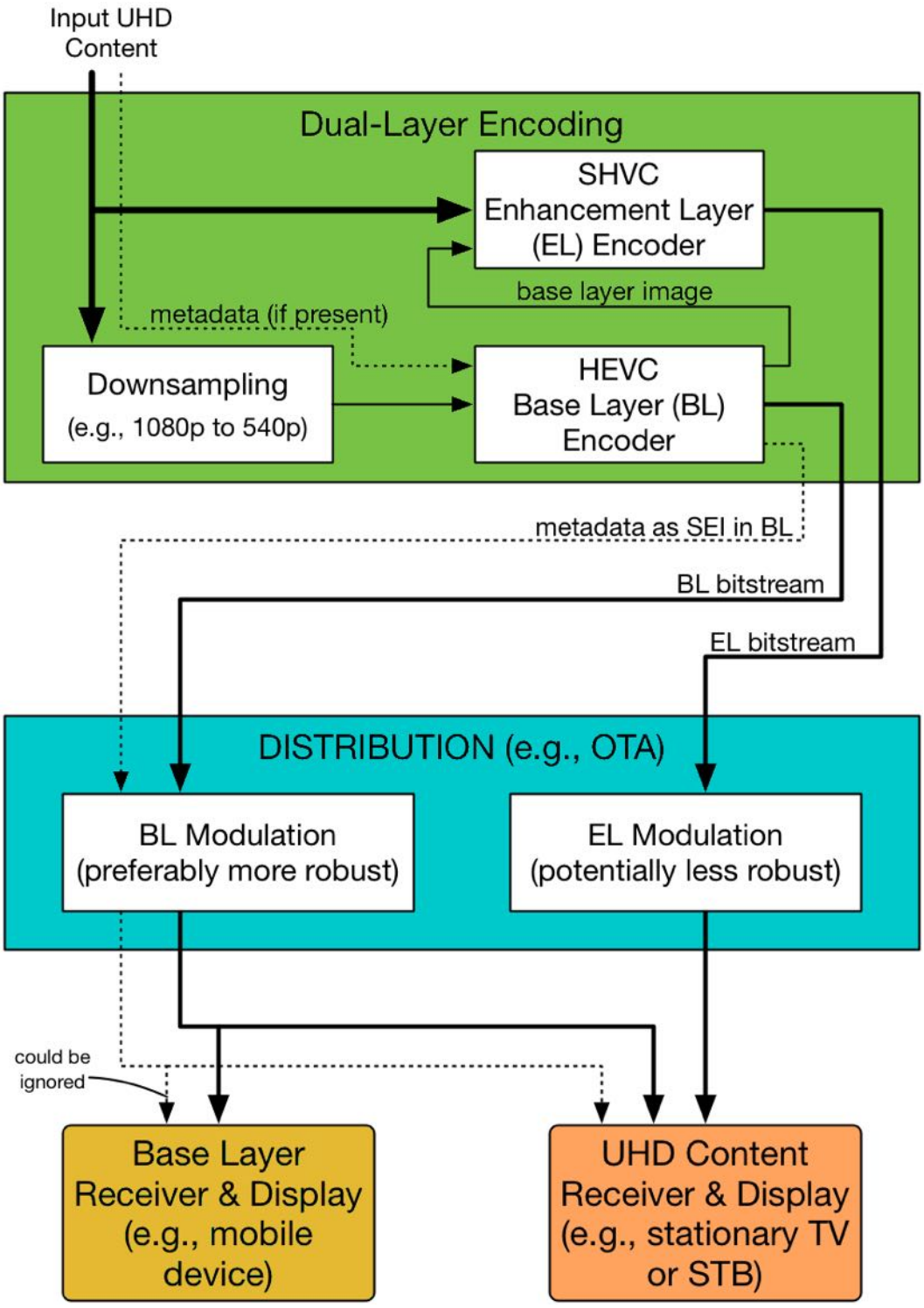

Figure 7 Example Phase B dual-Layer encoding and distribution

Figure 7 shows one configuration of the functional blocks for SHVC encoding, including the routing and embedding of metadata, which might be static or dynamic, into the preferably more robust BL bitstream. Other configurations (not shown) may embed the metadata into the EL bitstream, which is a case for which SL-HDR (see [18]) is well-suited, given that its error-

concealment process (described in Annex F of [18]) means that a loss of the less robust EL won't have as significant an effect as it might otherwise: When switching to the BL alone, the resulting image would lose detail, but the general HDR characteristics would remain, though ceasing to be dynamic.

In this example, distribution is by terrestrial broadcast (OTA) where the different bitstreams are separately modulated. Receiving stations may receive only the BL, or both the BL & EL as appropriate. Some receivers might ignore metadata provided in either bitstream (for example, as suggested for the BL-only receiver). As described above, for a hybrid distribution service, the BL would be distributed via OTA as shown, while the EL would be distributed via broadband connection. While SHVC is also supported by DASH, so that when connection bandwidth is limited, a DASH client may select only the BL, but as the connection bandwidth increases, the DASH client may additionally select the EL, so while not specifically a Phase B distribution mechanism, dual layer distribution is suitable for OTT distribution as well, both for VOD and linear programs.

5.3 SL-HDR1

As pointed out in Section 7.2 of the Phase A Guidelines [1], ETSI TS 103 433-1 [18] describes a method of down-conversion to derive an SDR/BT.709 signal from an HDR/WCG signal. The process supports PQ, HLG, and other HDR/WCG formats (see Section 6.3.2 of [1]) and may optionally deliver SDR/BT.2020 as the down-conversion target.

This ETSI specification additionally specifies a mechanism for generating an SL-HDR information SEI message (defined in Annex A.2 of [18]) to carry dynamic color volume transform metadata created during the down-conversion process. A receiver may use the SL-HDR information in conjunction with the SDR/BT.709 signal to reconstruct the HDR/WCG video.

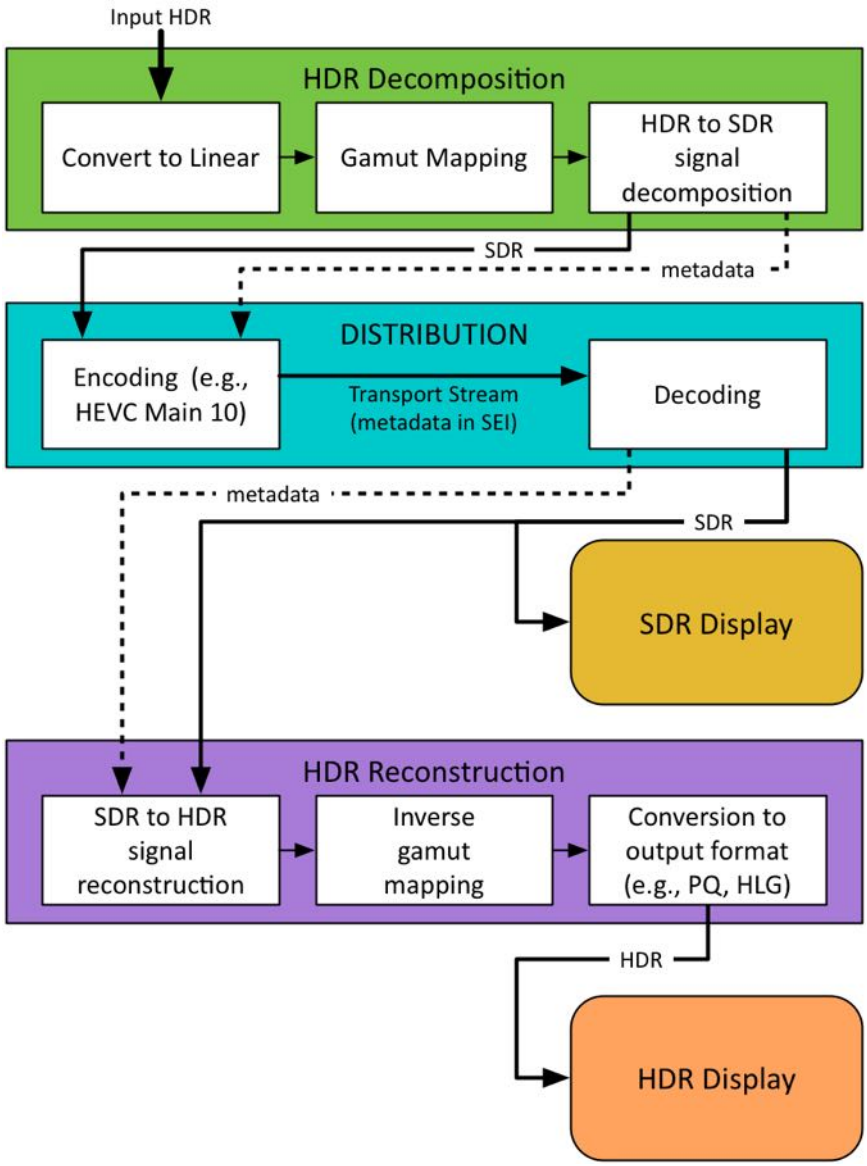

Figure 8 SL-HDR processing, distribution, reconstruction, and presentation

Figure 8 represents a typical use case of SL-HDR being used for distribution of HDR content. The down-conversion process applied to input HDR content occurs immediately before distribution encoding and comprises an HDR decomposition step and an optional gamut mapping step, which generates reconstruction metadata in addition to the SDR/BT.709 signal, making this down-conversion invertible.

For distribution, the metadata is embedded in the HEVC bitstream as SL-HDR information SEI messages, defined in [18], which accompany the encoded SDR/BT.709 content. The resulting stream may be used for either primary or final distribution. In either case, the SL-HDR metadata enables optional reconstruction of the HDR/WCG signal by downstream recipients.

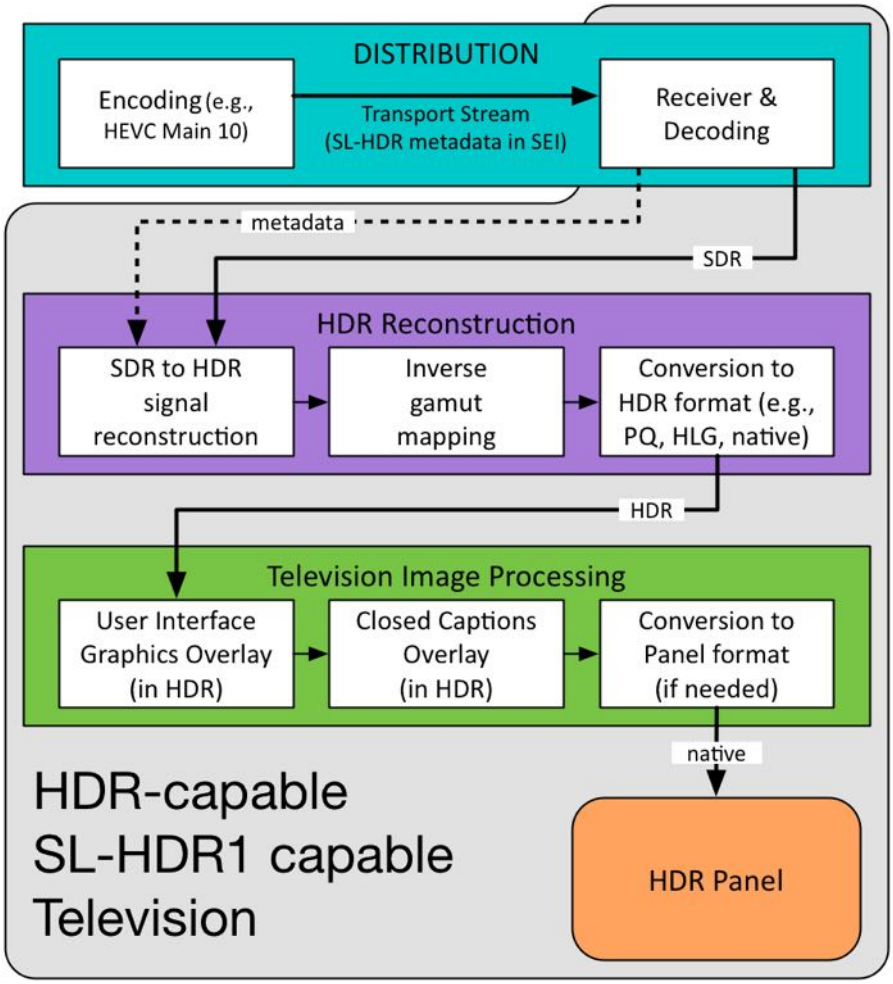

Figure 9 Direct reception of SL-HDR signal by an SL-HDR1 capable television

Upon receipt of an SL-HDR distribution, the SDR/BT.709 signal and metadata may be used by legacy devices by using the SDR/BT.709 format for presentation of the SDR/BT.709 image and ignoring the metadata, as illustrated by the SDR display in Figure 8 if received by a decoder that recognizes the metadata and is connected to an HDR/WCG display, the metadata may be used by the decoder to reconstruct the HDR/WCG image, with the reconstruction taking place as shown by the HDR reconstruction block of Figure 8.

This system addresses both integrated decoder/displays and separate decoder/displays such as a STB connected to a display.

In the case where an SL-HDR capable television receives a signal directly, as shown in Figure 9, the decoder recognizes metadata to be used to map the HDR/WCG video to an HDR format suitable for subsequent internal image processing (e.g., overlaying graphics and/or captions) before the images are supplied to the display panel.

If the same signal is received by a television without SL-HDR capability (not shown), the metadata is ignored, an HDR/WCG picture is not reconstructed, and the set will output the SDR/BT.709 picture.

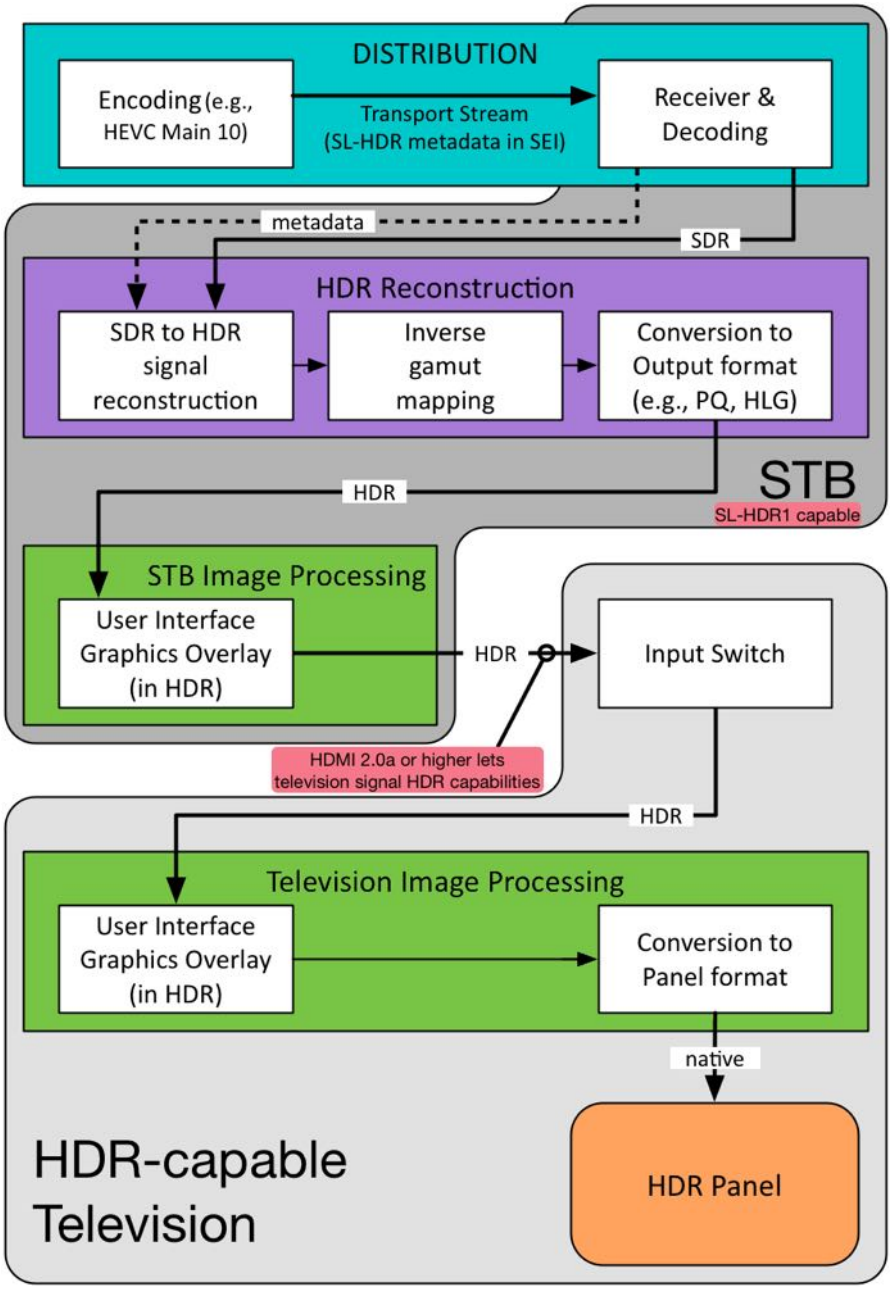

Figure 10 STB processing of SL-HDR signals for an HDR-capable television

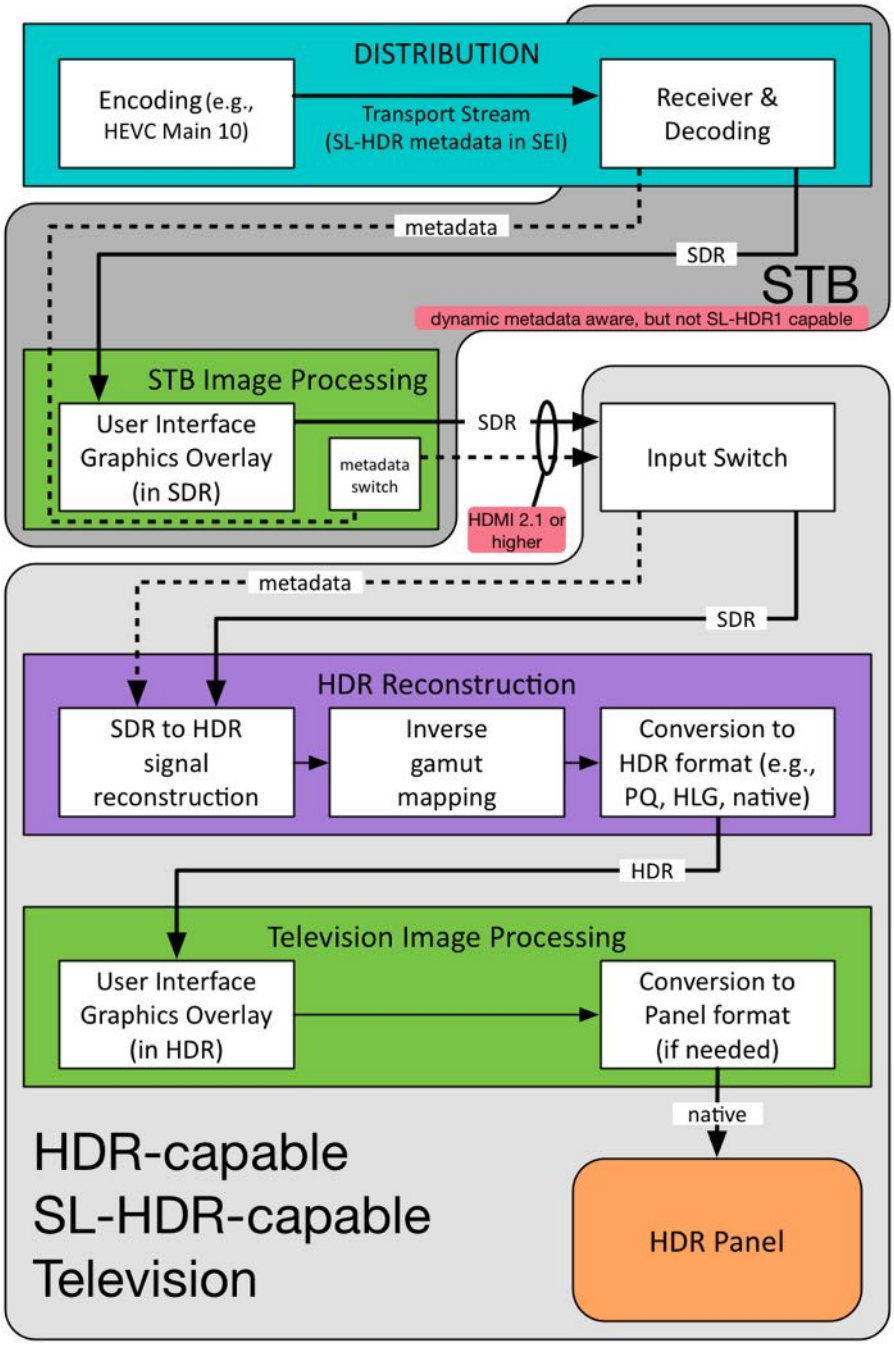

Figure 11 STB passing SL-HDR to an SL-HDR1 capable television

STBs will be used as OTA conversion boxes for televisions unable to receive appropriate OTA signals directly, and for all television sets in other distribution models. In the case of an STB implementing a decoder separate from the display, where the decoder is able to apply the SL-HDR metadata, as shown in Figure 10, then the STB may query the interface with the display device (e.g., via HDMI 2.0a or higher, using the signaling described in [10]) to determine whether the display is HDR-capable, and if so, may use the metadata to reconstruct, in an appropriate gamut, the HDR image to be passed to the display. If graphics are to be overlaid by the STB (e.g. captions,

user interface menus or an EPG), the STB overlays graphics after the HDR reconstruction, such that the graphics are overlaid in the same mode that is being provided to the display.

A similar strategy, that is, reconstructing the HDR/WCG video before image manipulations such as graphics overlays, is recommended for use in professional environments and is discussed below in conjunction with Figure 13.

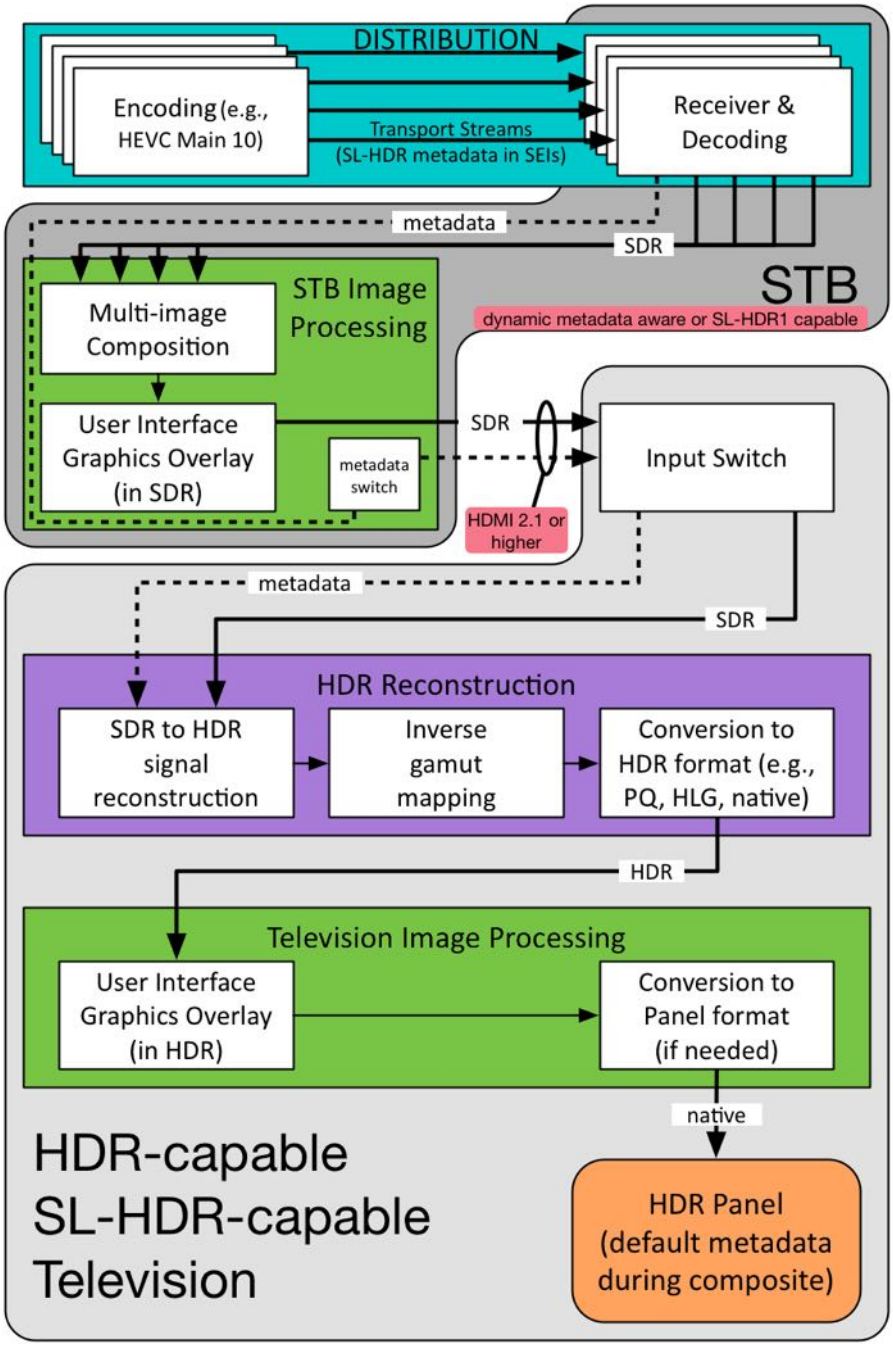

Figure 12 Multiple SL-HDR channels received and composited in SDR by an STB

If, as in Figure 11, an STB is not capable of using the SL-HDR information messages to reconstruct the HDR/WCG video, but the display has indicated (here, via HDMI 2.1 or higher, signaled as in [10]) that such information would be meaningful, then the STB may pass the SL-HDR information to the display in conjunction with the SDR video, enabling the television to reconstruct the HDR/WCG image.

In this scenario, if the STB were to first overlay SDR graphics (e.g., captions, user interface or EPG) before passing the SDR video along to the display, the STB has two options, illustrated as the "metadata switch" in Figure 11. The first option is to retain the original SL-HDR information, which is dynamic. The second option is to revert to default values for the metadata as prescribed in Annex F of [18]. Either choice allows the display to maintain the same interface mode and does not induce a restart of the television's display processing pipeline, thereby not interrupting the user experience. The former choice, the dynamic metadata, may in rare cases produce a "breathing" effect that influences the appearance of only the STB-provided graphics. Television-supplied graphics are unaffected. Switching to the specified default values mitigates the breathing effect, yet allows the SL-HDR capable television to properly adapt the reconstructed HDR/WCG image to its display panel capabilities

Another use for the default values appears when multiple video sources are composited in an STB for multi-channel display, as when a user selects a multiple sports or news channels that all play simultaneously (though typically with audio only from one). This requires that multiple channels are received and decoded individually, but then composited into a single image, perhaps with graphics added, as seen in Figure 12. In such a case, none of the SL-HDR metadata provided by one incoming video stream is likely to apply to the other sources, so the default values for the metadata is an appropriate choice. If the STB is SL-HDR1 capable, then each of the channels could be individually reconstructed with the corresponding metadata to a common HDR format, with the compositing taking place in HDR and the resulting image being passed to the television with metadata already consumed.

Where neither the STB nor the display recognize the SL-HDR information messages, the decoder decodes the SDR/BT.709 image, which is then presented by the display. Thus, in any case, the SDR/BT.709 image may be presented if the metadata does not reach the decoder or cannot be interpreted for any reason. This offers particular advantages during the transition to widespread HDR deployment.

Figure 8 shows HDR decomposition and encoding taking place in the broadcast facility immediately before emission. A significant benefit to this workflow is that there is no requirement for metadata to be transported throughout the broadcast facility when using the SL-HDR technique. For such facilities, the HDR decomposition is preferably integrated into the encoder fed by the HDR signal but, in the alternative, the HDR decomposition may be performed by a pre-processor from which the resulting SDR video is passed to an encoder that also accepts the SL-HDR information, carried for example as a message in SDI vertical ancillary data (as described in [43]) of the SDR video signal, for incorporation into the bitstream. Handling of such signals as contribution feeds to downstream affiliates and MVPDs is discussed below in conjunction with Figure 13 and Figure 14.

Where valuable to support the needs of a particular workflow, a different approach may be taken, in which the HDR decomposition takes place earlier and relies on the SDR video signal and metadata being carried within the broadcast facility. In this workflow, the SDR signal is usable by legacy monitors and multi-viewers, even if the metadata is not. As components within the broadcast facility are upgraded over time, each may utilize the metadata when and as needed to

reconstruct the HDR signal. Once the entire facility has transitioned to being HDR capable, the decomposition and metadata are no longer needed until the point of emission, though an HDR-based broadcast facility may want to keep an SL-HDR down-converter at various points to produce an SDR version of their feed for production QA purposes.

An SL-HDR-based emission may be used as a contribution feed to downstream affiliate stations. This has the advantage of supporting with a single backhaul those affiliates ready to accept HDR signals and those affiliates that have not yet transitioned to HDR and still require SDR for a contribution feed. This is also an advantage for MVPDs receiving an HDR signal but providing an SDR service.

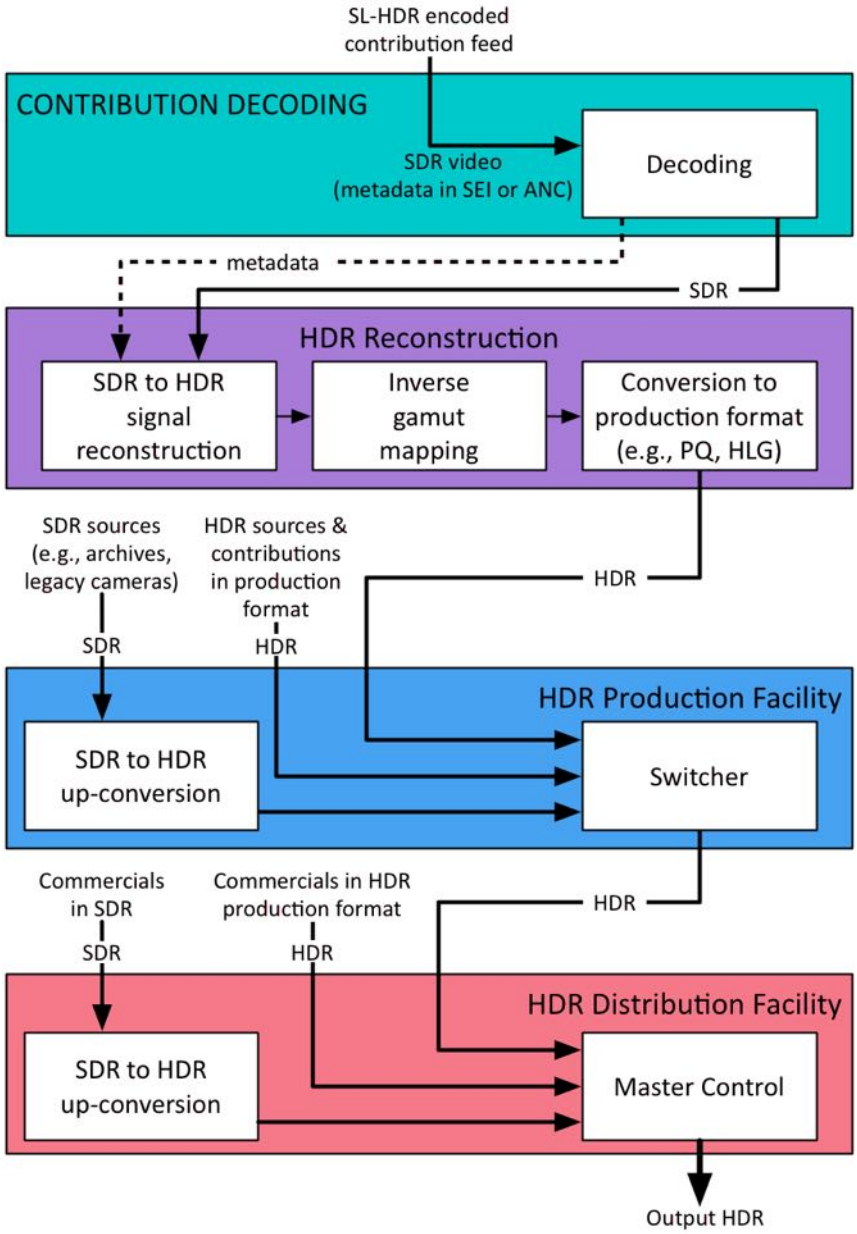

Figure 13 SL-HDR as a contribution feed to an HDR facility

The workflow for an HDR-ready affiliate receiving an SDR video with SL-HDR metadata as a contribution feed is shown in Figure 13. The decoding block and the HDR reconstruction block resemble the like-named blocks in Figure 8, with one potential exception: In Figure 13, the inverse gamut mapping block should use the invertible gamut mapping described in Annex D of [18] as this provides a visually lossless round-trip conversion.

In HDR-based production and distribution facilities, such as shown in the example of Figure 13, facility operations should rely as much as possible on a single HDR format. In the example facility shown, production and distribution does not rely on metadata being transported through the facility, as supported by such HDR formats as PQ10, HLG, Slog3, and others. Where metadata may be carried through equipment and between systems, e.g., the switcher, HDR formats requiring metadata, such as HDR10, may be used.

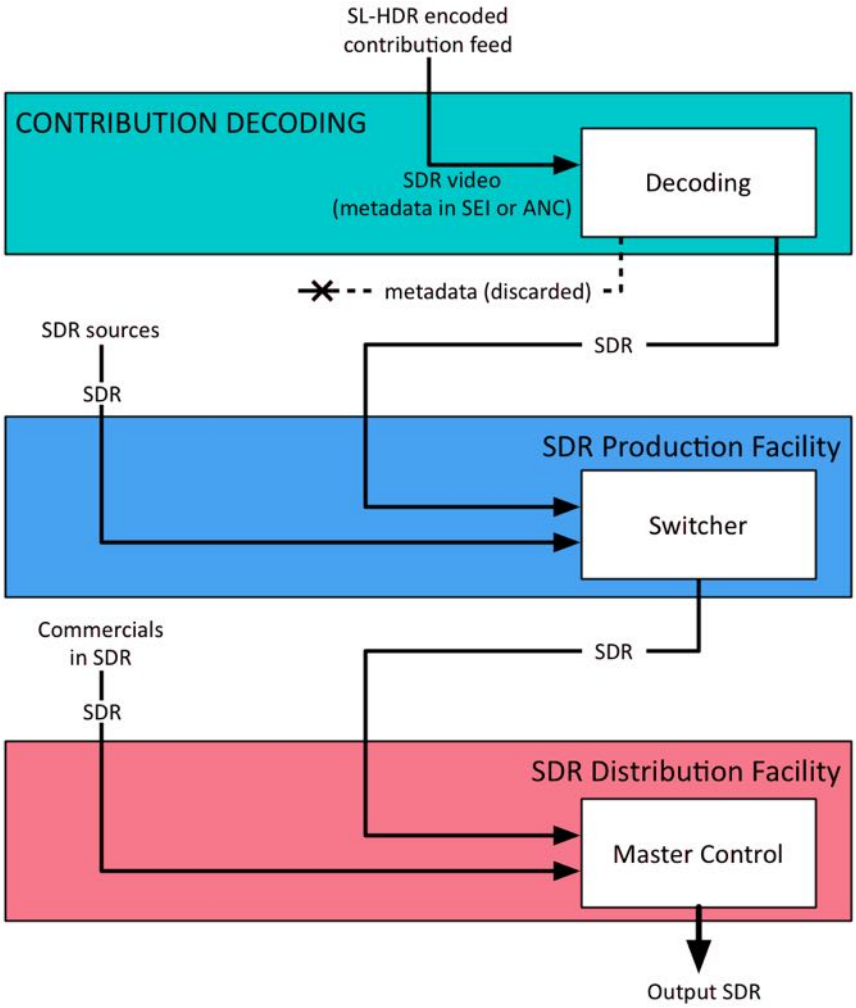

Figure 14 SL-HDR as a contribution feed to an SDR facility

In an HDR-based facility, the output HDR is complete immediately prior to the emission encode. As shown in Figure 8, this HDR signal is passed through the HDR decomposition and encode processes. With this architecture, a distribution facility has available the signals to distribute to an HDR-only channel using the Input HDR (though this may exhibit black screens for non-HDR-compatible consumer equipment), an SDR-only channel by encoding the SDR

signal, but no metadata (upon which no equipment may take advantage of the HDR production), and a channel that carries SDR video with SL-HDR metadata, which may address consumer equipment of either type with no black screens.

Figure 14 shows an SDR-based affiliate receiving an SL-HDR encoded contribution feed. Upon decode, only SDR video is produced, while the SL-HDR information carried in the contribution feed is discarded. This facility implements no HDR reconstruction and all customers downstream of this affiliate will receive the signal as SDR video with no SL-HDR information. This mode of operation is considered suitable for those downstream affiliates or markets that will be late to convert to HDR operation.

In the case of an MVPD, distribution as SDR with SL-HDR information for HDR reconstruction is particularly well suited, because the HDR decomposition process shown in Figure 8 and detailed in Annex C of [18] is expected to be performed by professional equipment not subject to the computational constraints of consumer premises equipment. Professional equipment is more likely to receive updates, improvements, and may be more easily upgraded, whereas STBs on customer premises may not be upgradeable and therefore may remain fixed for the life of their installation. Further, performance of such a down-conversion before distribution more consistently provides a quality presentation at the customer end. The HDR reconstruction process of Figure 8, by contrast, is considerably lighter weight computationally, and as such well suited to consumer premises equipment, and widely available for inclusion in hardware.

6. High Frame Rate

6.1 Introduction

For the purpose of this document, High Frame Rate (HFR) refers to frame rates of 100fps or higher, including 100, 120/1.001[4] and 120, Standard Framer Rate (SFR) refers to frame rates of 60fps or lower, which are commonly used including 24/1.001, 24, 25, 30/1.001, 30, 50, 60/1.001 and 60.

According to a SMPTE/HPA paper authored by Mark Schubin[5], frame rates of 100, 120/1.001 and 120 fps add significant clarity to high motion video such as sports or action scenes. Schubin also notes that high dynamic range put new demands on temporal resolution. He notes that, "Viewers of HDR imagery sometimes report increased perception of motion judder... Increased frame rate, therefore, might be necessary to accompany HDR."

Citing Schubin again, "... the [EBU[6]] found ... that in going from 60 frames per second (fps) to 120 fps or from 120 fps to 240 fps — a doubling of the frame rate — it is possible to achieve a full grade of improvement." Further, doubling the frame rate from 50/60 fps to 100/120 fps is a very efficient means of gaining that full grade of improvement when compared to going from 2K to 4K spatial resolution, as illustrated in Figure 15.

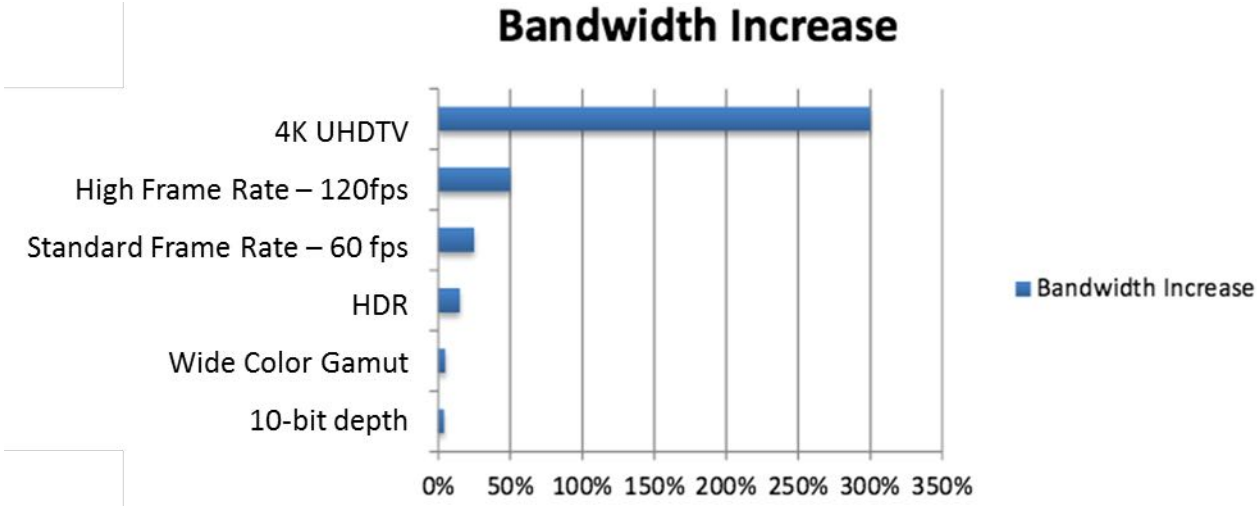

Figure 15 Bandwidth increases for various video format improvements

HFR has been included in newer over-the-air television standards including ATSC 3.0 [5] and DVB [15]. As such, the Ultra HD Forum considers HFR to be a Phase B technology for over-the-

[4] Although 120/1.001 is considered an example of HFR, he Ultra HD Forum recommends using integer frame rates for all UHD content whenever possible.

[5] "Higher Resolution, Higher Frame Rate, and Better Pixels in Context The Visual Quality Improvement Each Can Offer, and at What Cost", SMPTE/HPA paper, Mark Schubin, 2014, https://www.smpte.org/publications/industry-perspectives/schubin-HPA2014

[6] Rep. ITU-R BT.2246-6 The present state of ultra-high definition television; https://www.itu.int/dms_pub/itu-r/opb/rep/R-REP-BT.2246-6-2017-PDF-E.pdf; p 26.

air (OTA). Both systems include a backward compatibility mechanism that enables 50/60 fps decoders to render a 50/60 fps version of the content while 100/120 fps decoders render the full HFR experience. See Section 6.3 for more information about backward compatibility.

Looking beyond Phase B, the Ultra HD Forum anticipates that more HFR content will become available via more distribution channels, including with 4K resolutions. HFR could potentially play an important role in Virtual Reality content.

6.2 Phase B HFR Video Format Parameters

In the Phase B timeframe, 4K HFR would exceed the capabilities of some portions of the end-to-end ecosystem. For example, while HDMI 2.1 supports 4K 120 fps or 8K 60 fps, most production environment transport systems currently support only 2K with 100/120 fps. Although this is likely to change in the future, the Ultra HD Forum describes HFR with 2K spatial resolution for Phase B in order to provide an HFR guideline for a full end-to-end system. The parameters for HFR content in Phase B are shown in Table 1.

Table 1 Phase B high frame rate content parameters

Frame Rate	Spatial Resolution	Scan Type	Dynamic Range	Color Space	Bit Depth	Distribution Codec
100, 120[7]	HD	Progressive	SDR, HDR	709, 2020	10	HEVC Main 10 Level 5.1

6.3 Backward Compatibility for HFR

Both DVB UHD-1 Phase 2 (ETSI TS 101 154 v. 2.3.1) [15] and ATSC 3.0 (A/341) [5] include framerates up to 120 fps. Both documents further include optional temporal sub-layering for backward compatibility to a frame rate half of the HFR. ATSC A/341 additionally includes optional temporal filtering for enhancing the standard framerate picture when temporal sub-layering is used.

Both DVB and ATSC make use of the HEVC [22] Temporal Sub-layers technology to label every other frame for use by a 50/60 fps decoder.

In the case that an HFR video stream is available, an SFR stream may be extracted by dropping every other picture. HEVC temporal sub-layering identifies every other picture, which enables division of the stream prior to decompression. Note that strobe effects may be present when dropping every other frame without applying any filtering. Filtering systems such as the one shown in Figure 16 can mitigate this effect.

The SFR frames are Temporal ID = 0 and the additional frames needed for HFR are Temporal ID = 1. SFR devices render the frames with ID = 0 and HFR devices render all frames, i.e., ID = 0 and ID = 1.

[7] Note that 120/1.001 may be used for backward compatibility; however, the Ultra HD Forum recommends using integer frame rates for all UHD content whenever possible.

In the case of DVB, separate MPEG-2 TS Packet Identifiers (PIDs) are used to carry the two sub-layers. SFR decoders completely ignore the PID carrying the frames with Temporal ID = 1.

ATSC 3.0, one video stream includes both temporal video sub-steams (for ROUTE/DASH protocol implementations). ATSC 3.0 is a non-backward compatible system, so that devices that are capable of decoding ATSC 3.0 content are by definition new devices, and thus SFR ATSC 3.0 devices can be designed to correctly render the SFR portion of a temporal sub-layered stream.

The process of recording of HFR content in either compressed or uncompressed form should take into account the possibility that the content may undergo transformations in downstream processing to make the content backward compatible with SFR TVs, using one of the mechanisms described in this Section. Information that will be consumed by an SFR TV must be embedded in the images that will form the base layer of the temporally layered stream. For example, if CTA-608/CTA-708 captions are stored in the uncompressed image data, this data should be stored in the image data that would become the base layer and not the enhancement layer (since the SFR TV will not have access to the latter).

ATSC 3.0 includes an additional feature for backward compatibility called Temporal Filtering. According to A/341 [5], achieving backward compatibility by rendering every other frame may cause unwanted strobing. Temporal Filtering is a method by which the SFR frames are filtered in order to optimize the SFR experience. The SFR device plays the filtered SFR frames. The HFR device recovers the original, pre-filtered SFR frames and renders all frames (i.e., the pre-filtered frames are optimized for HFR). See Figure 16.

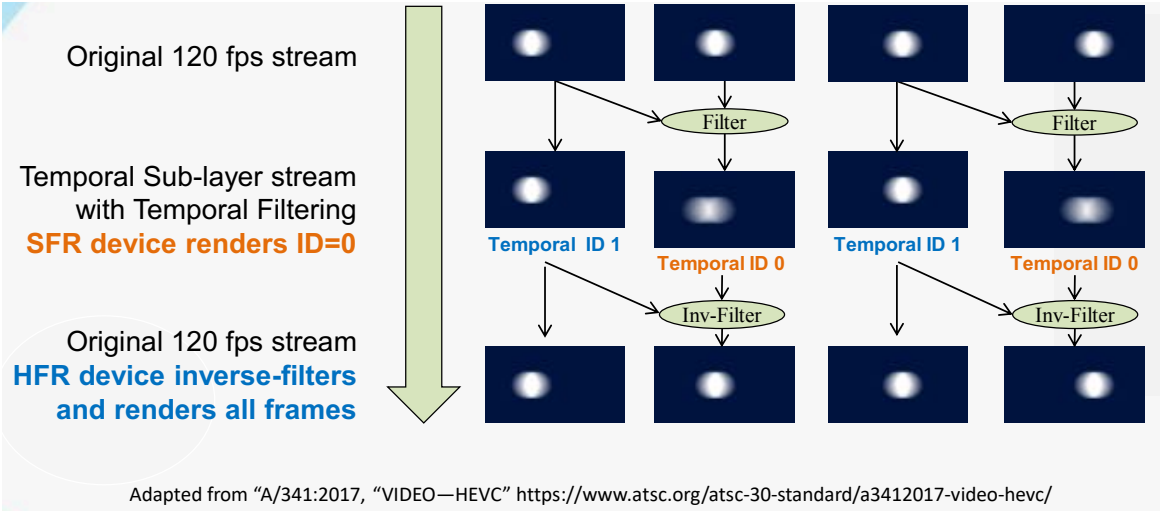

Figure 16 ATSC 3.0 temporal filtering for HFR backward compatibility

For ABR services, if a temporal layering scheme is used for backward compatibility, maintain temporal layering, and adjust overall bit rate. We do not recommend changing the frame rate to compensate for decrease in network bandwidth, viz., by eliminating the enhancement layer. HFR TV behavior is not predictable if a stream transitions between a temporally layered stream and a single layer stream.

A DVB broadcaster may ingest HFR content in compressed format that uses ATSC 3.0 temporal layering. The converse may also occur, where an ATSC 3.0 broadcaster ingests content in DVB dual layer format. Since these use cases will typically also involve frame rate conversion (between say 100p and 120p), which will require transcoders to be used, these transcoders can also

convert the temporal layering from one format to the other, or convert between a temporally layered HFR stream and a single-layered HFR stream. When transcoding to an ATSC 3.0 temporally layered HFR stream, temporal filtering for judder reduction can also be implemented by the transcoder if so desired.

6.4 Production Considerations for HFR

Since Phase-B limits HFR to 1080p 100 or 120 frame rates, interfaces between production systems such as cameras, switchers, storage and playout servers require no more that 3G SDI capability. Current state-of-the-art systems either incorporate these interfaces or are in the process of being revised to incorporate this capability.

The payload of HFR 1080p content can be carried via the SDI interface as described in SMPTE ST 425 [35], 2081 [36] and 2082 [39] document families using 3G or 12G interfaces. (6G is also possible, but not used in common practice.) Some examples include:

- 10-bit 4:2:2 over dual link 3G or single 12G
- 10-bit 4:4:4 over quad link 3G or single 12G

ST 2082-10 [39] includes information about signaling HFR content over the SDI interface. Experiments using 3G dual link with every other frame carried on each of the two interfaces are underway.

Carriage of HFR 1080p content via IP networks is described in SMPTE ST 2022 [36]. SMPTE ST 2022-6 [37] describes how to map SDI info to IP. The SMPTE ST 2110 [44] standards suite specifies the carriage, synchronization, and description of separate elementary essence streams over IP for real-time production, playout, and other professional media applications; i.e., it describes how each element of essence is mapped to IP.

7. Next Generation Audio

7.1 Terms and Definitions

The below terms and definitions pertain to Next Generation Audio systems (NGA). See also Table 2, which contains additional terms and Section 7.2.6 Audio Element Formats, which are provided courtesy of the Advanced Television Systems Committee from Standard A/342 Part 1, Audio Common Elements [6]. Readers may find the current version of the document at www.atsc.org.

Access Unit (AU)	Self contained audio stream packet.
Associated Services	A companion audio stream that can be combined with the Main Program.
Binaural Audio	Process that reproduces 3D audio for headphones to provide the listener with an immersive experience.
Commentary	Audio program element assigned to voice/announcer information.
Convergence/ Divergence	For audio object, the amount of the 'spread' of the audio in acoustic space.
Core Decode	Minimal decode specification, usually limited to stereo or 5.1 audio programs.
Dialog Enhancement	Feature for the hearing challenged viewer or where there is high ambient noise to enhance the intelligibility of the dialog or commentary audio.
Downmixing	For Channel-based audio formats, the ability for the decoder to reproduce the higher order speaker arrangement to a lesser speaker arrangement (i.e. 5.1 to 2.0).
Full Decode	Decode specification that provides for full sound program reproduction.
ISO Base Media File Format	File format for media as defined by ISO/IEC 14496-12 [21].
Loudness Normalization	Process within the encoding algorithm that ensures consistent audio loudness across all renders, downmixes and presets.
Next Generation Audio	Immersive, interactive and personalized audio delivery system.
Parametric	Audio encoding method that uses side-information to reconstruct a voice channel.

Preselection	Set of Audio Program Components representing a version of the Audio Program that may be selected for simultaneous decoding. An Audio Preselection is a sub-set of available Audio Program Components of one Audio Program.
Random Access Point	An audio data packet that provides an entry into an audio stream.
Renderer	A part of an NGA receive device, post decoding, that merges the various sound program components into the available reproduction channels.
Side Information	Data associated with features such as Dialog Enhancement, including metadata.

Table 2 Common terms related to NGA codecs[8]

Term	Description
2.0	Nomenclature for stereo audio, with two audio channels (L, R), as found in legacy television audio systems.
5.1	Nomenclature for surround audio, with five full-range audio channels (L, C, R, LS, RS) and one low-frequency effects (LFE) channel, as found in the existing ATSC digital television audio system.
7.1+4	Nomenclature for a particular 11.1 loudspeaker arrangement suitable for Immersive Audio, consisting of three frontal loudspeakers (L, C, R) and four surround loudspeakers (left side [LS], left rear [LR], right side [RS], right rear [RR]) on the listener's plane, and four speakers placed above the listener's head height (arranged in LF, RF, LR and RR positions).
Audio Element	The smallest addressable unit of an *Audio Program*. Consists of one or more *Audio Signals* and associated *Audio Element Metadata,* and can be configured as any of three different *Audio Element Formats*. (See Figure 18)
Audio Element Format	Description of the configuration and type of an *Audio Element*. Notes: There are three different types of Audio Element Formats. Depending on the type, different kinds of properties are used to describe the configuration: Channel-based audio: e.g., the number of channels and the channel layout Object-based audio: e.g., dynamic positional information Scene-based audio: e.g., HOA order, number of transport channels
Audio Element Metadata	Metadata associated with an *Audio Element*. Notes: Some examples of Audio Element Metadata include positional metadata (spatial information describing the position of objects in the reproduction space, which may dynamically change over time, or channel assignments), or personalization metadata (set by content creator to enable certain personalization options such as turning an element "on" or "off," adjusting its position or gain, and setting limits within which such adjustments may be made by the user). (See Table 3 for alternate nomenclature used for this term in other documents.)
Audio Object	An *Audio Element* that consists of an *Audio Signal* and *Audio Element Metadata,* which includes rendering information (e.g., gain and position) that may dynamically change. Audio Objects with rendering information that does not dynamically change may be called "static objects."

[8] Table 2 is provided courtesy of the Advanced Television Systems Committee from Standard A/342 Part 1, Audio Common Elements [6]. Readers may find the current version of the document at www.atsc.org.

Audio Presentation	A set of *Audio Program Components* representing a version of the *Audio Program* that may be selected by a user for simultaneous decoding. Notes: An Audio Presentation is a sub-selection from all available *Audio Program Components* of one *Audio Program*. (See Figure 18) A Presentation can be considered the NGA equivalent of audio services in predecessor systems, which each utilized complete mixes (e.g., “SAP” or “VDS”) (See Table 3 for alternate nomenclature used for this term in other documents.)
Audio Program	The complete collection of all *Audio Program Components* and a set of accompanying *Audio Presentations* that are available for one Audio Program. (See Figure 18.) Notes: Not all *Audio Program Components* of one Audio Program are necessarily meant to be presented at the same time. An Audio Program may contain *Audio Program Components* that are always presented, and it may include optional *Audio Program Components*. (See Table 3 for alternate nomenclature used for this term in other documents.)
Audio Program Component	A logical group of *Audio Elements* that is used to define an *Audio Presentation* and may consist of one or more *Audio Elements.* (See Figure 18.) (See Table 3 for alternate nomenclature used for this term in other documents.)
Audio Program Component Type	Characterization of an *Audio Program Component* with regard to its content. Notes: Examples for Audio Program Component Types are: Complete Main Music & Effects (M&E): the background signal that contains a Mix of various Audio Signals except speech. Dialog: one or more Audio Signals that contain only speech Video Description Service
Audio Signal	A mono signal. (See Figure 18.)
Bed	An *Audio Element* that is intended to be used as the foundational element of an *Audio Presentation* (e.g., Music & Effects), to which other complementing Audio Elements (e.g., Dialog) are added.
Channel Set	A group of *Channel Signals* that are intended to be reproduced together.
Channel Signal	An *Audio Signal* that is intended to be played back at one specific nominal loudspeaker position.
Complete Mix	All *Audio Elements* of one *Audio Presentation* mixed together and presented as a single *Audio Program Component*.
Elementary Stream	A bit stream that consists of a single type of encoded data (audio, video, or other data). Notes: The *Audio Elements* of one *Audio Program* may be delivered in a single audio Elementary Stream or distributed over multiple audio Elementary Streams. (See Table 3 for alternate nomenclature used for this term in other documents.)
Higher-Order Ambisonics	A technique in which each produced signal channel is part of an overall description of the entire sound scene, independent of the number and locations of actually available loudspeakers.
Immersive Audio	An audio system that enables high spatial resolution in sound source localization in azimuth, elevation and distance, and provides an increased sense of sound envelopment.
LFE	Low-frequency effects channel. A limited frequency response channel that carries only low frequency (e.g., 100 Hz and below) audio.
Mix	A number of *Audio Elements* of one *Audio Program* that are mixed together into one *Channel Signal* or into a *Bed*.
Rendering	The realization of aural content for acoustical presentation.
Track	Representation of an *Elementary Stream* that is stored in a file format like the ISO Base Media File Format. Notes: For some systems, it may be possible to directly store the unmodified data from the *Elementary Stream* into a Track, whereas for other systems it may be necessary to re-format the data for storage in a Track.

7.2 Common Features of NGA

Complementing the visual enhancements that Ultra HD will bring to consumers, Next Generation Audio (NGA) provides a compelling new audio experiences:

- Immersive – An audio system that enables high spatial resolution in sound source localization in azimuth, elevation and distance, and provides an increased sense of sound envelopment
- Personalized – Enabling consumers to tailor and interact with their listening experience, e.g. selecting alternate audio experiences, alternate languages, dialogue enhancement
- Consistent – Playback experience automatically optimized for each consumer device, e.g. home and mobile
- Object-based Audio – Audio elements are programmed to provide sound from specific locations in space, irrespective of speaker location. By delivering audio as individual elements, or objects, content creators can simplify operations, reduce bandwidth, and provide a premium experience for every audience
- Scene-Based Audio – An arbitrarily large number of directional audio elements composing a 3D sound field are mixed in a fixed number of PCM signals according to the Higher-Order Ambisonics format. Once in the HOA format, the Audio Scene can be efficiently transmitted, manipulated, and rendered on loudspeaker layouts/headphones/soundbars.
- Flexible Delivery - NGA can be delivered to consumers over a number of different distribution platforms including terrestrial, cable, and satellite broadcast, IPTV, OTT, and mobile. It could also be delivered over a hybrid of broadcast and OTT
- Flexible Rendering - NGA can be experienced by consumers through headphones or speakers (e.g., TV speakers, home theater systems including ceiling speakers, sound bars) as shown in Figure 17

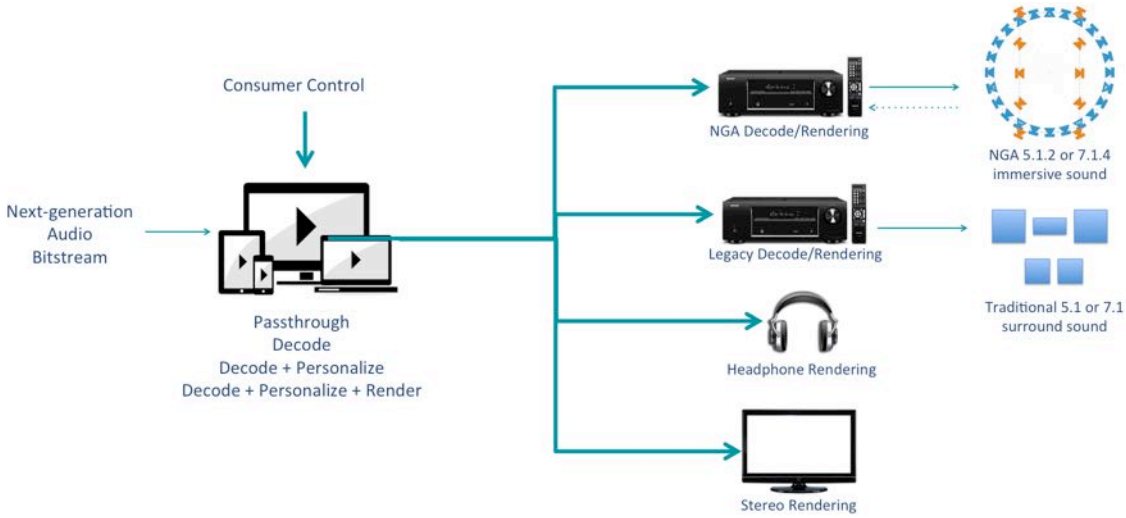

Figure 17 NGA in the consumer domain

NGA improves current use cases by supporting conventional Channel-based Audio (CBA) at lower data rates than were previously possible, which is preferred by next-generation broadcast and streaming services, along with a number of system level advancements over existing solutions. For example, these enhancements allow improved accessibility solutions compared with current broadcast systems.

7.2.1 NGA Use Cases

- **Home Theater --** Consumers can experience audio coming from overhead as well as around them while listening through speakers (including soundbars) at home
- **Headphones (Mobile) –** Consumers can experience audio through headphones, giving them a richer more immersive experience in constrained listening environments
- **Language Selection -** Consumers can select their language preference from many more choices than they have had in the past and enjoy audio program without compromise
- **Personalization -- Sports fans** are able to use interactive features to select their team announcer, which crowd noises they prefer, or maybe even add in the overhead public address feed so they feel like they are at the game
- **Accessibility Features**
    - **Visually impaired users** can select a descriptive audio track while enjoying television to be added to the main dialog (voice over) to better understand what is happening on-screen
    - **Hearing impaired users** may choose to use dialog enhancement as well as the ability to control the volume of the dialog independently of other sounds to improve their listening experience
- **Dialogue Enhancement -** Viewers can 'boost' specific elements of an audio program like dialog or the ambient sound when listening in high noise environments (e.g., train stations, crowds, airports) to better understand what's happening in a piece of audio content, or can reduce dynamic range in the evening to avoid disturbing others

7.2.2 Audio Program Components and Preselections

Audio Program Components are separate pieces of audio data that are combined to compose an Audio Preselection. A simple Audio Preselection may consist of a single Audio Program Component, such as a Complete Main Mix for a television program. Audio Preselections that are more complex may consist of several Audio Program Components, such as ambient music and effects, combined with dialog and video description. For example, a complete audio with English dialog, a complete audio with Spanish dialog, a complete audio (English or Spanish) with video description, or a complete audio with alternate dialog may all be selectable Preselections for a Program.

NGA systems enable user control of certain aspects of the Audio Scene (e.g., adjusting the relative level of dialogue with respect to the ambient music and effects) by combining the Audio Program Components, present in one or more NGA streams, at the receiver side in user-selectable modes. In this way several Audio Program Components can be shared between different Audio Preselections, allowing more efficient delivery of additional services compared to legacy broadcast systems. For example, the same music & effects component can be used with a Spanish and an

English dialog component, whereas a legacy broadcast would need to send two complete mixes, both including music and effects. This is a major advantage of the NGA systems, where one stream contains more than one complete audio main, or multiple streams contain pieces of a complete audio main.

7.2.3 Carriage of NGA

Audio Program Components corresponding to one or more Audio Preselections can be delivered in a single elementary stream (i.e., NGA single-stream delivery) or in multiple elementary streams (i.e., NGA multi-stream delivery).

In case of single-stream delivery, all Audio Program Components of one Audio Program are carried in a single NGA stream, together with the signaling information of the available Audio Preselections. The method of doing this is codec-specific, but in general, the different component streams are multiplexed into one single stream along with appropriate signaling information.

In the case of multi-stream delivery, the Audio Program Components of one Audio Program are not carried within one single NGA stream, but in two or more NGA streams, the main NGA stream contains at least all the Audio Program Components corresponding to one Audio Preselection. The auxiliary streams may contain additional Audio Program Components (e.g., additional language tracks). The multi-stream delivery also allows a hybrid distribution approach where one stream is delivered via OTA and another via OTT.

7.2.4 Metadata

NGA codec systems have a rich set of audio metadata features and functions. Each codec has its own set of definitions; however, there is a common framework for audio metadata developed by the EBU called the Audio Definition Model (ADM) [13].

In general, there are three types of audio metadata:

1. Descriptive - Provides information regarding the available audio program features (i.e., Channel configuration, alternate languages, VDS).
2. Functional - Provides information regarding how the audio should be rendered or presented (i.e., preselections, object audio locations, loudness controls, downmix coefficients).
3. Control - Allows for personalization and user preferences (e.g. Dialog enhancement, language preference, program preselection)

7.2.5 Overview of Immersive Program Metadata and Rendering

Immersive programming requires generating and delivering dynamic metadata to playback devices. For immersive programming, object position and rendering control metadata are essential for enabling the optimum set of experiences regardless of playback device or application. This section provides an overview of these important metadata parameters and how they are utilized during the creative process.

An important consideration for a spatial (immersive) audio description model supporting audio objects is the choice of the spatial frame of reference. This will be utilized by the core Object-based Audio rendering algorithm (in playback devices) to map the source audio objects to the

active speaker configuration/layout based on the positional metadata generated upstream in production.

In many cases (e.g. psychoacoustic research) sound source locations in 3-dimensional space can be represented with an *egocentric* model, where the listening position is the point of origin and the sound location expressed relative to this point (e.g. using azimuth and elevation angles). If used for sound scene description, this suggests that preserving the relative direction of incidence of a particular sound at the listening point should be a primary objective of the audio rendering algorithm and therefore is generally associated with direction-based rendering algorithms.

However, A/V production sound mixers do not generally author spatial content relative to the listening position but rather position the sound elements (object) in the room, relative to the action on the screen, known as an *allocentric* model. The ultimate goal is not necessarily to position the sound object consistently at the same direction for each seat, but to ensure that the perceived direction at each seat is consistent with the position of the sound element (object) in the room. Thus, it is more natural for mixers to author spatial audio content in terms of the balance of left/right, front/back and up/down position relative to the screen or room rather than in terms of the direction relative to their own listening position. In addition, the use of an allocentric frame of reference for sound source location ensures consistency between object- and Channel-based Audio elements because both the channels and objects are referenced to the listening environment.

Object position is therefore defined as an abstracted (unit) room where each object(s) 3-dimensional coordinates, (x,y,z) in [-1,1] × [-1,1] × [-1,1], correspond to the traditional balance controls found in mixing consoles (left/right, front/back and by extension to 3D bottom/top).

7.2.6 Audio Element Formats

The information contained in this section 7.2.6 is provided courtesy of the Advanced Television Systems Committee from Standard A/342 Part 1, Audio Common Elements. Readers may find the current version of the document at www.atsc.org.

The NGA systems support three fundamental Audio Element Formats (see also Table 2):

1. Channel Sets are sets of Audio Elements consisting of one or more Audio Signals presenting sound to speaker(s) located at canonical positions. These include configurations such as mono, stereo, or 5.1, and extend to include non-planar configurations, such as 7.1+4.
2. Audio Objects are Audio Elements consisting of audio information and associated metadata representing a sound's location in space (as described by the metadata). The metadata may be dynamic, representing the movement of the sound.
3. Scene-based audio (e.g., HOA) consists of one or more Audio Elements that make up a generalized representation of a sound field.

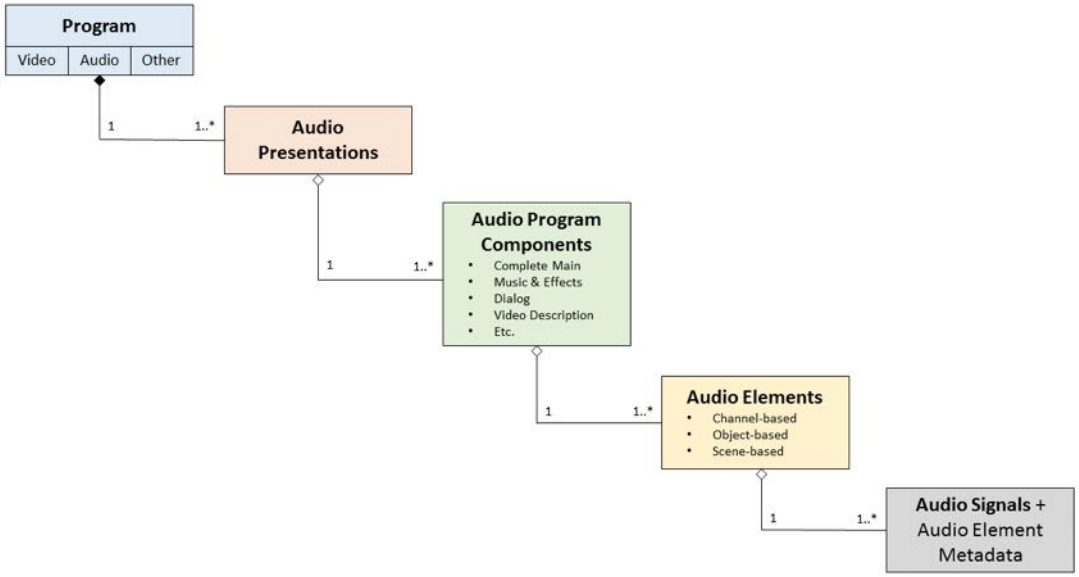

Figure 18 Relationship of key audio terms

Table 3 Mapping of terminology between NGA technologies

Common Term	DASH-IF Term [11]	AC-4 Term [7]	MPEG-H Audio Term [8]
Audio Element Metadata		Metadata, Object Audio Metadata	Metadata Audio Elements (MAE), Object Metadata (OAM)
Audio Presentation	Preselection	Presentation	Preset
Audio Program	Bundle	Audio Program	Audio Scene
Audio Program Component	Referred to as Audio Element	Audio Program Component	Group
Elementary Stream	Representation in an Adaptation Set	Elementary Stream	Elementary Stream

7.2.7 Audio Rendering

Audio Rendering is the process of composing an Audio Preselection and converting all the Audio Program Components to a data structure appropriate for the audio outputs of a specific receiver. Rendering may include conversion of a Channel Set to a different channel configuration, conversion of Audio Objects to Channel Sets, conversion of Scene-based sets to Channel Sets, and/or applying specialized audio processing such as room correction or spatial virtualization. In addition, the application of Dialog Enhancement as well as Loudness Normalization are parts of the audio rendering functionality.

7.2.7.1 Video Description Service (VDS)

Video Description Service is an audio service carrying narration describing a television program's key visual elements. These descriptions are inserted into natural pauses in the program's dialog. Video description makes TV programming more accessible to individuals who are blind or visually impaired. The Video Description Service may be provided by sending a collection of “Music and Effects” components, a Dialog component, and an appropriately labeled Video Description

component, which are mixed at the receiver. Alternatively, a Video Description Service may be provided as a single component that is a Complete Mix, with the appropriate label identification.

7.2.7.2 Multi-Language

Traditionally, multi-language support is achieved by sending Complete Mixes with different dialog languages. For NGA systems, multi-language support can be achieved through a collection of "Music and Effects" streams combined with multiple dialog language streams that are mixed at the receiver.

7.2.7.3 Personalized Audio

Personalized audio consists of one or more Audio Elements with metadata, which describes how to decode, render, and output "full" Mixes. Each personalized Audio Preselection may consist of an ambience "bed", one or more dialog elements, and optionally one or more effects elements. Multiple Audio Preselections can be defined to support a number of options such as alternate language, dialog or ambience, enabling height elements, etc.

There are two main concepts of personalized audio:

1. Personalization selection – The bitstream may contain more than one Audio Preselection where each Audio Preselection contains pre-defined audio experiences (e.g., "home team" audio experience, multiple languages, etc.). A listener can choose the audio experience by selecting one of the Audio Preselections.
2. Personalization control – Listeners can modify properties of the complete audio experience or parts of it (e.g., increasing the volume level of an Audio Element, changing the position of an Audio Element, etc.).

7.3 MPEG-H Audio

7.3.1 Introduction

MPEG-H Audio is a Next Generation Audio (NGA) system offering true immersive sound and advanced user interactivity features. Its object-based concept of delivering separate audio elements with metadata within one audio stream enables personalization and universal delivery. MPEG-H Audio is an open international ISO standard and standardized in ISO/IEC 23008-3 [23]. The MPEG-H 3D Audio Low Complexity Profile Level 3 is adopted by DVB in ETSI TS 101 154 v.2.3.1 [15] and is one of the audio systems standardized for use in ATSC 3.0 Systems as defined in A/342 Part 3 [8]. SCTE has included the MPEG-H Audio System into the suite of NGA standards for cable applications as specified in SCTE 242-3 [32].

The MPEG-H Audio system was selected by the Telecommunications Technology Association (TTA) in South Korea as the sole audio codec for the terrestrial UHDTV broadcasting specification TTAK.KO- 07.0127 [45] that is based on ATSC 3.0. On May 31, 2017, South Korea launched its 4K UHD TV service using the MPEG-H Audio system.

As shown in Figure 19, MPEG-H Audio can carry any combination of Channels, Objects and Higher-Order Ambisonics (HOA) signals in an efficient way, together with the metadata required for rendering, advanced loudness control, personalization and interactivity.

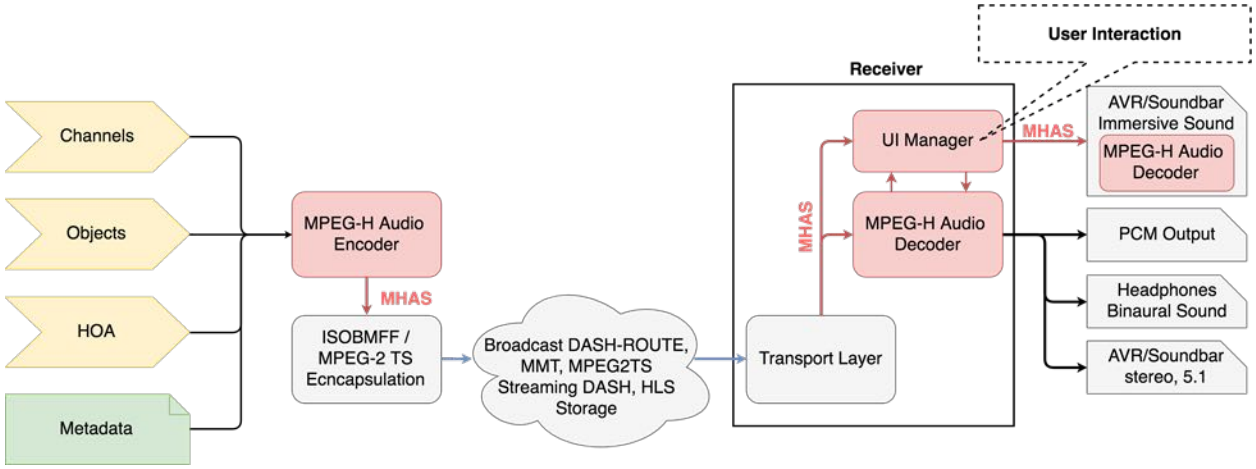

Figure 19 MPEG-H Audio system overview

The MPEG-H Audio Stream (MHAS), described in 7.3.3, contains the audio bitstream and various types of metadata packets and represents common layer for encapsulation into any transport layer format (e.g., MPEG-2 TS, ISOBMFF). The MPEG-H Audio enabled receiver can decode and render the audio to any loudspeaker configuration or a Binaural Audio representation for headphones reproduction. For enabling the advanced user interactivity features in cases where external playback devices are used, the UI Manger can supply the user interactions by inserting new MHAS packets into the MHAS stream and further deliver this over HDMI to the subsequent immersive AVR/Soundbar with MPEG-H Audio decoding capabilities. This is described in more detail in 7.3.1.4.

All MPEG-H Audio features that are described in the following sections are supported by the MPEG-H 3D Audio Low Complexity Profile Level 3 and are thus available in all broadcast systems based on the DVB and ATSC 3.0 specifications. See Table 4 for the characteristics of the Low Complexity Profile and levels.

Table 4 Levels for the Low Complexity Profile of MPEG-H Audio

Profile Level	1	2	3	4	5
Max Sample Rate (kHz)	48	48	48	48	96
Max Core Codec Channels in Bit Stream	10	18	32	56	56
Max Simultaneous decoded core codec channels	5	9	16	28	28
Max Loudspeaker outputs	2	8	12	24	24
Example loudspeaker configurations	2	7.1	7.1 + 4H	22.2	22.2
Max Decoded Objects	5	9	16	28	28

7.3.1.1 Personalization and Interactivity

MPEG-H Audio enables viewers to interact with the content and personalize it to their preference. The MPEG-H Audio metadata carries all the information needed for personalization such as attenuating or increasing the level of objects, disabling them, or changing their position. The

metadata also contains information to control and restrict the personalization options such as setting the limits in which the user can interact with the content, as illustrated in Figure 20. (See also section 7.4.3 MPEG-H Audio Metadata.)

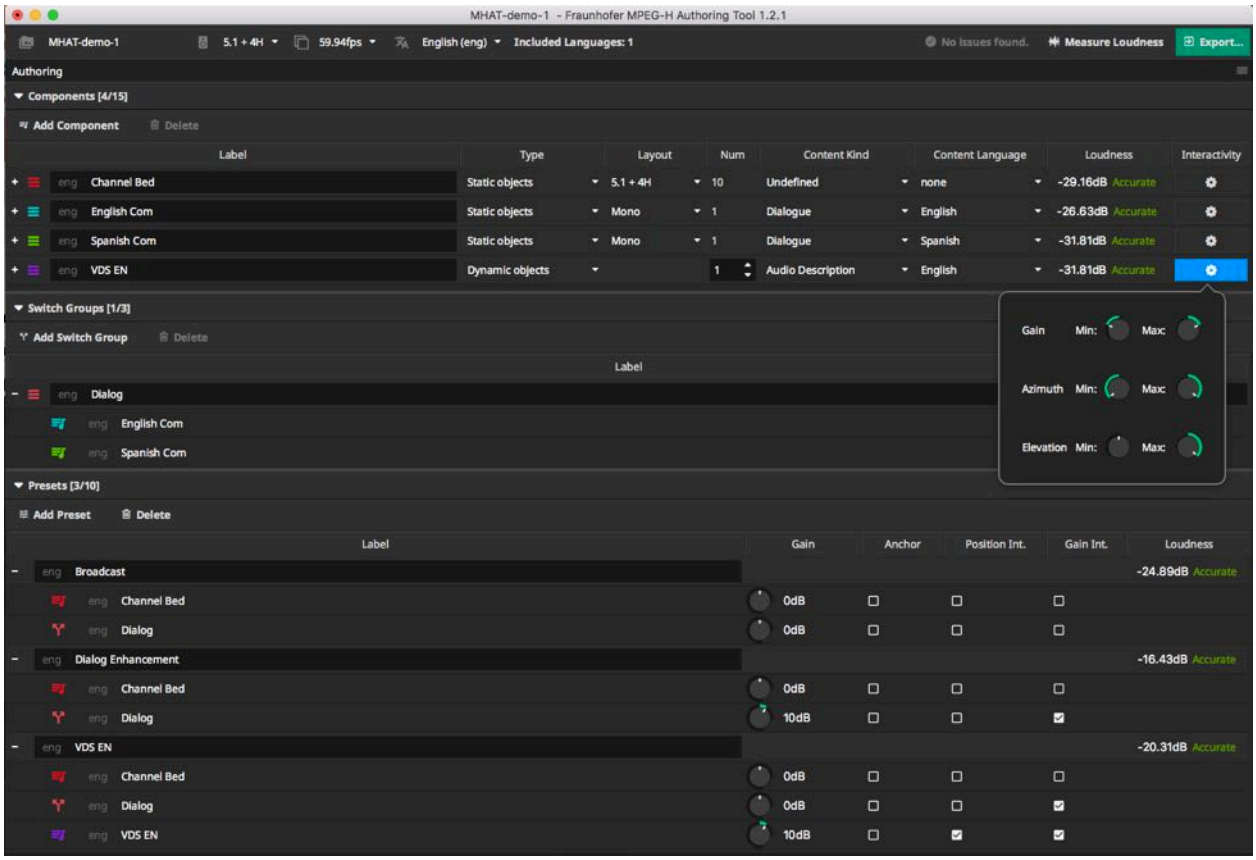

Figure 20 MPEG-H Authoring Tool example session

7.3.1.2 Universal delivery

MPEG-H Audio provides a complete integrated audio solution for delivering the best possible audio experience, independently of the final reproduction system. It includes rendering and downmixing functionality, together with advanced Loudness and Dynamic Range Control (DRC).

The loudness normalization module ensures consistent loudness across programs and channels, for different presets and playback configurations, based on loudness information embedded in the MPEG-H Audio stream. Providing loudness information for each preset allows for instantaneous and automated loudness normalization when the user switches between different presets. Additionally, downmix-specific loudness information can be provided for artist-controlled downmixes.

7.3.1.3Immersive Sound

MPEG-H Audio provides Immersive sound (i.e., the sound can come from all directions, including above or below the listener's head), using any combination of the three well-established audio formats: Channel-based, Object-based, and Higher-Order Ambisonics (Scene-Based Audio).

The MPEG-H 3D Audio Low Complexity Profile Level 3 allows up to 16 audio elements (channels, objects or HOA signals) to be decoded simultaneously, while up to 32 audio elements can be carried simultaneously in one stream (see Table 4).

7.3.1.4 Distributed User Interface Processing

In order to take advantage of the advanced interactivity options, MPEG-H Audio enabled devices require User Interfaces (UIs). In typical home set-ups, the available devices are connected in various configurations such as:

- a Set-Top Box connecting to a TV over HDMI
- a TV connecting to an AVR/Soundbar over HDMI or S/PDIF

In all cases, it is desired to have the user interface located on the preferred device (i.e., the source device).

For such use cases, the MPEG-H Audio system provides a unique way to separate the user interactivity processing from the decoding step. Therefore, all user interaction tasks are handled by the "UI Manager", in the source device, while the decoding is done in the sink device. This feature is enabled by the packetized structure of the MPEG-H Audio Stream, which allows for:

- easy stream parsing on system level
- insertion of new MHAS packets on the fly (e.g., "USERINTERACTION" packets).

Figure 21 provides a high-level block-diagram of such a distributed system between a source and a sink device connected over HDMI. The detached UI Manager has to parse only the MHAS packets containing the Audio Scene Information and provides this information to an UI Renderer to be displayed to the user. The UI Renderer is responsible for handling the user interactivity and passes the information about every user's action to the detached UI Manager, which embeds it into MHAS packets of type USERINTERACTION and inserts them into the MHAS stream.

The MHAS stream containing the USERINTERACTION packets is delivered over HDMI to the sink device which decodes the MHAS stream, including the information about the user interaction, and renders the Audio Scene accordingly.

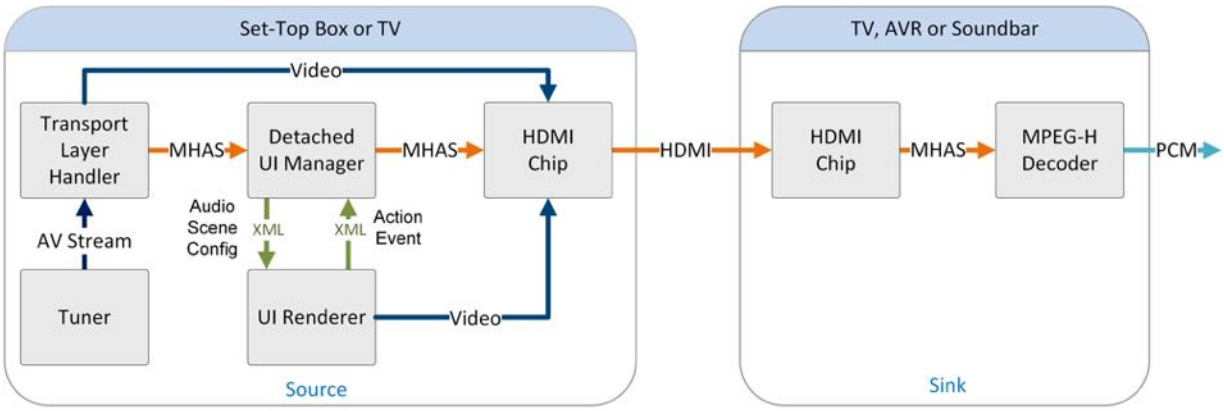

Figure 21 Distributed UI processing with transmission of user commands over HDMI

The USERINTERACTION packet provides an interface for all allowed types of user interaction. Two interaction modes are defined in the interface.

- An advanced interaction mode – where the interaction can be signaled for each element group that is present in the Audio Scene. This mode enables the user to freely choose which groups to play back and to interact with all of them (within the restrictions of allowances and ranges defined in the metadata and the restrictions of switch group definitions).

- A basic interaction mode – where the user may choose one preset out of the available presets that are defined in the metadata audio element syntax.

7.3.2 MPEG-H Audio Metadata

MPEG-H Audio enables NGA features such as personalization and interactivity with a set of static metadata, the “Metadata Audio Elements” (MAE). Audio Objects are associated with metadata that contain all information necessary for personalization, interactive reproduction, and rendering in flexible reproduction layouts. This metadata is part of the overall set-up and configuration information for each piece of content.

7.3.2.1 Metadata Structure

The metadata (MAE) is structured in several hierarchy levels. The top-level element is the Audio Scene Information or the "AudioSceneInfo" structure as shown in Figure 22. Sub-structures of the AudioSceneInfo contain descriptive information about "Groups", "Switch Groups", and "Presets." An "ID" field uniquely identifies each group, switch group or preset, and is included in each sub-structure.

The group structures ("mae_GroupDefinition") contain descriptive information about the audio elements, such as:

- the group type (channels, objects or HOA),
- the content type (e.g., dialog, music, effects, etc.),
- the language for dialogue objects, or
- the channel layout in case of Channel-based content.

User interactivity can be enabled for the gain level or position of objects, including restrictions on the range of interaction (i.e., setting minimum and maximum values for gain and position offset). The minimum and maximum values can be set differently for each group.

Groups can be combined into switch groups ("mae_SwitchGroupDefinition"). All members of one switch group are mutually exclusive, i.e., during playback, only one member of the switch group can be active or selected. As an example, using a switch group for dialog objects ensures that only one out of multiple dialog objects with different languages is played back at the same time. Additionally, one member of the switch group is always marked as default to be used if there is no user preference setting and to make sure that the content is always played back with dialog, for example.

The preset structures ("mae_GroupPresetData") can be used to define different "packages" of audio elements within the Audio Scene. It is not necessary to include all groups in every preset definition. Groups can be "on" or "off" by default and can have a default gain value. Describing only a sub-set of groups in a preset is allowed. The audio elements that are packaged into a preset are mixed together in the decoder, based on the metadata associated with the preset, and the group and switch group metadata.

From a user experience perspective, the presets behave as different complete mixes from which users can choose. The presets are based on the same set of audio elements in one Audio Scene and thus can share certain audio objects/elements, like a channel-bed. This results in bitrate savings compared to a simulcast of a number of dedicated complete mixes.

Textual descriptions ("labels") can be associated with groups, switch groups and presets, for instance "Commentary" in the example below for a switch group. Those labels can be used to enable personalization in receiving devices with a user interface.

7.3.2.2 Metadata Example

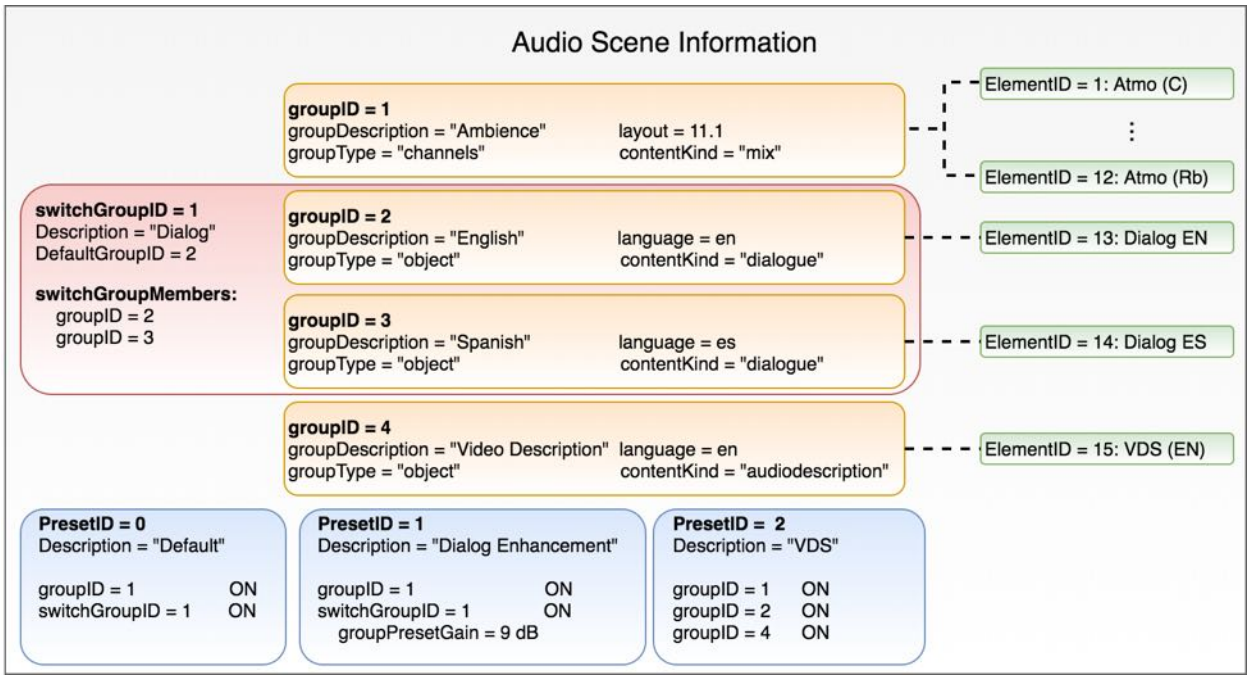

Figure 22 Example of an MPEG-H Audio Scene information

Figure 22 contains an example of MPEG-H Audio Scene Information with four different groups (orange), one switch group (red) and three presets (blue). In this example, the switch group contains two dialogs in different languages that the user can choose from, or the device can automatically select one dialog based on the preference settings.

The "Default" preset ("PresetID = 0") for this Audio Scene contains the "Ambience" group ("groupID = 1") and the "Dialog" switch group ("switchGroupID = 1") wherein the English dialog ("groupID = 2") is the default. Both the ambience group and the dialog switch group are active ("ON"). This preset is automatically selected in the absence of any user or device automatic selection. The additional two presets in this example enable the advanced accessibility features as described in the following sub-sections.

The "Dialog Enhancement" preset contains the same elements as the default preset, with the same status ("ON") with the addition that the dialog element (i.e., the switch group) is rendered with a 9 dB gain into the final mix. The gain parameter, determined by the content author, can be any value from -63 to +31 in 1 dB steps.

The "VDS" preset contains three groups, all active: the ambience ("groupID = 1"), the English dialog ("groupID = 2") and the Video Description ("groupID = 4").

The "VDS" preset can be manually selected by the user or automatically selected by the device based on the preference settings (i.e., if Video Description Service is enabled in the device's settings).

7.3.2.3 Personalization Use Case Examples

Advanced Accessibility

Object-based audio delivery with MPEG-H Audio together with the MPEG-H Audio Metadata offer advanced and improved accessibility services, especially:

- Video Descriptive Services (VDS, also known as Audio Description) and
- Dialog Enhancement (DE).

As described in the previous section, the dialog elements and the Video Description are carried as separate audio objects ("groups") that can be combined with a channel bed element in different ways and create different presets, such as a "default" preset without Video Description and a "VDS" preset.

Additionally, MPEG-H Audio allows the user to spatially move the Video Description object to a user selected position (e.g., to the left or right), enabling a spatial separation of main dialog and the Video Description element, as shown in Figure 23. This results in a better intelligibility of the main dialog as well as the Video Description (e.g., in a typical 5.1 set-up the main dialog is assigned to the center speaker while the Video Description object could be assigned to a rear-surround speaker).

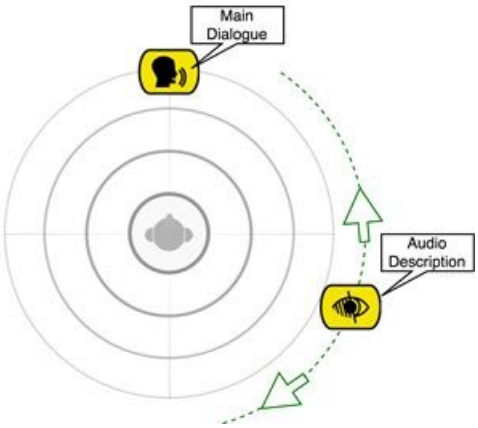

Figure 23 Audio description re-positioning example

Dialog Enhancement (DE)

MPEG-H Audio includes a feature of DE that enables automatic device selection (prioritization) as well as user manipulation. For ease of user selection or for automatic device selection (e.g., enabling TV "Hard of Hearing" TV setting), a Dialog Enhancement preset can be created, as illustrated in Figure 22 using a broadcaster defined enhancement level for the dialog element (e.g., 10dB as shown in Figure 20).

Moreover, if the broadcaster allows personalization of the enhancement level, MPEG-H Audio supports advanced DE which enables direct adjustment of the enhancement level via the user interface. The enhancement limitations (i.e., maximum level) are defined by the broadcaster/content creator as shown in Figure 20 and carried in the metadata. This maximum value for the lower and upper end of the scale can be set differently for different elements as well as for different content.

The advanced loudness management tool of the MPEG-H Audio system automatically compensates loudness changes that result from user interaction (e.g., switching presets or enhancement of dialogue) to keep the overall loudness on the same level, as illustrated in Figure 24. This ensures constant loudness level not only across programs but also after user interactions.

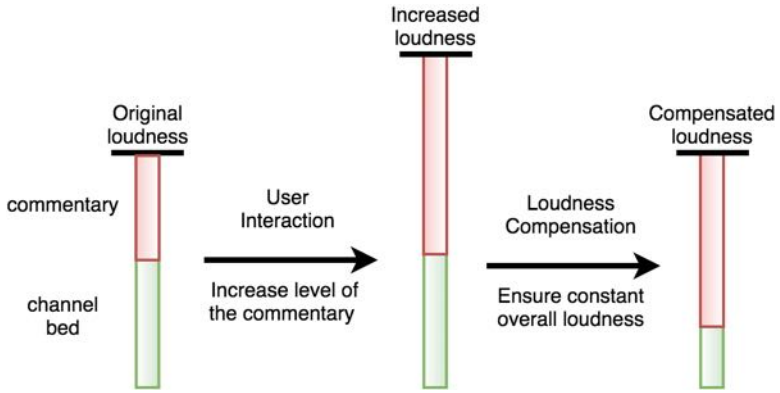

Figure 24 Loudness compensation after user interaction

Multi-language services

With a common channel bed and individual audio objects for dialog in different languages as well as for Video Description MPEG-H Audio results in more efficient broadcast delivery than non-NGA audio codecs in which common components must be duplicated to create multiple complete mixes.

Furthermore, all features (e.g., VDS and DE in several languages) can be enabled in a single audio stream, simplifying the required signalling and selection process on the receiver side.

Personalization for Sport Programs

For various program types, such as sport programs, MPEG-H Audio provides additional advanced interactivity and personalization options, such as choosing between 'home team' and 'away team' commentaries of the same game, listening to the team radio communication between the driver and his team during a car race, or listening only to the crowd (or home/away crowd) with no commentary during a sports program.

7.3.3 MPEG-H Audio Stream

The MPEG-H Audio Stream (MHAS) format is a self-contained, packetized, and extensible byte stream format to carry MPEG-H Audio data. The basic principle of the MHAS format is to separate encapsulation of coded audio data, configuration data and any additional metadata or control data into different MHAS packets. Therefore, it is easier to access configuration data or other metadata on the MHAS stream level without the need to parse the audio bitstream.

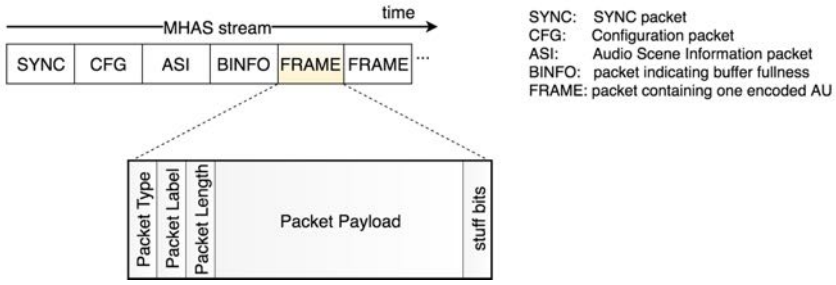

Figure 25 MHAS packet structure

Figure 25 shows the high-level structure of an MHAS packet, which contains the header with the packet type to identify each MHAS packet, a packet label and length information, followed by the payload and potential stuffing bits for byte alignment.

The packet label has the purpose of differentiating between packets that belong either to different configurations in the same stream, or different streams in a multi-stream environment.

7.3.3.1 Random Access Point

A Random Access Point (RAP) consists of all MHAS packets that are necessary to tune to a stream and enable start-up decoding: a potential sync packet, configuration data and an independently decodable audio data frame.

If the MHAS stream is encapsulated into an MPEG-2 Transport Stream, the RAP also needs to include a sync packet. For ISOBMFF encapsulation, the sync packet is not necessary, because the ISO file format structure provides external framing of file format samples.

The configuration data is necessary to initialize the decoder, and consists of two separate packets, the audio configuration data and the Audio Scene information metadata.

The encoded data frame of a RAP has to contain an “Immediate Playout Frame” (IPF), i.e., an Access Unit (AU) that is independent from all previous AUs. It additionally carries the previous AU’s information, which is required by the decoder to compensate for its start-up delay. This information is embedded into the Audio Pre-Roll extension of the IPF and enables valid decoded PCM output equivalent to the AU at the time instance of the RAP.

7.3.3.2 Configuration Changes and A/V Alignment

When the content set-up or the Audio Scene Information changes (e.g., the channel layout or the number of objects changes), a configuration change can be used in an audio stream for signalling the change and ensure seamless switching in the receiver.

Usually, these configuration changes happen at program boundaries (e.g., corresponding to ad insertion), but may also occur within a program. The MHAS stream allows for seamless configuration changes at each RAP.

Audio and video streams usually use different frame rates for better encoding efficiency, which leads to streams that have different frame boundaries for audio and video. Some applications may require that audio and video streams are aligned at certain instances of time to enable stream splicing.

MPEG-H Audio enables sample-accurate configuration changes and stream splicing using a mechanism for truncating the audio frames before and after the splice point. This is signaled on MHAS level through the AUDIOTRUNCATION packet.

An AUDIOTRUNCATION packet, indicating that the truncation should not be applied, can be inserted at the time when the stream is generated. The truncation can be easily enabled at a later stage on a systems level.

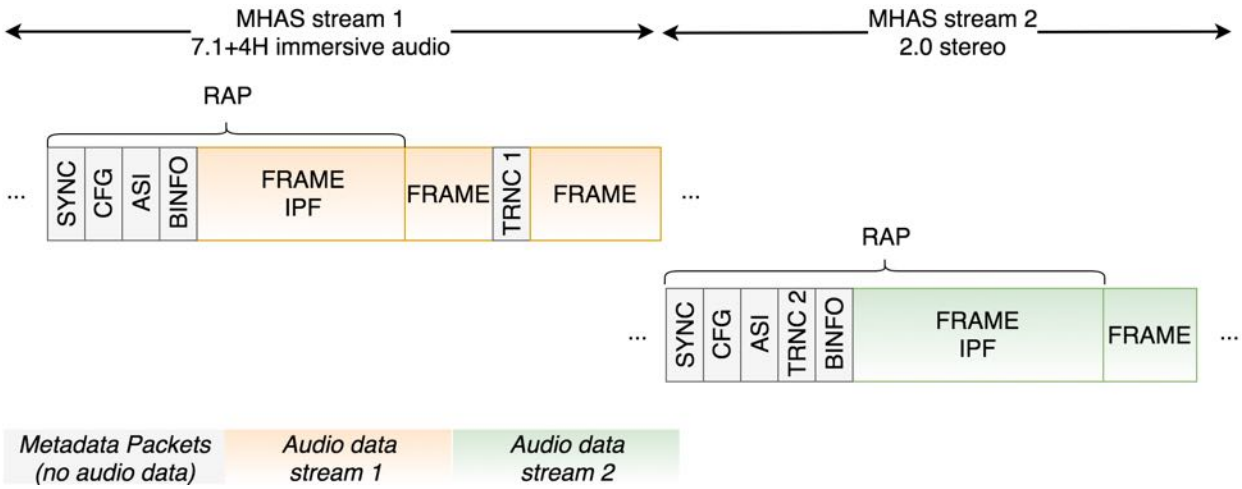

Figure 26 Example of a configuration change from 7.1+4H to 2.0 in the MHAS stream

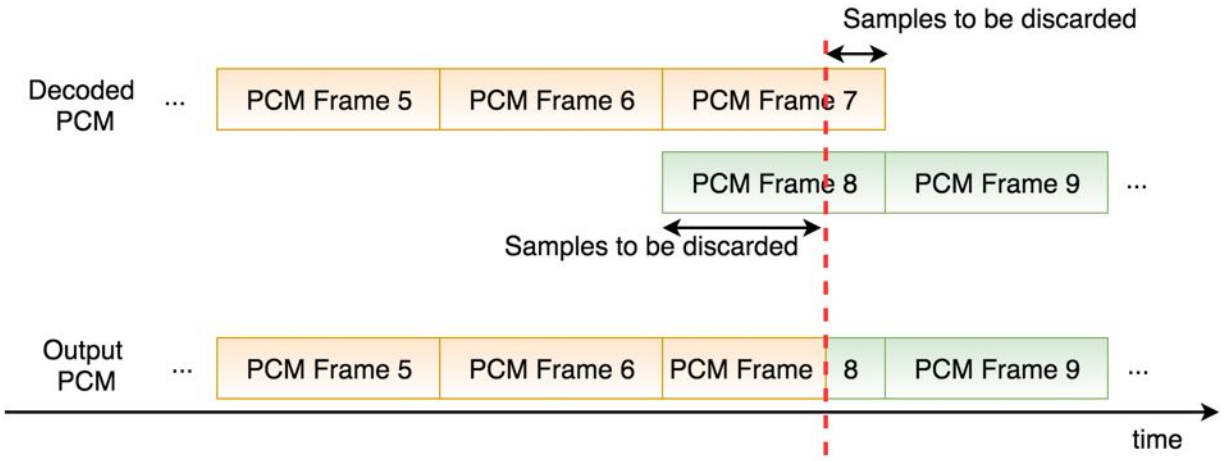

Figure 27 Example of a configuration change from 7.1+4H to 2.0 at the system output

Figure 26 and Figure 27 show an example of a sample-accurate configuration change from an immersive audio set-up to stereo inside one MHAS stream. (I.e., in the ad-insertion use case the inserted ad is stereo, while the rest of the program is in 7.1+4H.)

The first AUDIOTRUNCATION packet ("TRNC 1") contained in the first stream indicates how many samples are to be discarded at the end of the last frame of the immersive audio signal, while the second AUDIOTRUNCATION packet (“TRNC 2) in the second stream indicates the number of audio samples to be discarded at the beginning of the first frame of the new immersive audio signal.

7.4 Dolby AC-4 Audio

AC-4 is a audio system from Dolby Laboratories, which brings a number of features beyond those already delivered by the previous generations of audio technologies, including Dolby Digital®

(AC-3) and Dolby Digital® Plus (EAC-3). Dolby AC-4 is designed to address the current and future needs of next-generation video and audio entertainment services, including broadcast and Internet streaming.

The core elements of Dolby AC-4 have been standardized with the European Telecommunications Standards Institute (ETSI) as TS 103 190 [17] and adopted by Digital Video Broadcasting (DVB) in TS 101 154 [15] and are ready for implementation in next generation services and specifications. AC-4 is one of the audio systems standardized for use in ATSC 3.0 Systems [7]. AC-4 is specified in the ATSC 3.0 next-generation broadcast standard (A/342 [6]) and has been selected for use in North America (U.S., Canada and Mexico) as described in A/300 [3].

Furthermore, Dolby AC-4 enables experiences by fully supporting Object-based Audio (OBA), creating significant opportunities to enhance the audio experience, including immersive audio and advanced personalization of the user experience. As shown in Figure 28, AC-4 can carry conventional Channel-based soundtracks as well as Object-based mixes. Whatever the source type, the decoder renders and optimizes the soundtrack to suit the playback device.

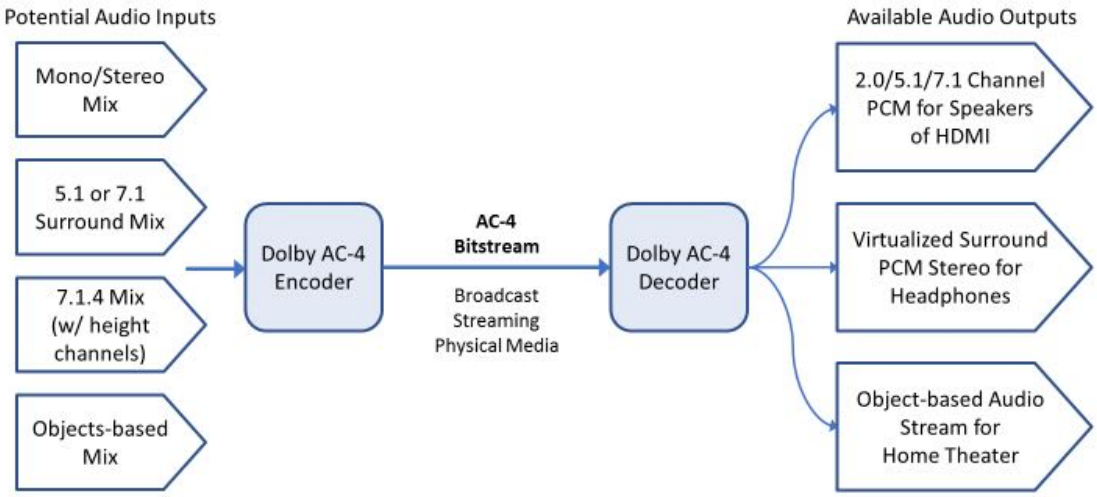

Figure 28 AC-4 Audio system chain

The AC-4 bitstream can carry Channel-based Audio, audio objects, or a combination of the two. The AC-4 decoder combines these audio elements as required to output the most appropriate signals for the consumer—for example, stereo pulse-code modulation (PCM) for speakers or headphones or stereo/5.1 PCM over HDMI. When the decoder is feeding a device with an advanced AC-4 renderer—for example, a set-top box feeding a Dolby Atmos® A/V receiver (AVR) in a home theater—the decoded audio objects can be sent to the AVR to perform sophisticated rendering optimized for the listening configuration.

Key features of the AC-4 audio system include:

1. **Core vs. Full Decode** and the concept of flexible **Input and Output Stages** in the decoder: The syntax and tools are defined in a manner that supports decoder complexity scalability. These aspects of the AC-4 coding system ensure that all devices, across multiple device categories, can decode and render the audio cost-effectively. It is important to note that the core decode mode does not discard any audio from the full decode but optimizes complexity for lower spatial resolution such as for stereo or 5.1 playback.
2. **Sampling Rate Scalable Decoding**: For high sampling rates (i.e., 96 kHz and 192 kHz), the decoder is able to decode just the 48kHz portion of the signal, providing decoded audio at a 48kHz sample rate without having to decode the full bandwidth audio track and downsampling. This reduces the complexity burden of having to decode the high sampling rate portion.
3. **Bitstream Splicing**: The AC-4 system is further designed to handle splices in bitstreams without audible glitches at splice boundaries, both for splices occurring at an expected point in a stream (controlled splice; for example on program boundaries), as well as for splices occurring in a non-predictable manner (random splice; for example when switching channels).
4. **Support for Separated Elements:** The AC-4 system offers increased efficiency not only from the traditional bits/channel perspective, but also by allowing for the separation of elements in the delivered audio. As such, use cases like multiple language delivery can be efficiently supported, by combining an M&E (Music and Effects) with different dialog tracks, as opposed to sending several complete mixes in parallel.
5. **Video Frame Synchronous Coding:** AC-4 supports a feature of video frame synchronous operation. This simplifies downstream splices, such as ad insertions, by using simple frame synchronization instead of, for example, decoding/re-encoding. The supported video frame synchronous frame rates are: 24 Hz, 30 Hz, 48 Hz, 60 Hz, 120 Hz, and 1000/1001 multiplied by those, as well as 25 Hz, 50 Hz, and 100 Hz.
   AC-4 also supports seamless switching of frame rates which are multiples of a common base frame rate. For example, a decoder can switch seamlessly from 25 Hz to 50 Hz or 100 Hz. A video random access point (e.g., an I-frame) is not needed at the switching point in order to utilize this feature of AC-4.
6. **Dialog Enhancement:** One important feature of AC-4 is Dialog Enhancement (DE) that enables the consumer/user to adjust the relative level of the dialogue to their preference. The amount of enhancement can be chosen on the playback side, while the maximum allowed amount can be controlled by the content producer.
   Dialogue Enhancement (DE) is an end-to-end feature, and the relevant bitrate of the DE metadata scales with the flexibility of the main audio information, from very efficient parametric DE modes up to modes where dialogue is transmitted in a self-contained manner, part of a so-called Music & Effects plus Dialog (M&E+D) presentation. Table 5 demonstrates DE modes and corresponding metadata information bitrates when dialogue is active, and the long-term average bitrate when dialog is active in only 50% of the frames.

Table 5 DE modes and metadata bitrates

DE mode	Typical bitrate during active dialog [kb/s]	Typical bitrate across a program (assumes 50% dialog) [kb/s]
Parametric	0.75 – 2.5	0.4 – 1.3
Hybrid	8 – 12	4.7 – 6.7
M&E+D	24 – 64	13 – 33

7.4.1 Dynamic Range Control (DRC) and Loudness

Loudness management in AC-4 includes a novel end-to-end signaling framework along with a real-time adaptive loudness processing mechanism that provide the service provider with an intelligent and automated system that ensures the highest quality audio while remaining compliant with regulations anywhere in the world. Compliant programming delivered to an AC-4 encoder with valid metadata will be encoded, preserving the original intent and compliance (see Figure 29). If the metadata is missing or the source cannot be authenticated, the system switches to an "auto pilot" mode, running a real-time loudness leveler (RTLL) to generate an ITU-R loudness-compliant gain offset value for transmission in the AC-4 bitstream. That gain offset value is automatically applied in the playback system. When compliant programming returns, the RTLL process is inaudibly bypassed. AC-4 is also designed to ensure that loudness compliance is maintained when several substreams are combined into a single presentation (see section III) upon decoding, e.g. M&E+D, or Main+Associated presentations.

The AC-4 system carries one or more dynamic range compression profiles (DRC), plus loudness information to the decoder. In addition to standard profiles, custom profiles can also be created for any type of playback device or content. This approach minimizes bitstream overhead compared to legacy codecs while supporting a more typical and desirable multiband DRC system that can be applied to the final rendered audio.

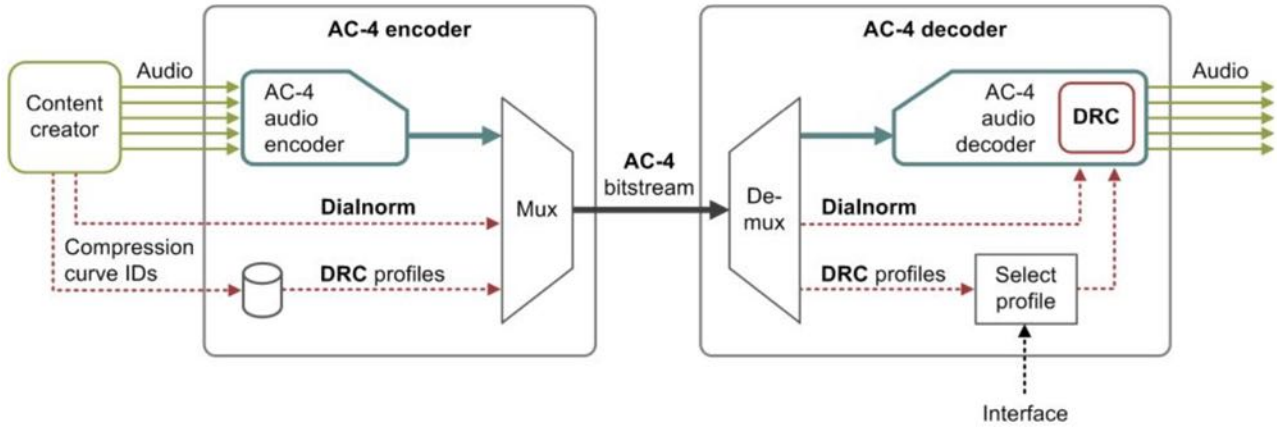

Figure 29 AC-4 DRC generation and application

A flexible DRC solution is essential to serve the wide range of playback devices and playback environments, from high-end Audio Video Receiver (AVR) systems and flat-panel TVs in living rooms to tablets, phones and headphones on-the-go. The AC-4 system defines four independent

DRC decoder operating modes that correspond to specific Target Reference Loudness, as shown in Table 6.

Table 6 Common target reference loudness for different devices

DRC Decoder mode	Target Reference Loudness [dB_{FS}]
Home Theater	-31..-27
Flat panel TV	-26..-17
Portable – Speakers	-16..0
Portable – Headphones	-16..0

7.4.2 Hybrid Delivery

AC-4 is designed to support hybrid delivery where, e.g. audio description or an additional language is delivered over a broadband connection, while the rest of the AC-4 stream is delivered as a broadcast stream.

The flexibility of the AC-4 syntax allows for easy signaling, delivery and mixing upon playback of audio substreams, which allows for splitting the delivery/transmission across multiple delivery paths. At the receiver side the timing information needed to combine the streams can be obtained from the AC-4 bitstream. In cases where DASH is used in both the broadcast and broadband transport, this information could be obtained from the transport layer.

7.4.3 Backward Compatibility

Dolby Atmos audio programs can be encoded using the AC-4 codec or the EAC-3 codec. When Atmos is used with E-AC-3 streams, backward compatibility is provided for existing non-Atmos E-AC-3 decoders. See Section 11.5, Phase A [1] for details. Backward compatibility is achieved in a different way: an AC-4 decoder (e.g., an ATSC 3.0 television or an advanced AVR) can provide a multichannel PCM audio (plus metadata) downmix which is delivered over HDMI and correctly interpreted by current Atmos-enabled devices (e.g., a soundbar) to produce a full Dolby Atmos immersive experience. If the destination renderer only supports stereo or 5.1 channel audio, the renderer will correctly provide a downmix to those legacy formats.

7.4.4 Next Generation Audio Metadata and Rendering

There are several metadata categories necessary to describe different aspects of next generation audio within AC-4:

- Immersive program metadata – informs Object-based Audio rendering and includes parameters such as position and speaker-dependencies
- Personalized program metadata – specifies audio presentations and defines the relationships between audio elements
- Essential Metadata Required for Next-Generation Broadcast:
    - Intelligent Loudness Metadata – metadata to signal compliance with regional regulations, dialogue loudness, relative-gated loudness, loudness correction type, etc.

- Program Synchronization – metadata to allow other sources/streams to be synchronized with the primary (emitted) presentation with frame-based accuracy.
- Legacy Metadata – traditional metadata including dialnorm, DRC, downmixing for Channel-based Audio, etc.

In the following three sections overview of the above three main metadata types are given.

7.4.5 Overview of Immersive Program Metadata and rendering

7.4.5.1 Object-based Audio Rendering

Object audio renderers also include control over the perceived object size (see "object width" metadata parameter in Section 7.4.7.3), which provides mixers with the ability to create the impression of a spatially extended source, which can be controlled within the same frame of reference (see Figure 30).

An audio object rendering engine is required to support Object-based Audio for immersive and personalized audio experiences. An audio renderer converts a set of audio signals with associated metadata to a different configuration of audio signals, e.g., speaker feeds, based on that metadata **AND** a set of control inputs derived from the rendering environment and/or user preference.

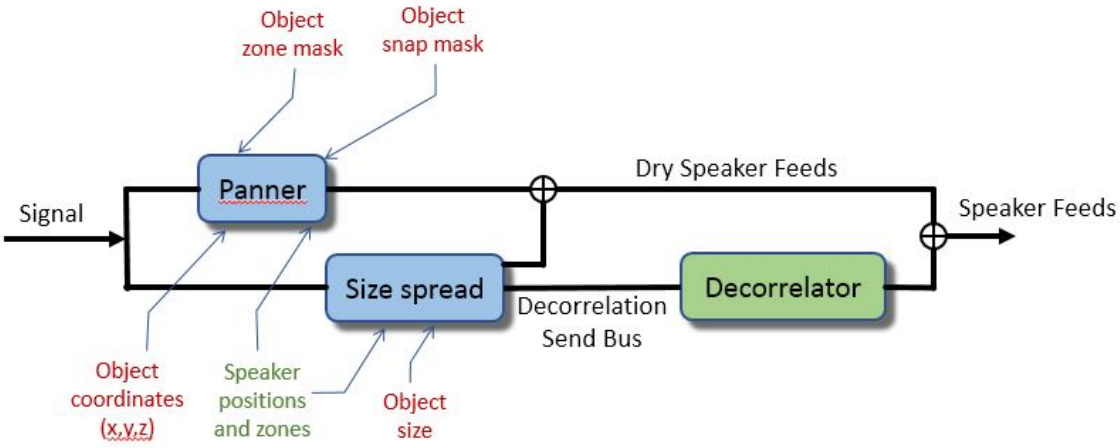

Figure 30 Object-based audio renderer

At the core of rendering are pan and spread operators, each executing a panning algorithm (see Figure 30) responsive to an audio object's coordinates (x,y,z). Most panning algorithms currently used in Object-based Audio production attempt to recreate audio cues during playback via amplitude panning techniques where, a gain vector G[1..n] is computed and assigned to the source signal for each of the *n* loudspeakers. The object audio signal s(t) is therefore reproduced by each loudspeaker *i* as $G_i(x,y,z) \times s(t)$ in order to recreate suitable localization cues as indicated by the object (x,y,z) coordinates and spread information as expressed in the metadata. There are multiple panning algorithms available to implement $G_i(x,y,z)$.

The design of panning algorithms ultimately must balance tradeoffs among timbral fidelity, spatial accuracy, smoothness and sensitivity to listener placement in the listening environment, all

of which can affect how an object at a given position in space will be perceived by listeners. For instance, Figure 31 illustrates how different speakers maybe utilized among various rendering (panning) algorithms to place an object's perceived position in the playback environment.

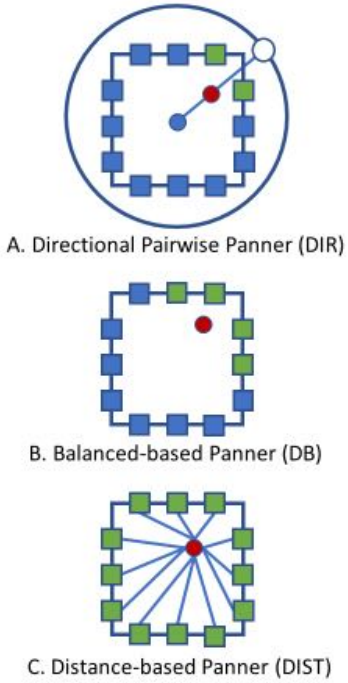

Figure 31 Common panning algorithms

Directional pairwise panning (DIR) (see Figure 31-A) is a commonly used strategy that solely relies on the directional vector from a reference position (generally the sweet spot or center of the room) to the desired object's position. The pair of speakers 'bracketing' the relevant directional vector is used to place (render) that object's position in space during playback. A well-documented extension of directional pairwise panning to support 3D loudspeaker layouts is vector-based amplitude panning which uses triplets of speakers. As this approach only utilizes the direction of the source relative to a reference position, it cannot differentiate between object sources at different positions along the same direction vector. It can also introduce instabilities as objects are panned near the center of the room. Moreover, some 3D implementations may constrain the rendered objects to the surface of a unit sphere and thus would not necessarily allow an object to cross inside the room without going 'up and over'. DIR can cause sharp speaker transitions as objects approach the center of the room with the result that rendering whips around from one side of a room to the opposite, momentarily tagging all the speakers in between.

The balanced-based (DB) panning algorithm, also known as "dual-balanced" is the most common approach used in 5.1/7.1-channel surround productions today (Figure 31-B). This approach utilizes left/right and front/back pan pot controls widely used for surround panning. As a result, dual-balance panning generally operates on the set of four speakers bracketing the desired 2D object position.

Extending to three-dimensions (e.g., when utilizing a vertical layer of speakers above the listener) yields a "triple-balance" panner. It generates three sets of one-dimensional gains

corresponding to left/right, front/back and top/bottom balance values. These values can then be multiplied to obtain the final loudspeaker gains:

$$Gi(x,y,z) = Gx\_i(x) \times Gy\_i(y) \times Gz\_i(z)$$

This approach is fully continuous for objects panned across the room in either 2D or 3D and makes it easier to precisely control how and when speakers on the base or elevation layer are to be used.

In contrast to the directional and balance-based approaches, distance-based panning (DIST) (Figure 31-C) uses the relative distance from the desired 2D or 3D object location to each speaker in use to determine the panning gains. Thus, this approach generally utilizes all the available speakers in use rather than a limited subset, which leads to smoother object pans but with the tradeoff of being prone to timbral artefacts, which can make the sound seem unnatural.

One aspect that both 'dual balance' and 'distance-based' panning share is the inherent smooth object pans in the sense that a small variation in an object's position will translate to a small change in loudspeaker gains.

The spread information (as defined by the ObjectWidth metadata) can be used to modify any of the panning algorithms, increasing the virtual 'size' of the object, modifying the signal strength at each speaker appropriately. That is, in the pairwise approach or balanced-based panning approaches, the spread operator modifies the signal strength of the more distant speakers providing a virtual sense of object width. In the distance-based panning approach (DIST), the actual object itself is sized as if it had the specified ObjectWidth. Speakers on each side of the virtual object would have their strength adjusted to represent the location and the size/spread of the virtual object.

The choice of mode and related trade-offs are up to the content creator.

7.4.5.2 Rendering-control metadata

As stated earlier, Object-based (immersive) Audio rendering algorithms essentially map a monophonic audio Object-based signal to a set of loudspeakers (based on the associated positional metadata) to generate the perception of an auditory event at an intended location in space.

While the use of a consistent core audio rendering algorithm is desirable, it cannot be assumed that a given rendering algorithm will always deliver consistent and aesthetically pleasing results across different playback environments. For instance, today the production community commonly remixes the same soundtrack for different Channel-based formats in use worldwide, such as 7.1/5.1 or stereo, to achieve their desired artistic goals for each format. With potentially over one hundred audio tracks competing for audibility, maintaining the discreteness of the mix and finding a place for all the key elements is a challenge that all theatrical/TV mixers face. Achieving success often requires mixing rules that are deliberately inconsistent with a physical model or a direct re-rendering across different speaker configurations.

To achieve this, AC-4 employs additional metadata to dynamically reconfigure the object renderer to "mask out" certain speaker zones during playback of a particular audio object. This is shown as the zone mask metadata in Figure 30. This guarantees that no loudspeaker belonging to the masked zones will be used for rendering the applicable object. Typical zone masks used in production today include: *no sides, no back, screen only, room only and elevation on/off*.

The main application of speaker zone mask metadata is to help the mixer achieve a precise control of which speakers are used to render each object in order to achieve the desired perceptual effect. For instance, the *no sides* mask guarantees that no speaker on the side wall of the room will

be used. This creates more stable screen-to-back fly-throughs. If the side speakers are used to render such trajectories, they will become audible for the seats nearest to the side walls and these seats will perceive a distorted trajectory "sliding" along the walls rather than crossing the center of the room.

Another key application of zone masks is to fine tune how overhead objects must be rendered in a situation where no ceiling speakers are available. Depending on the object and whether it is directly tied to an on-screen element, a mixer can choose, e.g., to use the *screen only* or *room only* mask to render this object, in which case it will be rendered only using screen speakers or using surround speakers, respectively, when no overhead speakers are present. Overhead music objects, for instance, are often authored with a *screen only* mask.

Speaker zone masks also provide an effective means to further control which speakers can be used as part of the process to optimize the discreteness of the mix. For instance, a wide object can be rendered only in the 2D plane by using the *elevation off* mask. To avoid adding more energy to screen channels, which could compromise dialogue intelligibility, the *room only* mask can be used.

Another useful aesthetic control parameter is the *snap-to-speaker mode* represented by snap mask metadata (see Figure 30). The mixer can select this mode for an individual audio object to indicate that consistent reproduction of timbre is more important than consistent reproduction of the object's position. When this mode is enabled, the object renderer does not perform phantom panning to locate the desired sound image. Rather, it renders the object entirely from the single loudspeaker nearest to the intended object location.

Reproduction from a single loudspeaker creates a pin-point (very discrete) and timbrally neutral source that can be used to highlight key effects in the mix, particularly more diffuse elements such as those being rendered utilizing the Channel-based elements.

A common use case for the *snap-to-speaker* parameter is for music elements, e.g., to extend the orchestra beyond the screen. When re-rendered to sparser speaker configurations (e.g., legacy 5.1 or 7.1), these elements will be automatically snapped to left/right screen channels. Another use of the snap metadata is to create "virtual channels", for instance to re-position the outputs of legacy multichannel reverberation plug-ins in 3D.

7.4.6 Overview of Personalized Program Metadata

Object-based Audio metadata defines how audio objects are reproduced in a sound field, and an additional layer of metadata defines the personalization aspects of the audio program. This personalization metadata serves two purposes: to define a set of unique audio "presentations" from which a consumer can select, and to define dependencies (i.e., constraints, e.g., maximum gain for music) between the audio elements that make up the individual presentations to ensure that personalization always sounds optimal.

7.4.6.1 Presentation Metadata

Producers and sound mixers can define multiple audio presentations for a program to allow users to switch easily between several optimally pre-defined audio configurations. For example, a sports event, a sound mixer could define a default sound mix for general audiences, biased sound mixes for supporters of each team that emphasize their crowd and favorite commentators, and a commentator-free mix. The defined presentations will be dependent on the content genre (e.g. sports, drama, etc.), and will differ from sport to sport. *Presentation* metadata defines the details that create these different sound experiences.

An audio *presentation* specifies which object elements/groups should be active along with the position and their absolute volume level. Defining a default audio presentation ensures that audio is always output for a given program. *Presentation* metadata can also provide conditional rendering instructions that specify different audio object placement/volume for different speaker configurations. For example, a dialogue object's playback gain may be specified at a higher level when reproduced on a mobile device as opposed to an AVR.

Each object or audio bed may be assigned a category such as dialogue or music & effects. This category information can be utilized later either by the production chain to perform further processing or used by the playback device to enable specific behavior. For example, categorizing an object as dialogue would allow the playback device to manipulate the level of the dialogue object with respect to the ambience.

Presentation metadata can also identify the program itself along with other aspects of the program (e.g., which sports genre or which teams are playing) that could be used to automatically recall personalization details when similar programs are played. For example, if a consumer personalizes their viewing experience to always pick a radio commentary for a baseball game, the playback device can remember this genre-based personalization and always select the radio commentary for subsequent baseball games.

The *presentation* metadata also contains unique identifiers for the program and each presentation.

Presentation metadata typically will not vary on a frame-by-frame basis. However, it may change throughout the course of a program. For example, the number of presentations available may be different during live-game-play but may change during a half-time presentation.

7.4.7 Essential Metadata Required for Next-Generation Broadcast

This section provides a high-level overview of the most essential metadata parameters required for enabling next-generation audio experiences. Essential metadata is capable of being interchanged for both file-based workflows (as per the ITU-R BWF/ADM formats [26], [27]) AND in serialized form for real-time workflows and interconnects utilizing SMPTE ST 337 [34] formatting/framing.

7.4.7.1 Intelligent Loudness Metadata

The following section highlights the essential loudness-related metadata parameters required for next-generation broadcast systems. Intelligent Loudness metadata provides the foundation for enabling automatic (dynamic) bypass of cascaded (real-time or file-based) loudness and dynamic range processing commonly found throughout distribution and delivery today. Intelligent Loudness metadata is supported for both channel- and object-based audio representations.

Dialogue Normalization Level – This parameter indicates how far the average dialogue level is below 0 LKFS.

Loudness Practice Type - This parameter indicates which recommended practice was followed when the content was authored or corrected. For example, a value of "0x1" indicates the author (or automated normalization process) was adhering to ATSC A/85 [2]. A value of "0x2" indicates the author was adhering to EBU R 128 [14]. A special value, "0x0" signifies that the loudness recommended practice type is not indicated.

Loudness Correction Dialogue Gating Flag - This parameter indicates whether dialogue gating was used when the content was authored or corrected.

Dialogue Gating Practice Type - This parameter indicates what dialogue gating practice was followed when the content was authored or corrected. This parameter is typically 0x02 – "Automated Left, Center and/or Right Channel(s)". However, there are values for signaling manual selection of dialogue, as well as other channel combinations, as detailed in the ETSI TS 103 190 [17].

Loudness Correction Type - This parameter indicates whether a program was corrected using a file-based correction process, or a real-time loudness processor.

Program Loudness, Relative Gated - This parameter is entered into the encoder to indicate the overall program loudness as per ITU-R BS.1770-4 [24]. In ATSC regions, this parameter would typically be -24.0 LKFS for short-form content as per ATSC A/85 [2]. In EBU regions, this parameter would typically indicate -23.0 LKFS (LUFS).

Program Loudness, Speech Gated - This parameter indicates the speech-gated program loudness. In ATSC regions, this parameter would typically be -24.0 LKFS for long-form content as per ATSC A/85 [2].

max_loudstrm3s - This parameter indicates the maximum short-term loudness of the audio program measured per ITU-R BS.1771 [25].

max_truepk - This parameter indicates the maximum true peak value for the audio program measured per ITU-R BS.1770 [24].

loro_dmx_loud_corr - This parameter is used to calibrate the downmix loudness (if applicable), as per the Lo/Ro coefficients specified in the associated metadata and/or emission bitstream, to match the original (source) program loudness. Note: this parameter is not currently supported in the pending ITU-R BWF/ADM [26] format.

ltrt_dmx_loud_corr - This parameter is used to calibrate the downmix loudness (if applicable), as per the Lt/Rt coefficients specified in the associated metadata and/or emission bitstream to match the original (source) program loudness. Note: This parameter is not currently supported in the pending ITU-R BWF/ADM [26] format.

Note regarding the loudness measurement of objects: The proposed system supports loudness estimation and correction of both Channel-based and Object-based (immersive) programs utilizing the ITU-R BS.1770-4 [24] recommendation.

AC-4 supports the carriage (and control) of program loudness at the presentation level. This ensures any presentation (constructed from one or more sets of program elements or substreams) available to the listener will maintain a consistent loudness.

7.4.7.2 Personalized Metadata

Personalized audio consists of one or more audio elements with metadata that describes how to decode, render and output "full" mixes defined as one or more presentations. Each personalized audio presentation typically consists of an ambience (often part of a Program Bed, a static audio element, defined below), one or more dialogue elements, and optionally one or more effects elements. For example, a presentation for a hockey game may consist of a 5.1 ambience bed, a mono dialogue element, and a mono element for the public-announcement speaker feed. Multiple presentations may be defined throughout the production system and emission (encoded) bitstream to support several options such as alternate language, dialogue, ambiance, etc. enabling height elements, and so on. As an example, the AC-4 bitstream always includes a default presentation that would replicate the default stereo or 5.1 legacy program that is delivered to downstream devices that are only capable of stereo or 5.1 audio.

The primary controls for personalization are:

- Presentation selection
- Dialogue element volume level

The content creator can have control over the options presented to the user. Moreover, they can choose to disable viewer dialogue control or limit the range of viewer control to address any content agreements and/or artistic needs.

While personalized audio metadata is typically static throughout an entire program, it could change dynamically at key points during the event. For example, options for personalization may differ during the half-time show of a sporting event as opposed to live game play.

7.4.7.3 Object Audio Metadata

This section provides an overview of the metadata parameters (and their application) essential for enabling next-generation immersive experiences in the AC-4 system.

Object-based audio consists of one or more audio signals individually described with metadata. Object-based audio can contain static bed objects (similar to Channel-based Audio) which have a fixed nominal playback position in 3-dimensional space and dynamic objects with explicit positional metadata that can change with time. Object-based Audio is closely linked to auditory image position rather than presumed loudspeaker positions. The object audio metadata contains information used for rendering an audio object.

The primary purpose of the object audio metadata is to:

- Describe the composition of the Object-based Audio program
- Deliver metadata describing how objects should be rendered
- Describe the properties of each object (for example, position, type of program element [e.g., dialog], and so on)

Within the production system, a subset of the object audio metadata fields is essential to provide the best audio experience and to ensure that the original artistic intent is preserved. The remaining non-essential metadata fields described in ETSI TS 103 190-2 [17] are used for either an enhanced playback applications or aiding in the transmission and playback of the program content.

Metadata critical to ensure proper rendering of objects and provide sufficient artistic control include:

- Object type / assignment
- Timing (timestamp)
- Object position
- Zone / elevation mask
- Object width
- Object snap
- Object divergence

Object Type / Object Assignment - To properly render a set of objects, both the object type and object assignment of each object in the program must be known.

For spatial objects, two object types defined for current Object-based Audio production.

Bed objects - This is an object with positional metadata that does not change over time and is described by a predefined speaker position. The object assignment for bed objects describe the intended playback speaker, for example, Left (L), Right (R), Center (C) … Right Rear Surround (Rrs) … Left Top Middle (Ltm).

Dynamic objects - A dynamic object is an object with metadata that may vary over time, for example, position.

Timing (timestamp) - Object audio metadata can be thought of a series of metadata events at discrete times throughout a program. The timestamp indicates when a new metadata event takes effect. Each metadata event can have, for example, updates to the position, width, or zone metadata fields.

Object Position - The position of each dynamic object is specified using three-dimensional coordinates within a normalized, rectangular room. The position is required to render an object with a high degree of spatial accuracy.

Zone / Elevation mask - The zone and elevation mask metadata fields describe which speakers, either on the listener plane or height plane of the playback environment, shall be enabled or disabled during rendering for a specific object. Each speaker in the playback environment can belong to either the screen, sides, backs or ceiling zones. The mask metadata instructs the renderer to ignore speakers belonging to a given zone for rendering. For instance, to perform a front to back panning motion, it might be desirable to disable speakers on the side wall. It might also be useful to limit the spread of a wide object to the two-dimensional surround plane by disabling the elevation zone mask. Otherwise, objects are spread uniformly in three-dimensions including the ceiling speakers. Finally, masking the screen would let an overhead object be rendered only by surround speakers for configurations that do not comprise ceiling channels. As such, zone mask is a form of conditional rendering metadata.

Object Width - Object width specifies the amount of spread to be applied to an object. When applied, object width increases the number of speakers used to render a particular object and creates the impression of a spatially wide source as opposed to a point source. By default, object width is isotropic and three-dimensional unless zone masking metadata is used.

Object Snap - The object snap field instructs the renderer to reproduce an object via single loudspeaker. When object snap is used, the loudspeaker chosen to reproduce the object is typically the one closest to the original position of the object. The snap functionality is used to prioritize timbral accuracy during playback.

Object Divergence - Divergence is a common mixing technique used in broadcast applications. It is typically used to spread a Center channel signal (for example, a commentator voice) across the speakers in screen plane instead of direct rendering to the center speaker. The spread of the Center channel signal can range from all center (full convergence), through equal level in Left, Right, and Center speakers, to full divergence where all the energy is in the Left and Right speakers with none in Center speaker. Regardless of how the center signal is spread, full convergence or full divergence, the spatial image of the center signal remains consistent. This can be applied to any signal, including objects (including bed objects) and channel-based selections.

The object divergence field controls the amount of direct rendering of the object compared with the rendering of two virtual sources spaced equidistantly to the left and right of the original object using identical audio. At full convergence, the object is directly rendered, as it would be normally. At full divergence, the object is reproduced by rendering the two virtual sources.

7.4.8 Metadata Carriage

In the production system, different methods are introduced for enabling the carriage of metadata described above within file-based and real-time (HD-SDI) contribution/distribution workflows to

address a wide range of industry needs related to interoperability and reliability necessary for day-to-day operations.

7.4.8.1 File-based carriage of Object- and Channel-based Audio with metadata

With the growing interest across the worldwide broadcast industry to enable delivery of both immersive and personalized (interactive) experiences, additional information (i.e., the metadata) must co-exist to describe fully these experiences. The EBU Audio Definition Model [13] has provided the foundation for the development of an international recommendation within ITU-R WP6B, which produced the ITU-R Audio Definition Model (ADM) [26].

The ITU-R ADM [26] specifies how XML data can be generated to provide definitions of tracks and associated metadata within Broadcast Wave (BWF), RF64 files or as a separate file that references associated essence files. In general, the ADM describes the associated audio program as two parts via the XML. The *content* part describes what is contained in the audio (e.g. language, loudness, etc.), while the *format* part describes the technical detail of the underlying audio to drive either decoding and/or rendering properly – including the rendering of Object-based Audio as well as signaling of compressed audio formats in addition to LPCM.

ITU-R BS.2088 [27] incorporates the ADM into the Broadcast Wave (BWF) and RF64 File formats (BW64) as well as well as incorporating metadata within the legacy BWF format as defined in Recommendation ITU-R BR.1352 [28]. ITU-R BS.2088 allows the ubiquitously supported audio file format, B-WAV, to carry numerous audio program representations including Object-based immersive along with audio programming containing elements that are intended to be used for personalization.

The ITU-R BWAV/ADM Recommendation is a critical element for enabling the Object-based Audio content pipeline and it is expected that regional application standards and recommendations will reference this format for Object-based program (file) interchange. Moreover, being an international and open recommendation, accelerated adoption from vendors supplying workflow solutions throughout the worldwide broadcast and post production industries is anticipated as immersive and personalized content creation becomes commonplace.

7.4.8.2 Real-time carriage of Object- and Channel-based Audio with metadata

The reliable carriage of audio metadata across real-time interfaces and workflows within HD-SDI has been a long-standing challenge for the industry over the years. Moreover, the existing method(s) could only describe a limited number of Channel-based Audio programs along with limitations in terms of extensibility to support future needs. In the production system, there is a framework and accompanying bitstream format to be carried across any AES3 channel pair within the HD-SDI format. One embodiment of this framework/bitstream is a SMPTE 337 [34] formatted derivative of the Extensible Format for the Delivery of Metadata (EMDF) originally defined in ETSI TS 102 366 Annex H[9] [16]. (See Figure 32)

[9] Note: the EMDF framework described in Annex H is also known as the Evolution Framework (EVO).

* Dynamic object metadata carried via this method embedded in HD-SDI is required to maintain sync to within ~ +/- 40 AES samples (@ 48kHz) with the associated audio object channel(s)/track(s).

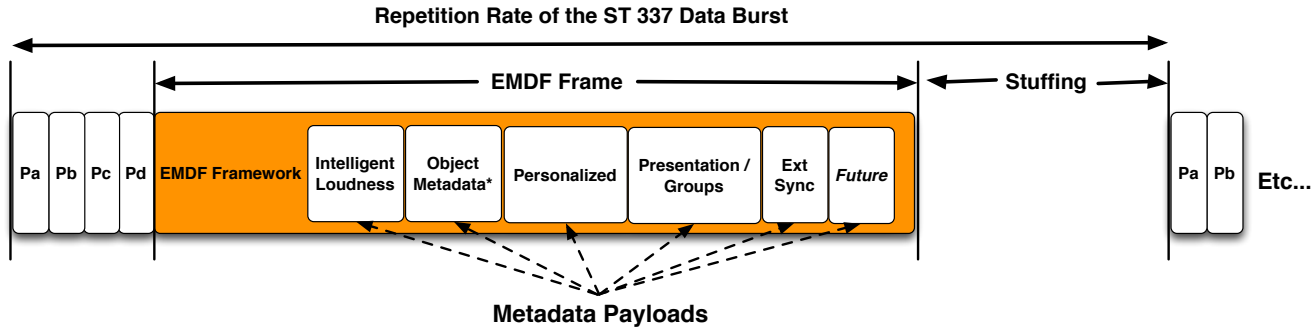

Figure 32 Serialized EMDF Frame formatted as per SMPTE ST 337 [34]

The EMDF specifies the carriage of metadata in a serialized (and efficient) form made up of 'payloads' each with a unique ID. Payload IDs can signal the carriage of several types of metadata (and associated DRM system-specific protection information) necessary for next-generation audio including immersive (object), personalized, intelligent loudness (i.e., as per ETSI TS 102 366 [16] Annex H payload_id 0x1), second-stream synchronization, and so on. Tools to translate to/from the metadata format defined in ITU-R BWF/ADM referenced earlier are necessary, including conversion(s) to support frame-based splicing and cloud-based distributed processing required by interchange, distribution and emission systems. The open standardization of the EMDF framework/bitstream and associated payloads allow efficient real-time interchange of immersive (object), personalized, intelligent loudness, second-stream sync metadata, etc. within the SMPTE 337 family of standards for use in today's HD-SDI environments, while also ensuring the design supports efficient transport of metadata payloads for IP-based environments going forward. Operational note: The SMPTE 337 EMDF bitstream is a critical component to enable automated (and seamless) switching of a broadcast emission encoder to accommodate day-to-day operations where legacy programming is interleaved with next-generation programming utilizing program-specific embedded audio channel (or audio object) layouts.

8. Content Aware Encoding

8.1 Introduction

Content Aware Encoding, also referred to as Content-Adaptive Encoding, or CAE, is a technique applied during the encoding process to improve the efficiency of encoding schemes. It can be used with any codec, but in the context of this document we will solely focus on HEVC.

We will describe in this chapter how CAE works, how it can be applied to Ultra HD and the benefits of using CAE for the transmission of UHD program material. CAE is not a standard, but a technique applied on the encoder side that is expected to be decoded by an HEVC Main 10 decoder. Regarding adaptive streaming, the only existing specification is iOS 11[10]. As of the publication of v1.0 of these Phase B Guidelines Ultra HD Forum is seeking DASH IF Guidance to support VBR encoded content on the client.

As opposed to other techniques such as HDR, WCG, NGA or HFR, where new devices or network equipment are required, CAE just requires an upgrade of the encoder and should work with any decoder. All networking and interoperability aspects are described in this Section.

8.1.1 Adaptive Bitrate Usage for UHD

For OTT, ABR is already the most common way to deliver content. CAE is applied on top of ABR in the encoding process. Currently only iOS11[11] has done that, but we expect a wider support such as from Android, DASH, and DVB in the future. For managed IP networks (Cable, Telcos), we also see ABR being used.

Cable operators can broadcast Live over either QAM or ABR or over IP (DOCSIS® 3.0 [32]). The IP delivery may be performed in Unicast as the traffic is not expected to be high, and may later be scaled using ABR Multicast CableLabs [9].

For Telco operators, they can use either IP Multicast or Unicast using ABR.

8.1.2 Per-title Encoding

Content aware encoding was introduced in production by Netflix®[12] in 2015 using “per-title encoding”[13]. In summary Netflix discovered that the ABR ladder defined for the video encoding was very much dependent on the content and that for each title they would consider an optimized ladder where each step provides a just noticeable difference (JND) in quality (originally using PSNR, Netflix developed the VMFA metric) at the lowest bitrate. In addition, as the content complexity changes during a movie, the bitrate per resolution should also vary. Netflix later refined that model, by changing the ladder not per-title, but per-segment.

---

[10] https://developer.apple.com/library/content/documentation/General/Reference/HLSAuthoringSpec/Requirements.html

[11] ibid

[12] Netflix is a registered trademark of Netflix, Inc.

[13] https://medium.com/netflix-techblog/per-title-encode-optimization-7e99442b62a2

The main drawback of the original method applied by Netflix is that all the different combinations of the encoding parameters (resolution, bitrate, etc.) were used to generate intermediate encodings, and only then was the optimization process applied to select best combinations of encodings to use in the final ABR ladder. This is a CPU intensive technique, possibly applicable to Cloud for VOD, but does not fit the Live use case.

Some of the more recent implementations of CAE ladder generators reviewed[14] do not require full additional transcodes to be done ahead of time, making them more practical, and applicable in both VOD and Live use cases.

8.1.3 VBR Encoding

VBR achieves bitrate savings by only using as many bits as are required to achieve the desired video quality for a given scene or segment. Simpler scenes are encoded at a much lower rate (e.g., 80 percent less) than complex ones, with no discernible difference in quality to viewers.

A drawback of traditional VBR streaming is that the bitrate of an encoded stream can be very high during complex scenes, putting OTT content providers at risk of exceeding the streaming bandwidth supported by the network. The maximum bitrate is chosen based on a combination of network bandwidth limitations and the video quality delivered during complex scenes. Setting a ceiling for the maximum bitrate of the stream, known as Capped VBR (CVBR), resolves this issue by protecting the streaming bandwidth. But the technique is not infallible.

CVBR may be thought of as a subset of CAE. CVBR cannot achieve the same performance as CAE because it does not include the same flexibility to change the profile ladder or resolution (as described for CAE in the next section). In addition, in practice, many older CVBR implementations used simple and inaccurate models of video quality which further limited the performance gain they could achieve in comparison to CBR.

Given the limitations of traditional encoding systems, content providers need a more effective method for measuring the ideal video quality and compression level of each video scene. CAE encoding techniques may be deployed in a Live or VOD environment with average savings over VBR and CVBR encoding in the range of 20-50%.

8.2 Content Aware Encoding Overview

Content Aware Encoding or Content-Adaptive Encoding (CAE) is a class of techniques for improving efficiency of encodings by exploiting properties of the content. By using such techniques, “simple” content, such as scenes with little motion, static images, etc. will be encoded using fewer bits than “complex” content, such as high-motion scenes, waterfalls, etc. By so doing, content-aware techniques aim to spend only a minimum number of bits necessary to ensure quality level needed for delivery. Since “simple” content is prevalent, the use of CAE techniques results in significant bandwidth savings and other benefits to operators (e.g., some systems may also reduce the number of encodings, deliver higher resolution in the same bits as the previous systems

[14] Jan Ozer, One Title at a Time: Comparing Per-Title Video Encoding Options, Oct 2017, Streaming Media magazine, http://www.streamingmedia.com/Articles/Editorial/Featured-Articles/One-Title-at-a-Time-Comparing-Per-Title-Video-Encoding-Options-121493.aspx

required for lower resolution, better overall quality, etc.). The CAE process is the "secret sauce" of an encoder company as described in several references[15].

8.2.1 Principles

The CAE can be applied to either or both VOD and Live use cases. From an operational point of view, it is recommended that this function be applied in the encoder, though it can be effective as a post process depending on the needs of the workflow and the architectural demands of the video encoding system.

CAE techniques can be applied at different levels, described in Table 7.

Table 7 CAE granularity

Level	Description	Application
Per ladder	Encoder looks at the entire file and decides: a) how many streams to include in the ABR ladder, b) which resolutions/framerates to use for each stream, c) how to allocate bits within each of the encoded streams	VOD
Per stream	Encoder looks at the entire file and decides where to allocate the bits	VOD
Per segment	The encoder looks at the complete segment horizon to allocate the bits	VOD, Live*
Per frame	The encoder allocates the bits within the frame	Live, VOD
Per Macroblock	The encoder allocates the bits within the frame	Live, VOD

*This might bring unacceptable additional delay (latency).

8.3 Content Aware Encoding applied to UHD

When applied to Ultra HD using any of the tools captured in the Ultra HD Forum Guidelines, CAE can provide significant savings. We will use CBR as a reference as this is the de-facto encoding mode used in the past for ABR encoding though the technology functions just the same with a VBR input.

[15] http://info.harmonicinc.com/Tech-Guide-Harmonic-EyeQ
http://media2.beamrvideo.com/pdf/Beamr_Content_Adaptive_Tech_Guide.pdf
https://www.brightcove.com/en/blog/2017/05/context-aware-encoding-improves-video-quality-while-cutting-costs
Jan Ozer, One Title at a Time: Comparing Per-Title Video Encoding Options, Oct 2017, Streaming Media magazine, http://www.streamingmedia.com/Articles/Editorial/Featured-Articles/One-Title-at-a-Time-Comparing-Per-Title-Video-Encoding-Options-121493.aspx

Table 8 provides examples of three ABR encodings ladders. Note that these are examples provided to give the reader an indication of the bitrates that may be possible; however, the nature of the content and other factors will affect bitrate. All ladders use the same set of DVB-DASH-recommended resolutions [12], ranging from HD (720p) to UHD (2160p), but they differ in rates. The first ladder (shown in column 4) is a fixed CBR encoding design, assigning bitrates that are chosen independently, regardless of the type of content being encoded. The ladders shown in columns 5 and 6 are examples of CAE ladders generated for two different types of content. The CAE ladder in column 5 is produced for easier-to-encode content resulting in an average savings of more than 50%. The CAE ladder in column 6 is produced for more difficult content, resulting in an average savings of 40-50% vs. CBR encoding, depending on the content complexity.

Note that the CAE technique is truly content dependent, while in a CBR mode; more artefacts would be visible with high complexity content. With CAE, the bitrate will fluctuate with the content complexity, and will therefore provide a higher quality at same average bitrates vs. CBR. When a CAE stream cap is the same level as a CBR bitrate stream, the CAE stream can be 40-50% lower average bitrate than the CBR stream, while retaining the same quality video.

Table 8 Examples of fixed and CAE encoding ladders for live sports

Stream	Resolution	Frame Rate	CBR bitrate (Mbps)	CAE Easy Content: Ave. bitrate (Mbps)	CAE Complex Content Ave. bitrate (Mbps)
1	3840x2160	60	25	12	15
2	3840x2160	60	15	8	9
3	3200x1800	60	12	6	7
4	2560x1440	60	8	4	5
5	1920x1080	60	5	2.5	3
6	1600x900	60	3.6	1.8	2.1
7	1280x720	60	2.5	1.2	1.5

We draw in Figure 33 the bitrate vs. resolution of CAE vs. CBR at the same quality level. For simplicity for CAE, we use a more conservative example ladder, resulting in 40% savings.

We are plotting in Figure 33 CAE vs. CBR bitrates, assuming the same visual quality at a given resolution.

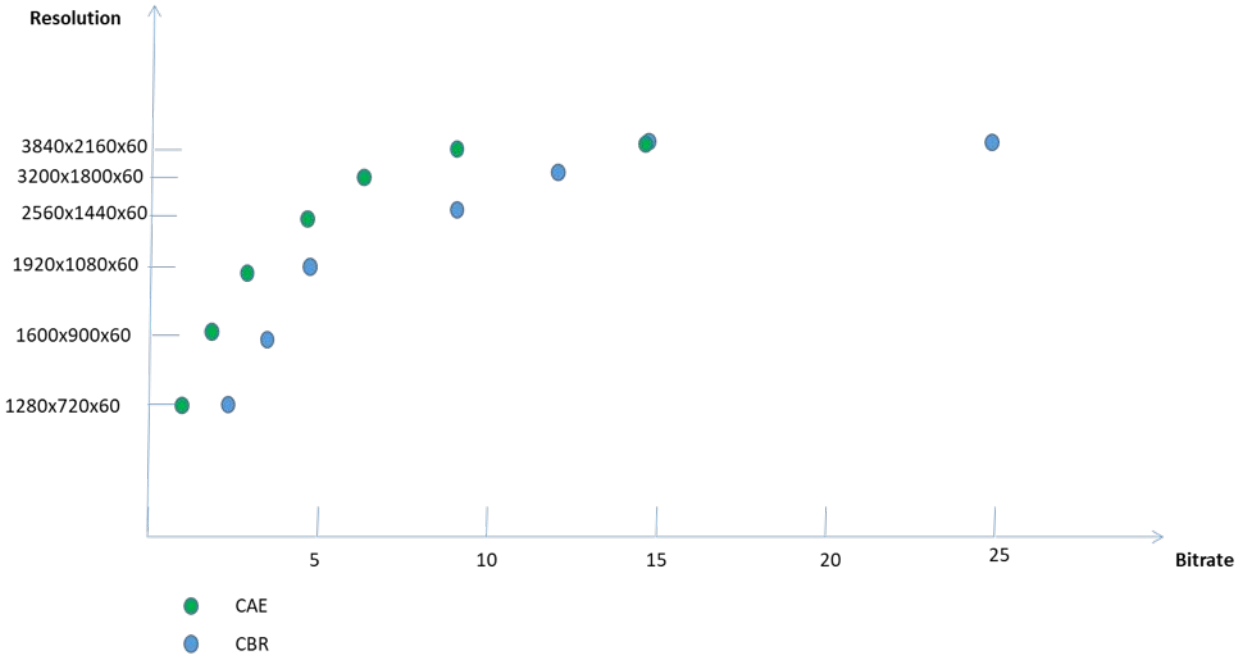

Figure 33 CAE encoding chart

8.4 Content Aware Encoding interoperability

The resulting bitstream from a CAE encoder is compliant with the guidelines for ABR delivery used in DVB-DASH [12] and Apple TV / HLS [20].

8.5 Application for Content Aware Encoding

We will describe in this section what can be the impact of CAE on Internet delivery of UHD.

8.5.1 Internet bandwidth

From Belson[16], Figure 34 shows the Internet speed distribution over various regions of the world.

[16] Belson D, "Akamai's state of the Internet, Q3 2016 report", https://content.akamai.com/PG7659-q3-2016-state-of-the-Internet-connectivity-report.html

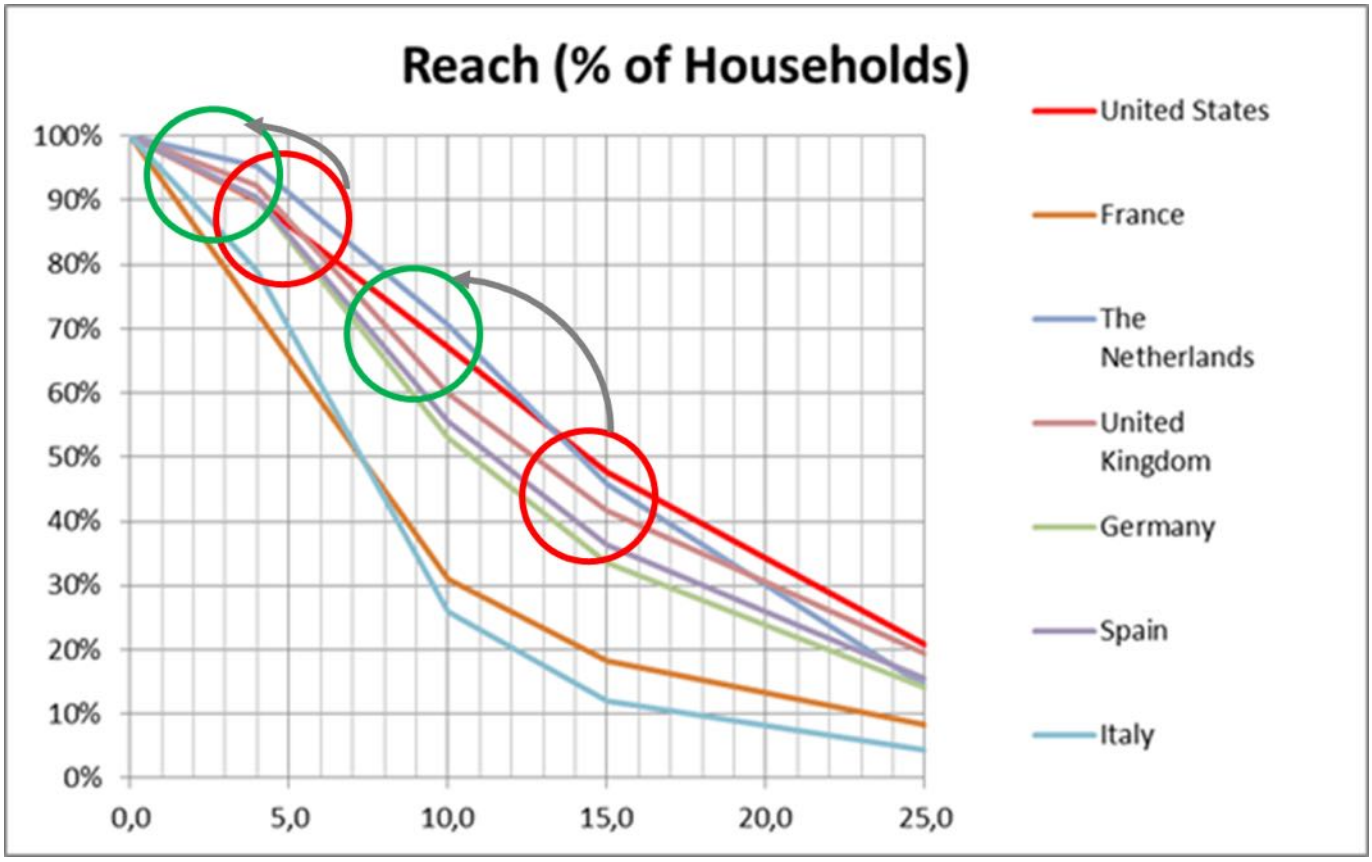

Figure 34 Internet speed distribution per countries (source Akamai)

We will look at CAE for two use cases: One to deliver the full UHD (2160p60) experience and the other one to deliver an HD (1080p60) experience. We will look at a group of countries who have a very homogeneous Internet speed distribution in their populations: Germany, France, Netherlands, UK and US.

2160p60 use case

At 15Mbps a CBR encoding of 2160p60 only reaches 40% of the population of those countries. CAE can offer 2160p60 at 9Mbps (on average) to 70% of the population. This is a significant 75% increase of the population that can be targeted.

1080p60 use case

At 5Mbps a CBR encoding of 1080p60 already reaches 85% of the population of those countries. CAE can offer 1080p60 at 3Mbps (on average) to 95% of the population. This is just an increase of 17% of the population that can be targeted.

From this chart, we can see that CAE has a larger impact on 2160p60 and this should push more OTT operators to deliver premium UHD experience at 2160p60 over the Internet.

8.5.2 CAE Sweet Spot for UHD

Based on the previous section finding, the CAE sweet spot is when 2160p60 can be delivered at a lower bitrate than the CBR case. We describe in Figure 35 the CAE sweet spot.

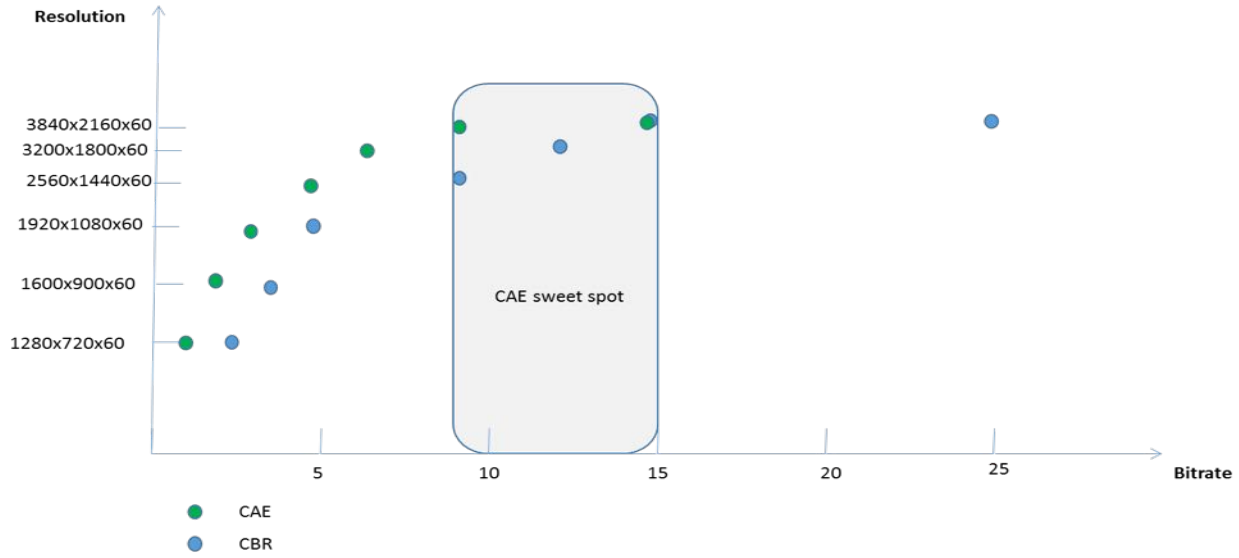

Figure 35 CAE sweet spot vs. CBR

The CAE sweet spot is between 9Mbps where CAE can deliver 2160p60 and 15Mbps, which we believe is the maximum quality CAE can provide for 2160p60.

8.6 Content Aware Encoding Benefits

8.6.1 CDN cost

Whatever the cost of the CDN for the OTT operator, CAE will reduce the cost by 40-50% in terms of streaming, storing vs. CBR for the streaming part, ingest to CDN and storage on CDN for VOD or catch up.

8.6.2 Quality of experience

Because the bandwidth required to carry CAE vs. CBR is reduced by 40-50%, the content will be transmitted in a smoother way across the delivery chain. Video services have reported up to a 50% reduction in re-buffering events and a 20% improvement in stream start times for VOD services. As the traffic is the same for Live or VOD, we expect the same network performances to apply for Live[17].

Due to the smaller size of the video bitrates, higher resolutions will become available to more viewers as compared with the traditional CBR encoding schemes in operation today.

As CAE bitrate is modulating vs. the complexity of video, the quality is guaranteed vs. the CBR encoding where the bitrate is guaranteed, but the quality always suffers on complex scenes.

From a purely qualitative point of view, at junction bitrates (i.e., bitrates where the CAE encoding is at a higher resolution than the CBR encoding), the quality will be improved as a higher resolution will be displayed. The junction bitrates are depicted in Figure 36.

[17] http://beamrvideomedia.s3.amazonaws.com/pdf/Beamr_M-GO_Case_Study_2015.pdf

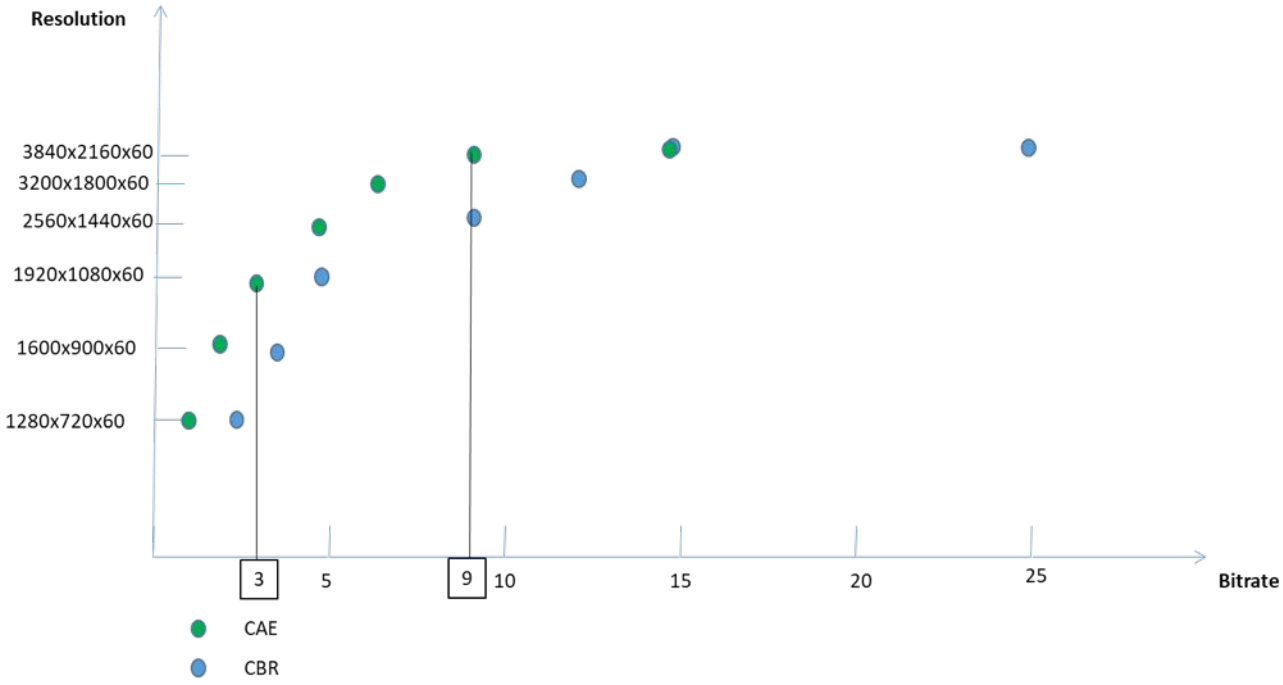

Figure 36 Junction bitrates chart

In a home Wi-Fi environment, transmitting bitrates higher than 10Mbps can be a challenge, therefore with 2160p60 being on average encoded at 9Mbps, the CAE experience will always be of better quality.

9. Annex A: AVS2

The Digital Audio and Video Coding Standard Working Group (AVS Workgroup) of China has delivered their latest generation Advanced Video Coding Standard (AVS2) to target UHD and HDR content for both broadcast and broadband communications and for storage. AVS2 standards[18,19] were published in 2016, with parallel work started by the IEEE P1857 workgroup[20]. An English-language standard of AVS2 is expected to be completed shortly (IEEE P1857.4).

9.1 Why AVS2

AVS2 is the successor to the earlier video coding standard AVS+[21], which was successor-in-turn to AVS1[22,23]. AVS2 has double the coding efficiency of AVS1. Testing hosted by the State Administration of Radio, Film, and Television (SARFT) determined that AVS2 compared favorably to HEVC, producing slightly less image degradation relative to source images at the same bitrate. Using 4K video sequences (2160p 10-bit) specified by China's National Film and Television Administration, a test identifying specific builds of reference software demonstrated AVS2 to have a 3.0% average performance advantage relative to HEVC[24], while the decoder complexity is similar.

Initially intended to support greater numbers of HD streams and the introduction of 4K content, the AVS2 architecture is also scalable for use with 8K images. The Main-10bit profile supports several levels from typical 60fps and up to 120fps for 4K and 8K content.

9.2 Deployment

AVS2 is already supported by chipsets from multiple manufacturers, for both set-top boxes and televisions. Further, licensable video coder technology is available for manufacturers wanting to design their own SoC. On the production side, encoders are available from multiple manufactures.

The predecessors to AVS2 have seen widespread deployment: AVS+ is presently used widely in China, Sri Lanka, Laos, Thailand, Kyrgyzstan, and other countries; while AVS1 is further used

[18] General Administration of Quality Supervision, Inspection and Quarantine (GAQSIQ) GB/T 33475.2-2016 "Information Technology - High Efficient Media Coding - Part 2: Video"

[19] State Administration of Press, Publication, Radio, Film and Television (SAPPRFT) GY/T 299.102016 "High Efficiency Coding of Audio and Video - Part 1: Video"

[20] IEEE 1857.4 "Standard for 2nd Generation IEEE 1857 Video Coding" [presently under development]

[21] GAQSIQ GB/T 20090.16-2016 "Information Technology - Advanced Audio and Video Coding Part 16: Radio and Television Video"

[22] GAQSIQ GB/T 20090.2-2006 "Advanced Video and Audio Coding for Information Technology Part 2: Video"

[23] IEEE 1857-2013 "IEEE Standard for Advanced Audio and Video Coding"

[24] Digital Media Research Center, Peking University, "Who will lead the next generation of video coding standards: HEVC, AVS2 and AV1 performance comparison report"

in Burma, Cuba, and Uzbekistan. In December 2017, Guangdong Radio and Television (GRT) announced a pilot for China's first 4K UHD channel and its use of AVS2. Growth of that channel is expected to reach 15M subscribers in the Guangdong province alone.

9.3 Technology

AVS2 uses a coding framework as shown in Figure 1. The residual between an image and a prediction of the image is compressed by the transform and quantization module. Thereafter, de-quantization and inverting the transform reconstitutes the residual and encoded image (green modules). Two classes of predictor are available: Intraprediction (orange modules) to detect and exploit similarities within the encoded image itself, and interprediction (gold modules) to detect and exploit similarities to other reconstituted images. Coefficients representing the compressed residual and the detected similarities from intra- or interprediction are collected and further compressed by entropy coding to produce the AVS2 bitstream.

Improvements offered in AVS2 over prior codecs include new intraprediction modes (e.g., chrominance derived from luminance) and long-term reference frames in interprediction modes. A significantly more detailed description of the AVS2 technologies and the advances in AVS2 can found in an IEEE paper published in 2015.

Where there are similarities between HEVC and AVS2, for example the overall processing flow, quad-tree partitioning, certain prediction modes, and motion vectors, transcoding from HEVC to AVS2 can be especially efficient[25].

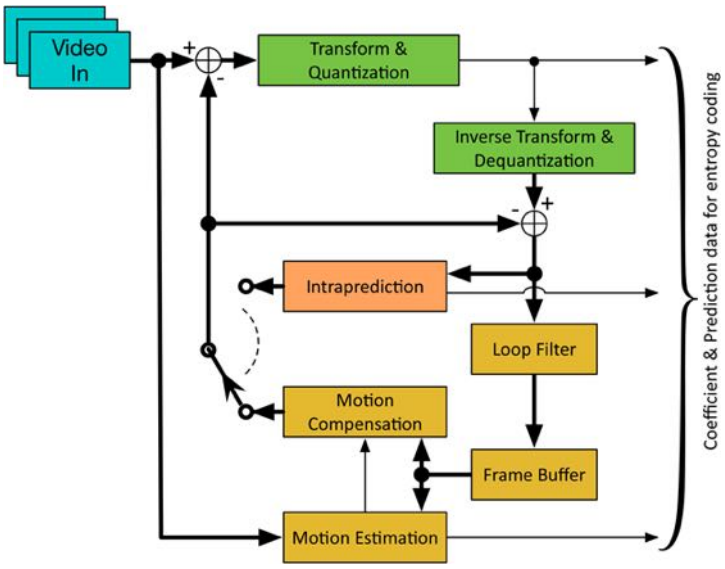

Figure 37 AVS2 coding framework

[25] Yucong CHEN, et al., "Efficient Software HEVC to AVS2 Transcoding", Information, 2016, 7, 53

End of Document