Ultra HD Forum Guidelines

Indigo Book – Ultra HD Technology Implementations

Ultra HD Forum

8377 Fremont Blvd., Suite 117

Fremont, CA 94538

UNITED STATES

VERSION 3.3.0

Sep 10, 2024

1. Foreword

The Ultra HD Forum Guidelines provides a holistic view of modern media systems, their mechanisms and workflows, and how those are impacted by the latest generation of improvements – the “Ultra HD” technologies, those that take media beyond the limits established at the start of this millennia, as characterized in large part by the video resolutions and the dynamic of “high definition” (i.e., ITU-R Rec. BT.709). The Forum considers Ultra HD to not only be any UHD media (i.e., 4K resolution, or higher), but also HD-resolution media with enhancements such as High Dynamic Range, Wide Color Gamut, etc. Ultra HD is a constellation of technologies that can provide significant improvements in media quality and audience experience. In addition, the Forum collaborates in promoting the understanding of the various deployments and delivery methods for Ultra HD media that continuously evolve around the world.

This work represents over eight years of collaborative effort by the membership of the Ultra HD Forum. The Ultra HD Forum’s Guideline books would not have been possible without the leadership of Jim DeFilippis, who represents Fraunhofer, and chair of our Guidelines Work Group with invaluable support from the co-chair, Pete Sellar of Xperi as well as technical assistance from Ian Nock of Fairmile West Consulting, chair of the Interop Working Group.

Our gratitude to all the companies listed in the Acknowledgments that have participated in this effort over the years and specifically to Nabajeet Barman (Brightcove), Elena Burdiel Pérez (Fraunhofer), Andrew Cotton (BBC), Jean Louis Diascorn (Harmonic), Richard Doherty (Dolby), Felix Nemirovsky (Dolby),Chris Johns (Sky UK), Katy Noland (BBC), Bill Redmann (InterDigital), Yuriy Reznik (Brightcove), Chris Seeger (Comcast/NBCUniversal), Adrian Murtaza (Fraunhofer) and Alessandro Travaglini (Fraunhofer).

This document, Indigo Book - *Ultra HD Technology Implementations,* is one of a series of books, referred to as the Rainbow Books, that compose the Ultra HD Forum Guidelines. If any of these terms sound unfamiliar, follow the link below to the Black Book. If a particular standard is of interest, links such as the one above are available to take you to the White Book, where references are collected.

The Rainbow Books are, in their entirety:

White Book	Guidelines Index and References
Red Book	Introduction to Ultra HD
Orange Book	Foundational Technologies for Ultra HD
Yellow Book	Beyond Foundational Technologies
Green Book	Ultra HD Distribution
Blue Book	Ultra HD Production and Post Production
Indigo Book	**Ultra HD Technology Implementations**
Violet Book	Real World Ultra HD
Black Book	Terms and Acronyms

Updates in this new version of the Ultra HD Forum Guidelines are described on the following page.

I hope you will enjoy reading today.

If you want to know more about Ultra HD, and join our discussions on how it can be deployed, I invite you to join the Ultra HD Forum. You can start by visiting our website: www.ultrahdforum.org.

Dr. Yasser Syed, President, Ultra HD Forum
Sept 2024

1.1 Changes from version 3.2 to 3.3

What's new in the September 2024 version of the UHDF Guidelines Indigo Book, *Ultra HD Technology Implementations* v3.3, edited by Bill Redmann..

The *Ultra HD Technology Implementations* is the sixth in the series of Rainbow Books on the Guidelines for Ultra HD. The scope and purpose of this book is to provide deeper background technical information on the range of Ultra HD technologies used in HDR, NGA, workflows, and distribution.

This version of the Indigo book has updates to Section 10.3, NBCU Single-Master Dual-Focused HDR-SDR Workflow Recommendations, including external references to NBCU production guidelines. Section 12.2, Brazilian Roadmap to UHD, includes new information on the adoption of ATSC 3.0 Physical Layer specifications and the demonstration of the Brazil 2.5 broadcast of the 2024 Paris Olympics.

This edition has updated references.

We hope this new format will be helpful in understanding Ultra HD technologies as well as planning for new or expanded UHD services.

Jim DeFilippis and Pete Sellar,

Guidelines Working Group Co-Chairs, Ultra HD Forum, Sept 2024

2. Acknowledgements

We would like to provide the acknowledgement to all the member companies, past and present, of the Ultra HD Forum who have contributed in some small or large part to the body of knowledge that has been contributed to the Guidelines Color Books, including the specific subject of this book.

ARRIS	ATEME	ATT DIRECTV
British Broadcasting Corporation	BBright	Beamr
Brightcove Inc.	Broadcom	B<>COM
Comcast / NBC Universal LLC	Comunicare Digitale	Content Armor
CTOIC	Dolby	DTG
Endeavor Streaming	Eurofins Digital Testing	Fairmile West
Fraunhofer IIS	Harmonic	Huawei Technologies
InterDigital	LG Electronics	Mediakind
MovieLabs	NAB	Nagra, Kudelski Group
NGCodec	Sky UK	Sony Corporation
Xperi	Technicolor SA	Verimatrix Inc.
V-Silicon		

3. 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 an 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.

4. Contents

5. Figures

6. Tables

List Of Tables

7. Introduction

This book is a compendium of technologies, methods and implementations of Ultra HD. Some of these are unique to a particular service or application. We have provided details of each, based on the best information available.

The Ultra HD Forum is providing this information for the benefit of the reader. The other Guideline books may make reference to these technologies and how they may be used to advance beyond the Foundational technologies for Ultra HD; however, the reader should not infer that these are recommendations or endorsements by the Ultra HD Forum of any of the included technologies, methods or implementations. We have included information on various methods and implementations being contemplated or deployed for UHD production and distribution.

8. Monographs on HDR

Several technologies exist that augment the HDR capabilities achievable with just an HDR transfer function such as PQ or HLG [5]. Each brings something different to the table and enables different use cases. The monographs on HDR are:

- Dolby Vision (Sec 8.1) emphasizes the ability to preserve artistic intent across a wide variety of distribution systems and consumer rendering environments, using dynamic metadata.
- Dual Layer (SHVC) (Sec 8.2) techniques provide a robust base layer and an optional, high-quality enhancement layer, given great bandwidth flexibility and hybrid delivery options.
- SL-HDR 7.3.SL-HDR (Sec 8.3.) offers a live, automatic, single-stream ability to deliver HDR productions to legacy displays in SDR, and use dynamic metadata to reconstruct the original for HDR displays.
- SL-HDR 7.4.SL-HDR (Sec 8.4.) facilitates frame-by-frame adaptation of a PQ signal to a receiving display's luminance capabilities by using real-time, automatically generated metadata.

8.1. 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 component signal format (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 [9], Recommendation ITU-R BT.2100 [5]
- Wide color gamut: Recommendation ITU-R BT.2100
- IC_TC_P: Recommendation ITU-R BT.2100
- Mastering display metadata: SMPTE ST 2086 [10] and CTA 861-I [31]
- Dynamic metadata: SMPTE ST 2094-10 [86] and CTA 861-I
- MaxFall/MaxCLL: CTA 861-I

8.1.1. 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 pre-recorded content in a file-based workflow).

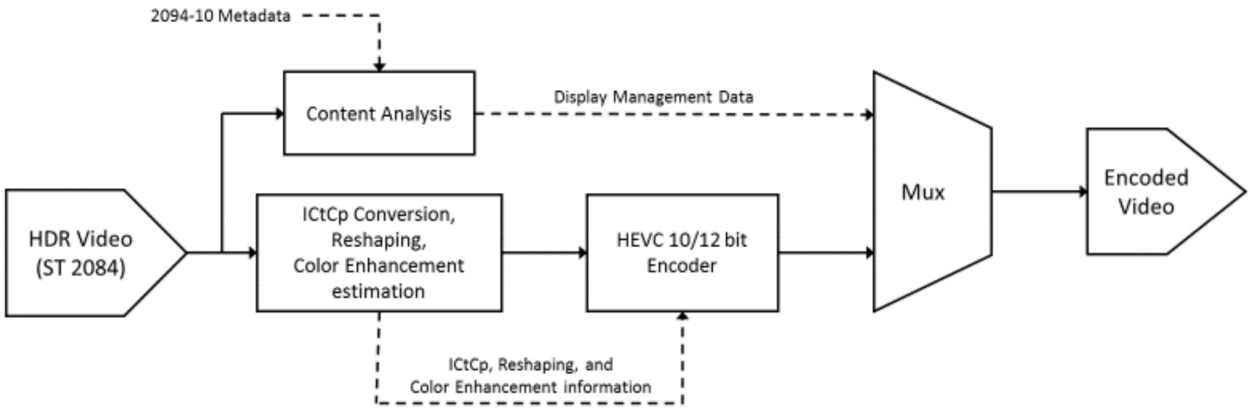

Figure 1. Encoder functional block diagram

If not natively in IC_TC_P signal format, it may be advantageous to convert the HDR video into IC_TC_P signal format. 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.

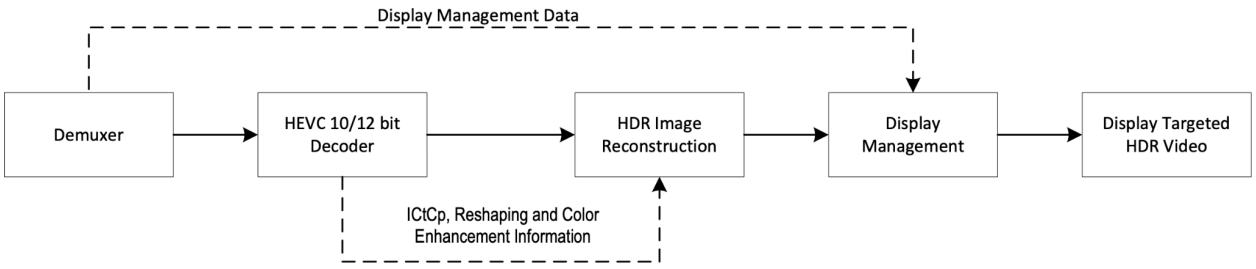

Figure 2. Decoder function block diagram

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 restored 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, display management is generally 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.

8.1.2. Dolby Vision Cross Compatibility

Dolby Vision constrained as described in these Guidelines is based on 2094-10 [86] metadata contained in SEI messages as described in Yellow Book Section 7.1.2 [Y01] and in ATSC A/341 [54], and when used in this method the streams are fully backwards compatible with HDR10 (assuming the underlying signal format remains YCbCr). A player receiving the stream can simply ignore the SMPTE 2094-10 dynamic metadata contained in the SEI messages and play the fully conforming HDR10 stream. In the case where the underlying signal format is ICtCp, the streams are generally not cross compatible, and the delivery system would need to deliver an alternate stream for non-Dolby Vision devices.

Note that Dolby Vision is also used in a wide variety of VOD services, and has a number of profiles to service this market (see Dolby Vision Profiles and Levels [90]). Profiles that rely on common underlying HDR10 streams (notably profile 8.1) can leverage the same cross stream compatibility advantage – the same stream can play back in HDR10 devices by simply discarding the dynamic metadata. In other profiles that are not cross compatible (notably profile 5, which is in wide use), service providers typically offer an alternate stream for non-Dolby Vision devices.

8.1.3. 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 [5], 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 [86] is the standardized mechanism to carry this metadata.

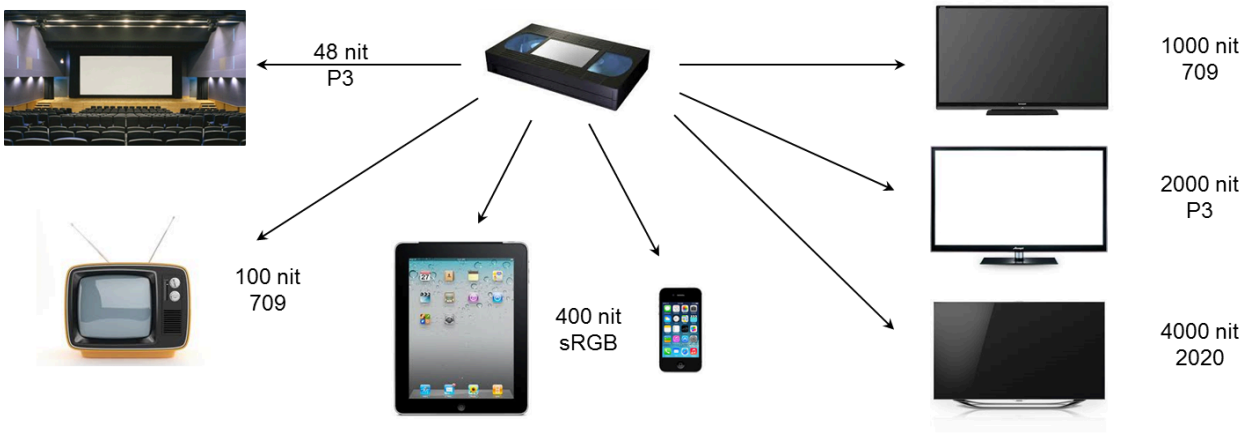

Figure 3. Example display device color volumes

While Dolby Vision works with the $Y'C'_BC'_R$ signal format 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 signal format model as defined in BT.2100. IC_TC_P isolates intensity from the color difference channels and may be a superior format in which to perform color volume mapping.

8.1.4. Dolby Vision in Broadcast

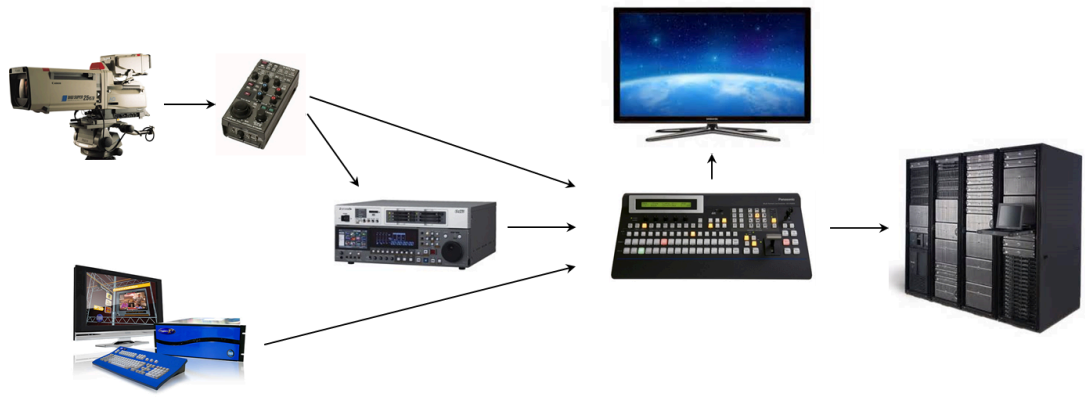

Figure 4. Example broadcast production facility components

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 5 shows a block diagram of the workflow in an HDR Broadcast facility using BT.2100 [5] 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 [86] metadata in real time in just prior to, or inside of, the emission encoder as shown (block labeled “HPU” in brown in Figure 39). 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. Note that the SEI message approach allows a production facility to utilize a common HDR10 bitstream, where one single stream is used for both HDR10 devices (which simply ignore the ST 2094-10 metadata) and Dolby Vision devices that correctly utilize the included metadata.

SMPTE ST 2110-40 [47] standardizes the carriage of HDR metadata via ANC packets in both SDI and IP interfaces. Once completed, this standard will allow the ST 2094-10 [86] 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 tan 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.

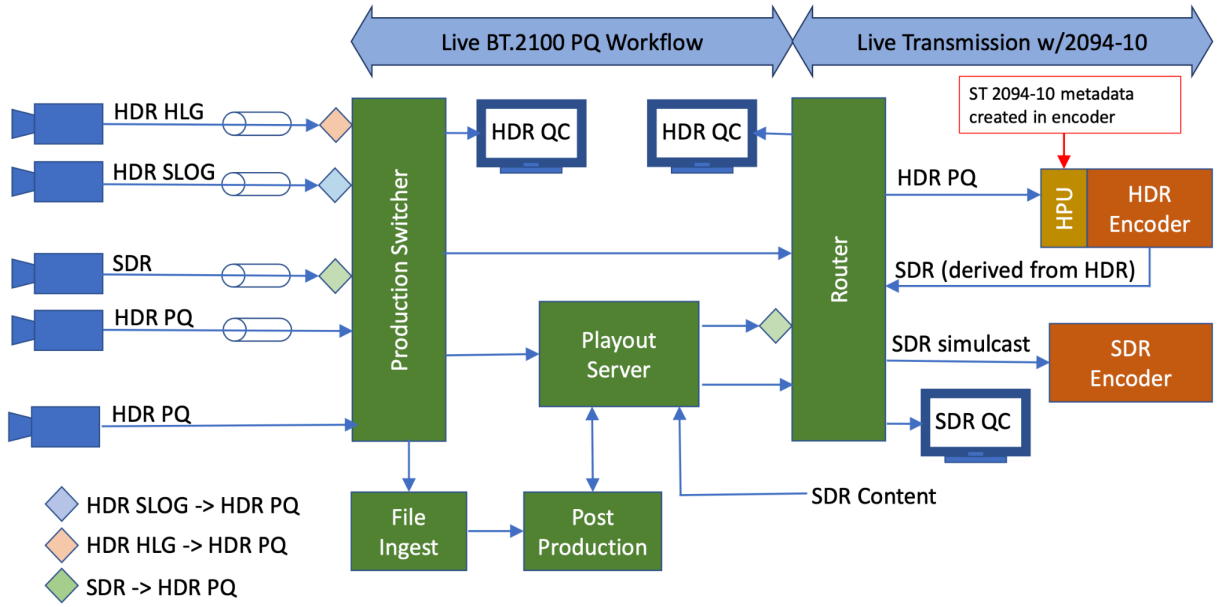

Figure 5. HDR broadcast production facility with BT.2100 PQ workflow- transition phase

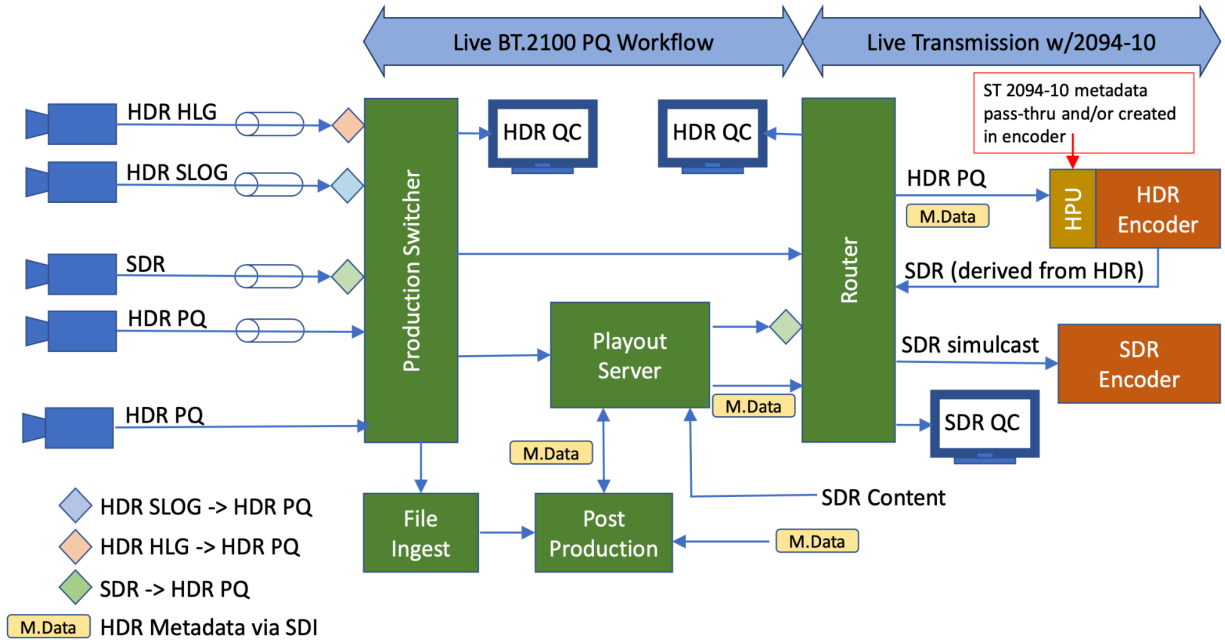

Figure 6. HDR broadcast production facility with BT.2100 PQ workflow- SDI metadata

8.2. Dual Layer - Scalable High-Efficiency Video Coding (SHVC)

Scalable High-Efficiency Video Coding (SHVC) is specified in Annex H of the HEVC specification [69]. 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 DTT 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 DTT 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 Section 5.1.6 of ATSC A/300 [51] , 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 ATSC A/322 [52] 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 system colorimetry, transfer function, bit depth, and frame rate, for this

document, only differences in spatial resolution and quality are supported. In addition, while SHVC permits use of multiple ELs, only a single EL is used herein.

The combined BL and ELs should provide Foundation Ultra HD 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 Foundation Ultra HD 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 Foundation Ultra HD content.

As an alternative, the base and ELs may be provided in an SDR format, which with metadata (see TS 103 433-1 [33]) 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.

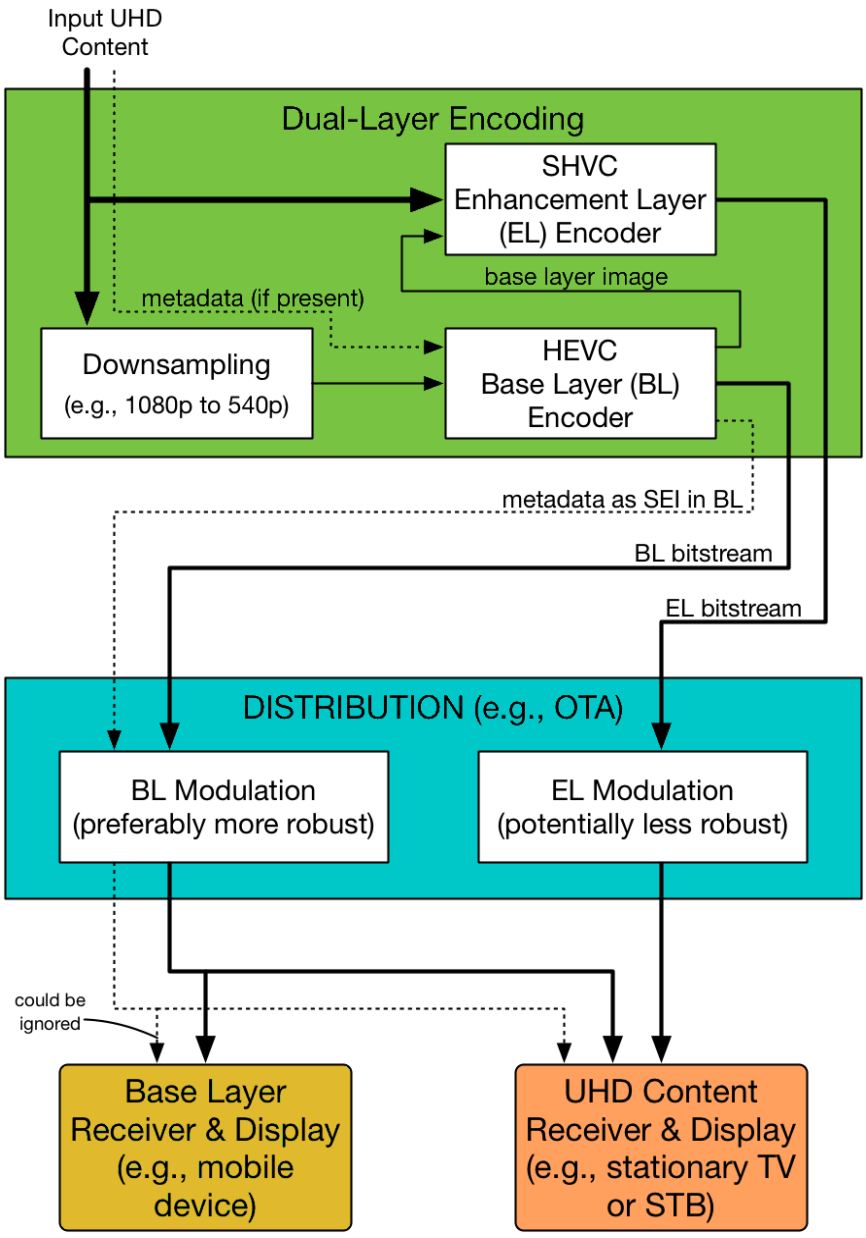

Figure 7. Example dual-layer encoding and distribution

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-HDR1 [33] is well-suited, given that its

error-concealment process (described in Annex F of the SL-HDR standard) 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 (DTT) 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 DTT as shown, while the EL would be distributed via broadband connection. 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. Thus, while not specifically described herein, dual layer distribution is suitable for OTT distribution as well, both for VOD and linear programs.

8.3. SL-HDR1

As pointed out in Section 8.4, ETSI TS 103 433-1 [33] 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 ITU-T H.222 [1]) and may optionally deliver SDR/BT.2020 [3] as the down-conversion target.

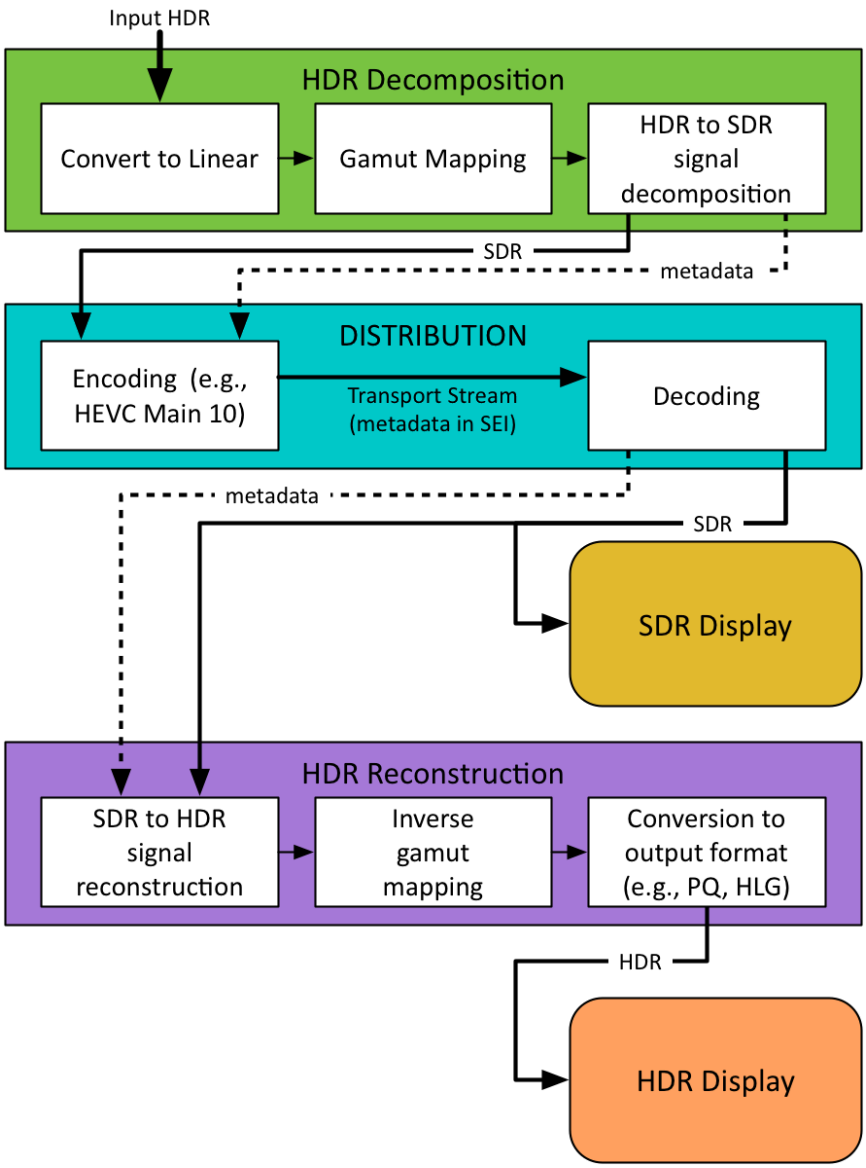

Figure 8. SL-HDR processing, distribution, reconstruction, and presentation

This ETSI specification additionally specifies a mechanism for generating an SL-HDR information SEI message (defined in Annex A.2 of ETSI TS 103 433 [33]) 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 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 ETSI TS 103 433-1, 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.

Upon receipt of an SL-HDR1 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.

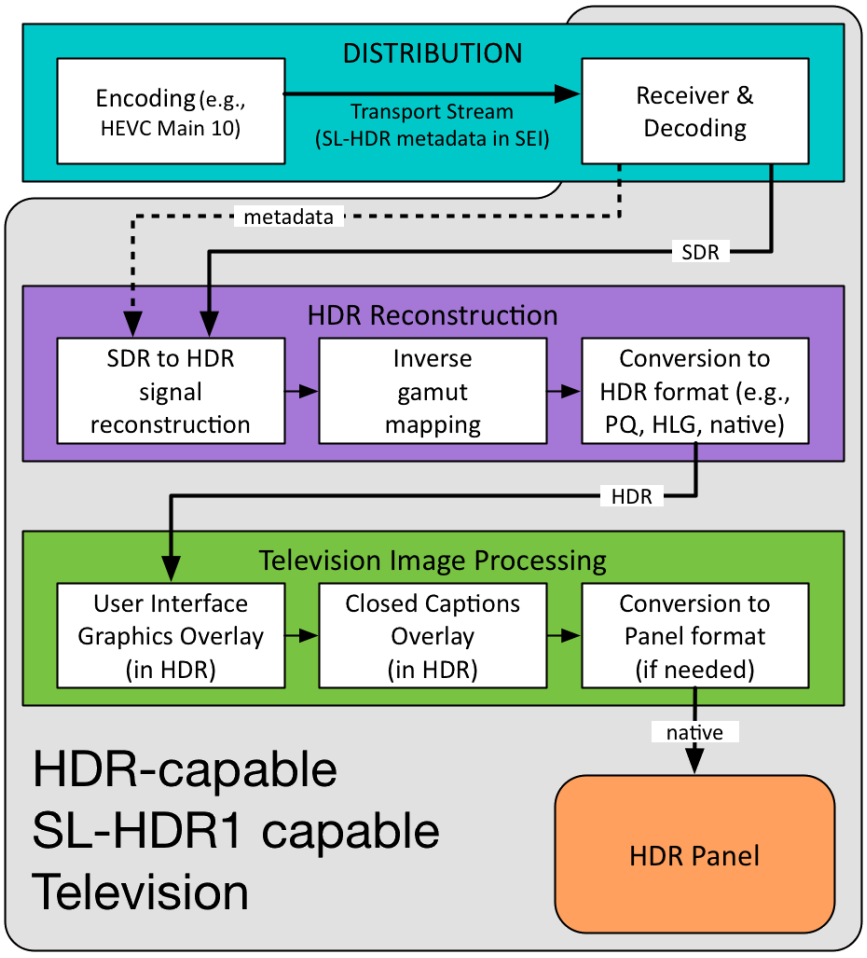

Figure 9. Direct reception of SL-HDR signal by an SL-HDR1 capable television

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 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.

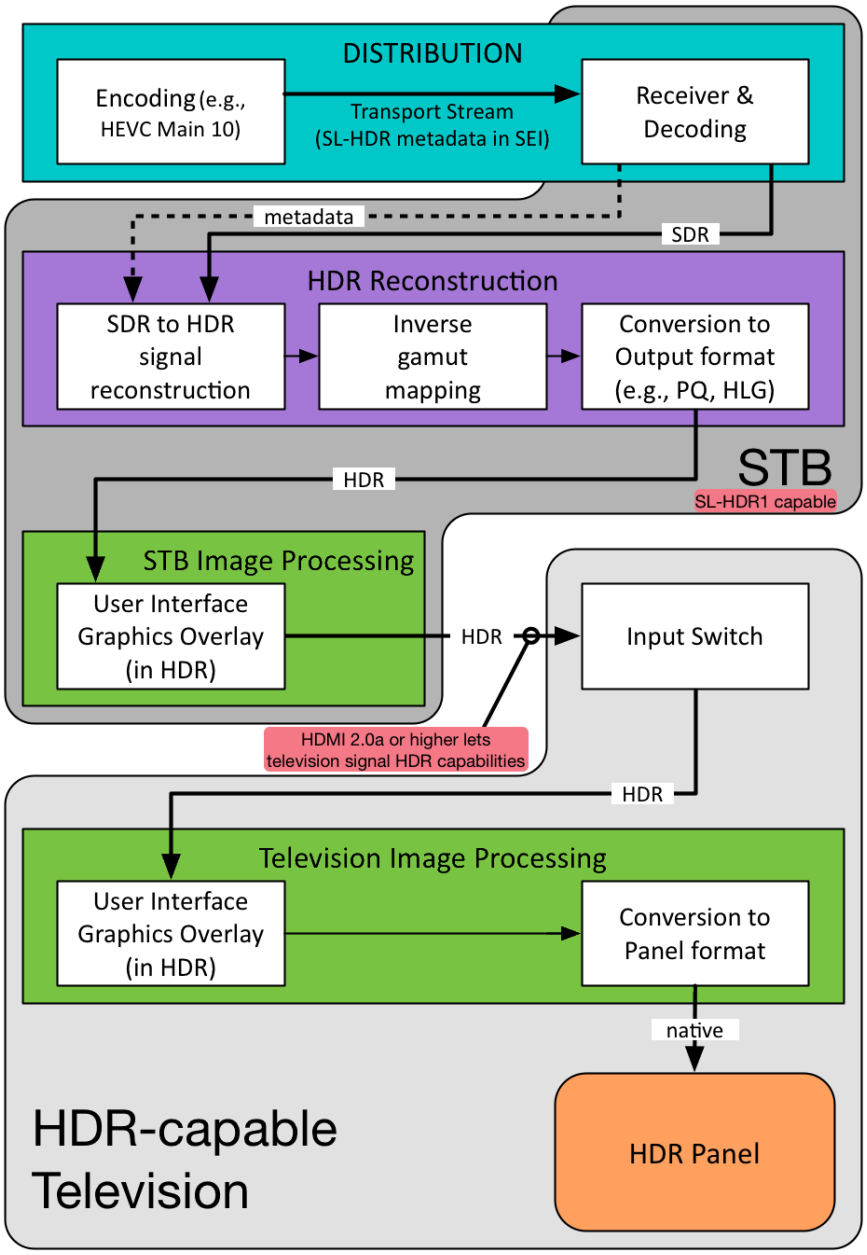

Figure 10. STB processing of SL-HDR signals for an HDR-capable television

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 [2].

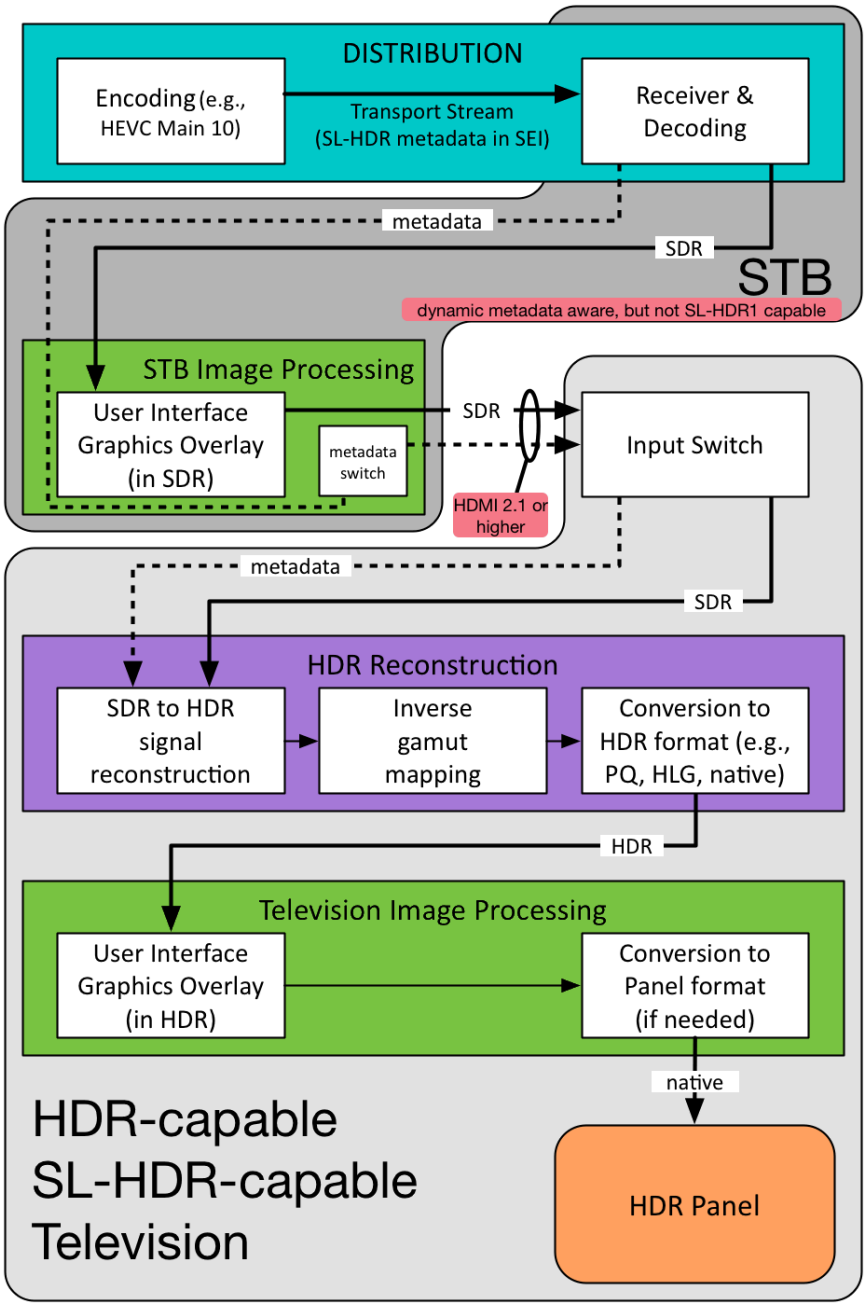

Figure 11. STB passing SL-HDR to an SL-HDR1 capable television

STBs will be used as DTT conversion boxes for televisions unable to receive appropriate DTT 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 ETSI TS 103 433-1 [33]) 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.

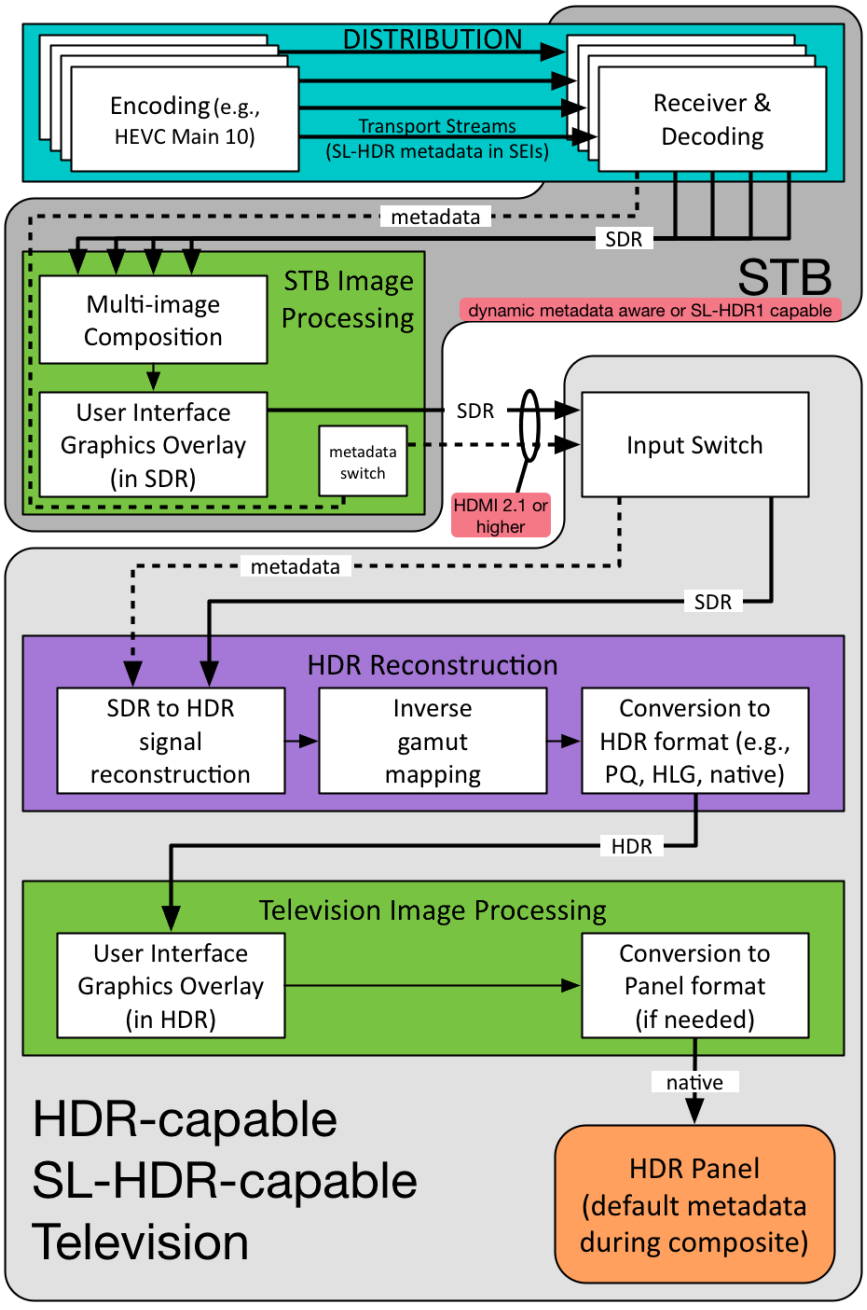

Figure 12. Multiple SL-HDR channels received and composited in SDR by an STB

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)

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 [33]. 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 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 [86]) 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.

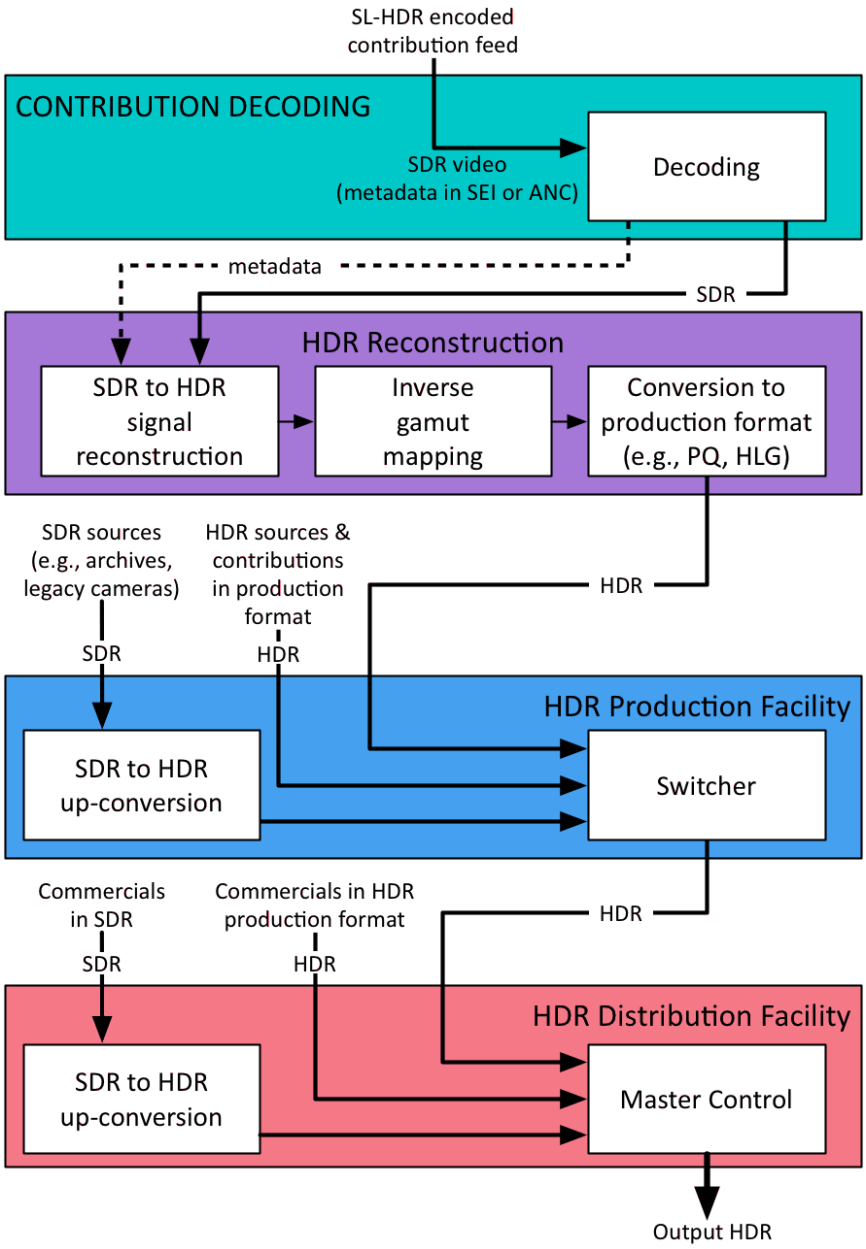

Figure 13. SL-HDR as a contribution feed to an HDR facility

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 TS 103 433-1[33] 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.

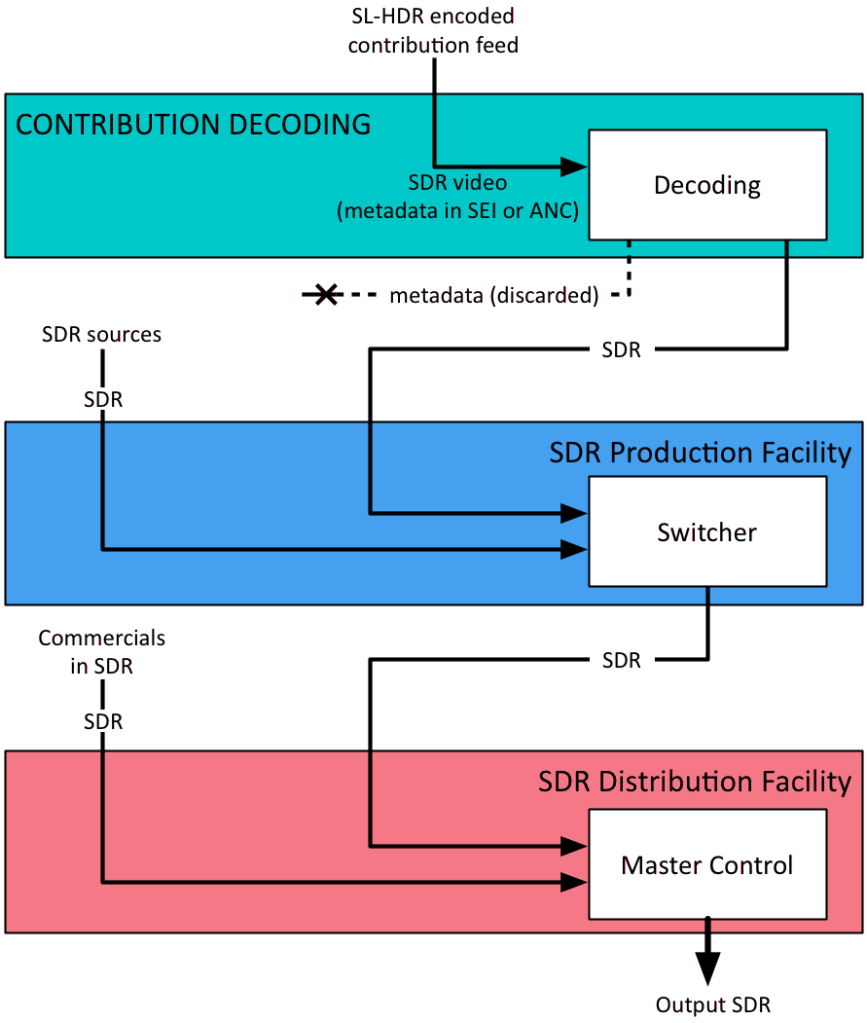

Figure 14. SL-HDR as a contribution feed to an SDR facility

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 ETSI TS 103 433-1 [33] 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.

8.4. SL-HDR2

SL-HDR2 is an automatically generated dynamic color volume transform metadata for HDR/WCG content that may be provided with a PQ signal to facilitate adaptation by a consumer electronic device of an HDR/WCG content to a presentation display having a different peak luminance than the display on which the content was originally mastered.

Generation and application of SL-HDR2 metadata is specified in ETSI TS 103 433-2 [34]. Typically, SL-HDR2 metadata is generated immediately prior to, or as a part of, distribution encoding, as shown in Figure 15, but SL-HDR2 metadata can also be generated upstream of the distribution encoder, e.g., as an encoding pre-process, and carried to the encoder as ST 2108-1 ANC messages [48] via SDI, or via IP using ST 2110-40 [47], or stored in file-based production infrastructures.

SL-HDR2 metadata may be carried on CE digital interfaces (e.g., HDMI) having dynamic metadata support as described in Annex G of ETSI TS 103 433-2 [34] and is optionally applied by consumer electronic devices before or as the content is displayed.

The SL-HDR information SEI message used to carry SL-HDR2 metadata is as specified in ETSI TS 103 433-1 (in Annex A.2 of [33]), but with the constraints specified in ETSI TS 103 433-2.

Figure 15 represents a typical use case of SL-HDR2 being used for distribution of HDR content. The input HDR content is analyzed to produce the SL-HDR2 metadata and is then converted to PQ format.

For distribution, the metadata is embedded in the HEVC bitstream as SL-HDR information SEI messages, defined in ETSI TS 103 433-1], which accompany the PQ encoded HDR/WCG content. The resulting stream may be used for either primary or final distribution. Whereas the SDR signal resulting from the down-conversion was the signal distributed with SL-HDR1, with SL-HDR2 it is the master PQ signal that is distributed. As a result, a legacy HDR display can receive the PQ signal and operate successfully without reference to the SL-HDR2 metadata. However, when recognized, the SL-HDR2 metadata enables an optional adaptation, by downstream recipients, of the HDR/WCG content for a particular presentation display.

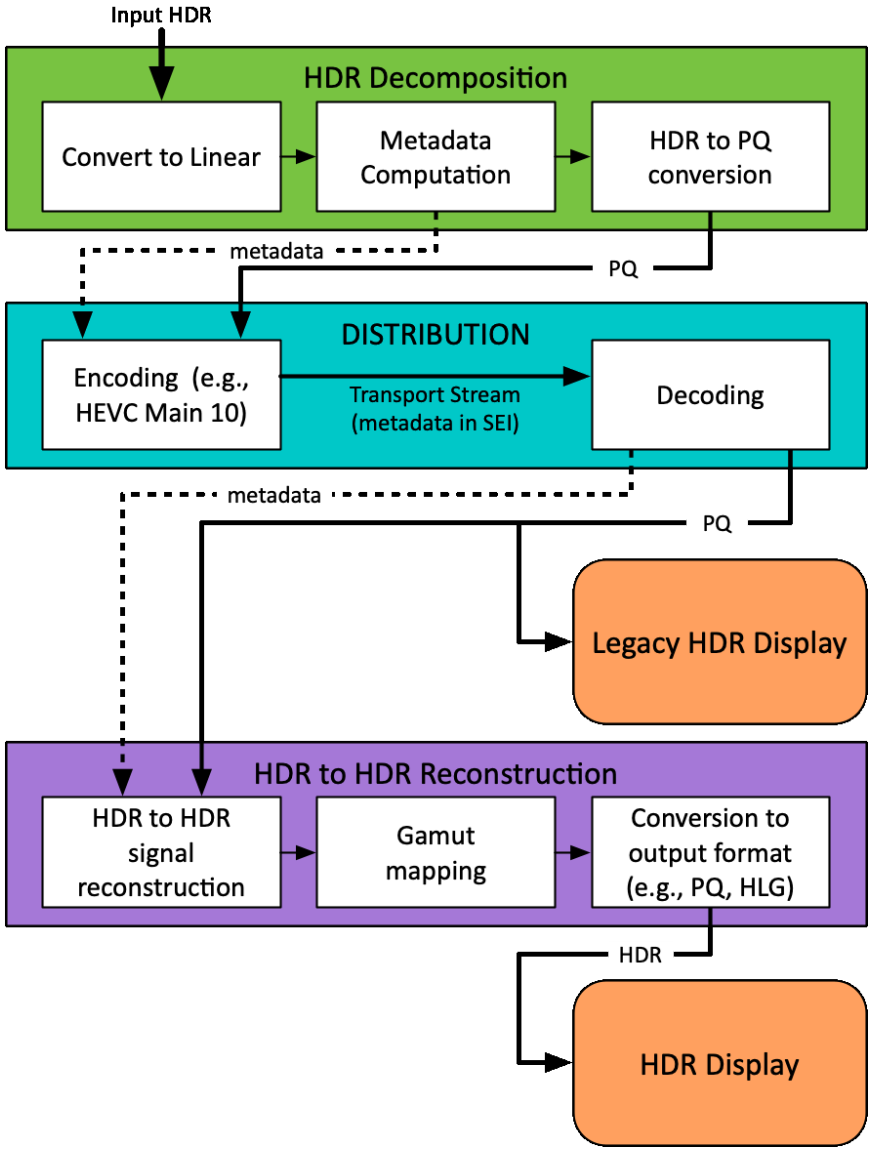

Figure 15. SL-HDR2 processing, distribution, reconstruction, for HDR presentation

Upon receipt of an SL-HDR distribution, the HDR/WCG signal and metadata may be used by legacy HDR devices by using the PQ format for presentation of the image and ignoring the metadata, as illustrated by the legacy HDR display in Figure 15 but if received by a decoder that recognizes the metadata, the metadata may be used by the decoder to reconstruct the image as appropriate for the peak brightness and transfer function of the presentation display to which it is connected, with the reconstruction taking place as shown by the HDR to HDR and HDR to

SDR reconstruction blocks in Figure 15 and Figure 16, respectively. An optional Gamut Mapping may be used during the reconstruction process if the presentation display is only able to support BT.709 images.

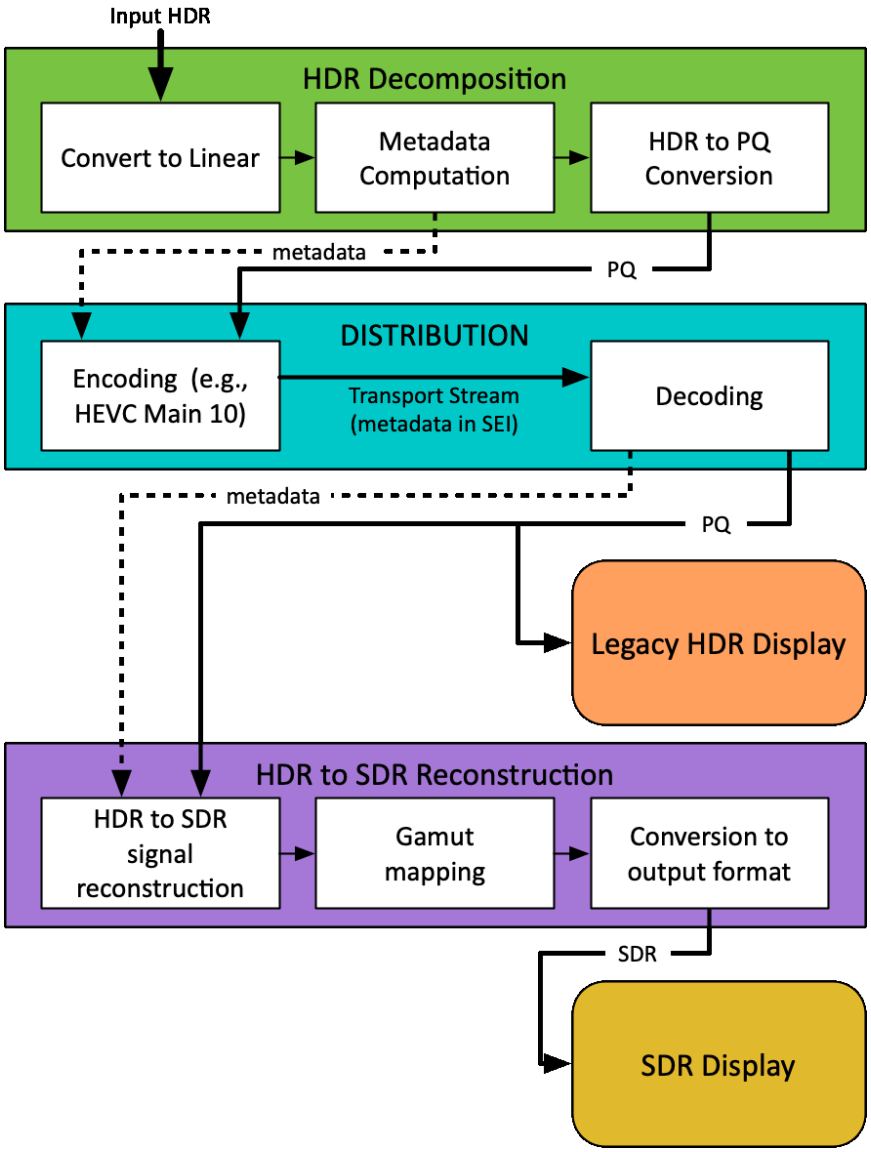

Figure 16. SL-HDR2 processing, distribution, reconstruction, and SDR presentation

The capability of this presentation display adaptation extends all the way to a downstream recipient having an SDR display, as shown in Figure 16 where the processing block labeled

HDR to SDR Reconstruction can also be used when redistributing or retransmitting to a legacy SDR network.

The HDR to HDR Reconstruction process of Figure 15, and HDR to SDR Reconstruction process of Figure 14 are considerably lighter weight computationally than is the HDR Decomposition process, and as such is well suited to consumer premises equipment, and widely available for inclusion in consumer electronic hardware, both in STBs and displays.

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 17, 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 PQ picture.

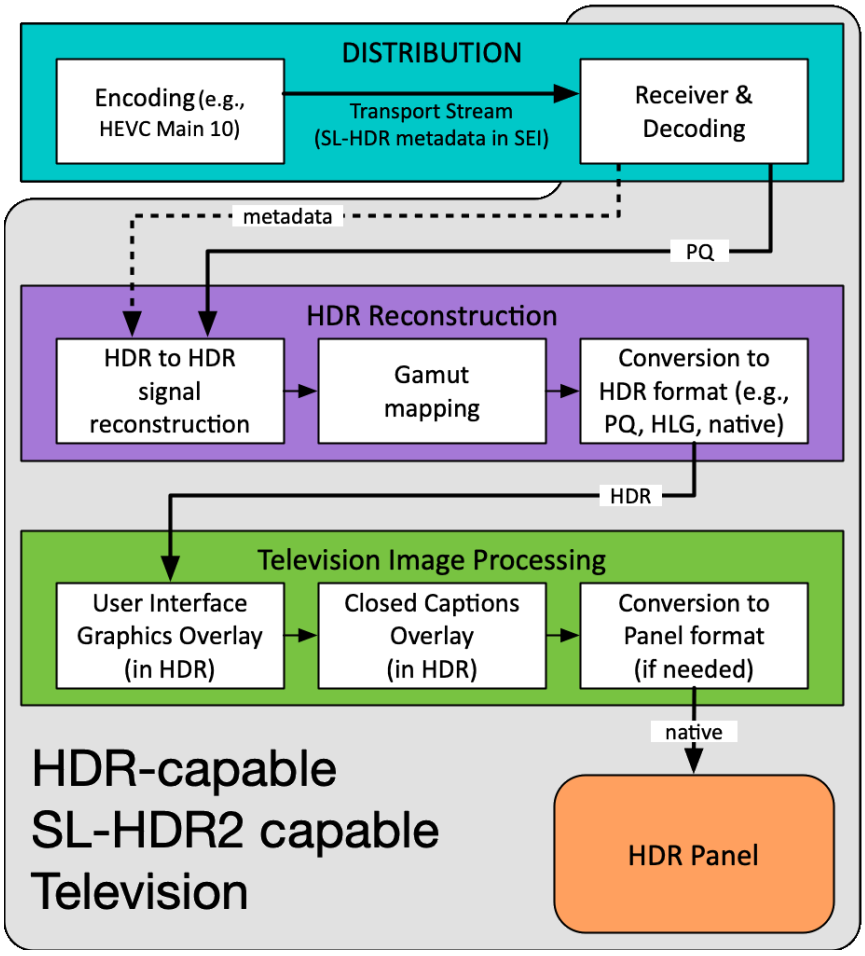

Figure 17. Direct reception of SL-HDR signal by an SL-HDR2 capable television

STBs will be used as DTT conversion boxes for televisions unable to receive appropriate DTT 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 18, then the STB may query the interface with the display device (e.g., via HDMI 2.0a or higher, using the signaling described in CTA 861-I [31]) to determine the display capabilities (HDR and corresponding peak luminance or SDR, gamut capabilities) that will serve in conjunction with the metadata to reconstruct, in an appropriate gamut and with an appropriate peak luminance, the HDR or SDR 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.

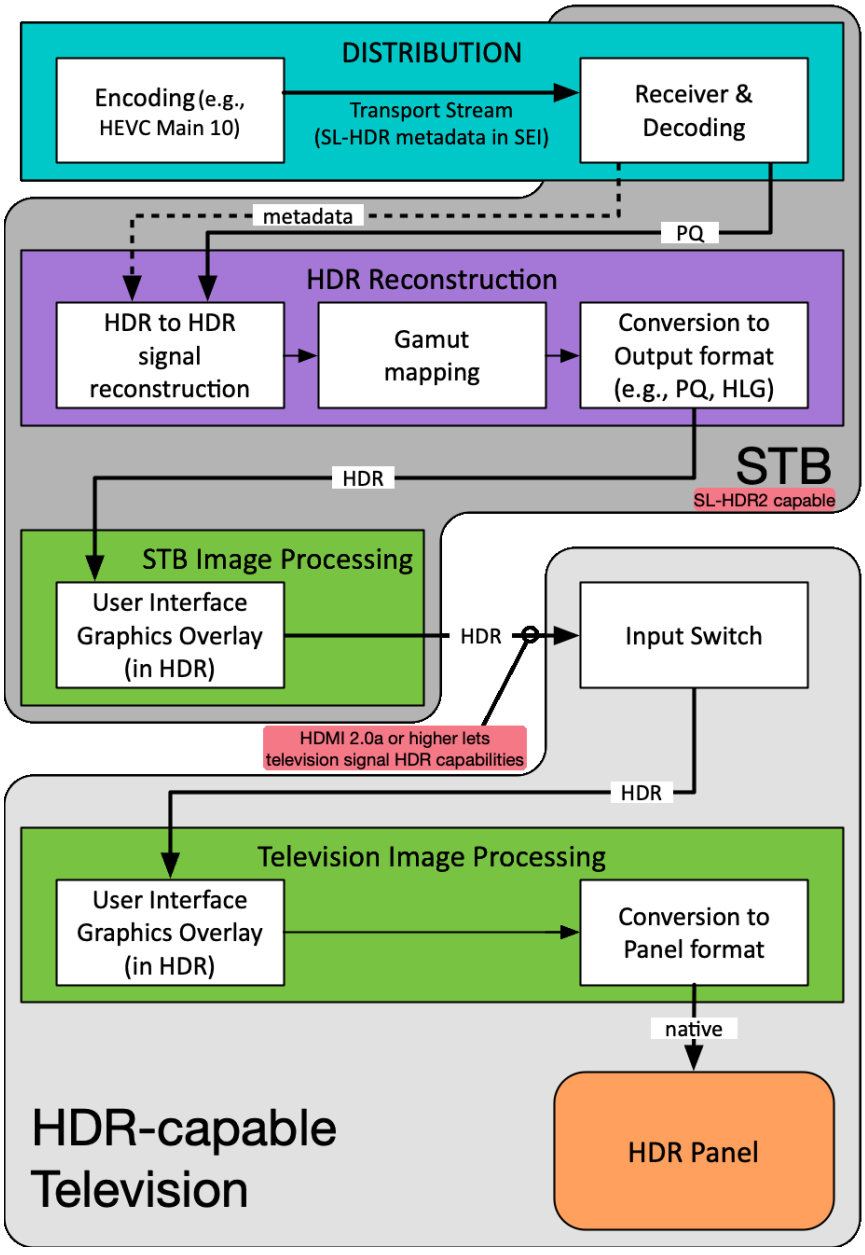

Figure 18. STB processing of SL-HDR signals for an HDR-capable television

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 21.

If, as in Figure 19, an STB is not capable of using the SL-HDR2 information messages to implement display adaptation of the PQ video, but the display has indicated (here, via HDMI 2.1 or higher, signaled as in CTA 861-I [31]) that such information would be meaningful, then the STB may pass the SL-HDR information to the display in conjunction with the PQ video, enabling the television to reconstruct the HDR/WCG image.

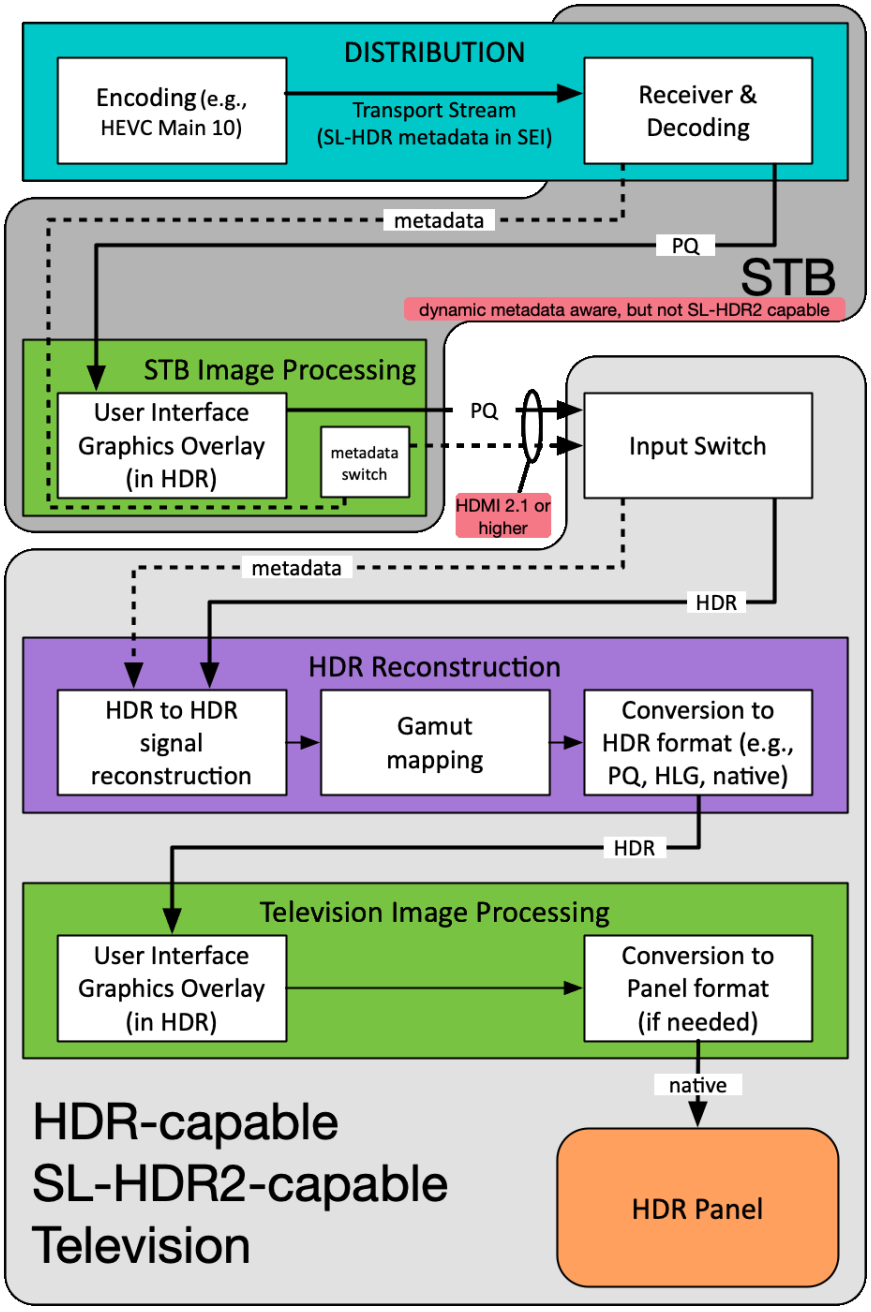

Figure 19. STB passing SL-HDR to an SL-HDR2 capable television

In this scenario, if the STB were to first overlay HDR graphics (e.g., captions, user interface or EPG) before passing the HDR video along to the display, the STB has two options, illustrated as

the “metadata switch” in Figure 19, 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 ETSI TS 103 433-2 [34]. 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

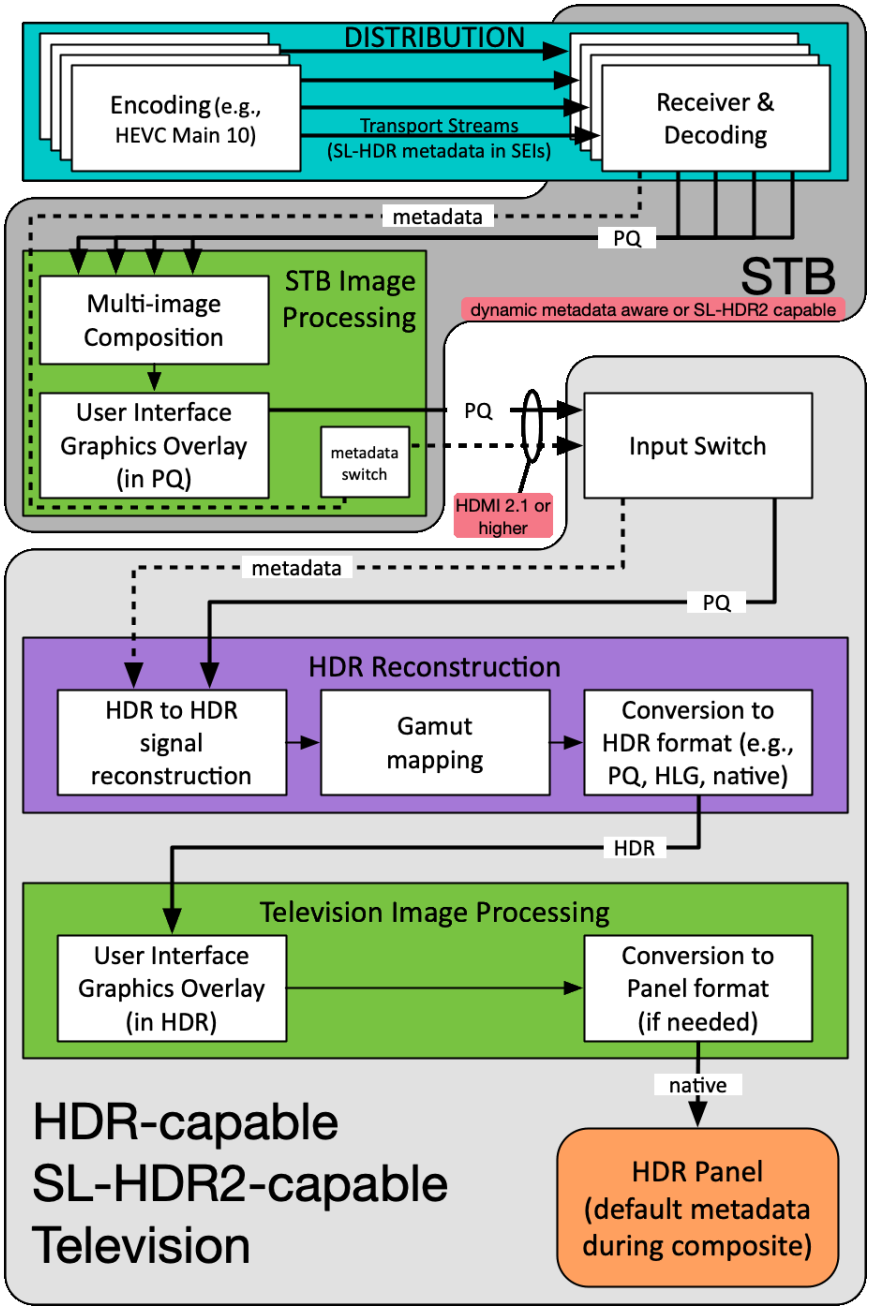

Figure 20. Multiple SL-HDR channels received and composited in HDR by an STB

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 multiplex of 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 20. 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-HDR2 capable, then each of the channels could be individually reconstructed with the corresponding metadata to a display-appropriate, common format (whether HDR or even SDR), with the compositing taking place in the common format and the resulting composite 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 PQ image, which is then presented by the display. Thus, in any case, the HDR image may be presented even if the metadata does not reach the decoder or cannot be interpreted for any reason.

Figure 15 shows HDR formatting 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 formatting is preferably integrated into the encoder fed by the HDR signal but, in the alternative, the HDR formatting may be performed by a preprocessor from which the resulting PQ 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 SMPTE ST 2108 [48]) of the HDR 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 21 and Figure 22.

Where valuable to support the needs of a particular workflow, a different approach may be taken, in which the HDR formatting takes place earlier and relies on the HDR video signal and metadata being carried within the broadcast facility. In this workflow, the HDR signal is usable by HDR 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 for adaptation of the HDR signal. Note that an HDR-based broadcast facility may still 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. Similarly, the distribution may be a down-converted HDR version (e.g., 1000 cd/m^2 while the original stream is 4000 cd/m^2) as the distributor may know the display capabilities of the client base or their equipment (STB) or may have low confidence in unaided down-conversion processes in consumer equipment.

The workflow for an HDR-ready affiliate receiving an HDR video with SL-HDR metadata as a contribution feed is shown in Figure 21.

In HDR-based production and distribution facilities, such as shown in the example of Figure 21, 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. Accordingly, the SL-HDR metadata carried in the Input HDR signal can be discarded. Alternatively, where metadata may be carried through equipment and between systems, e.g., the switcher, HDR formats requiring metadata, such as HDR10, may be used.

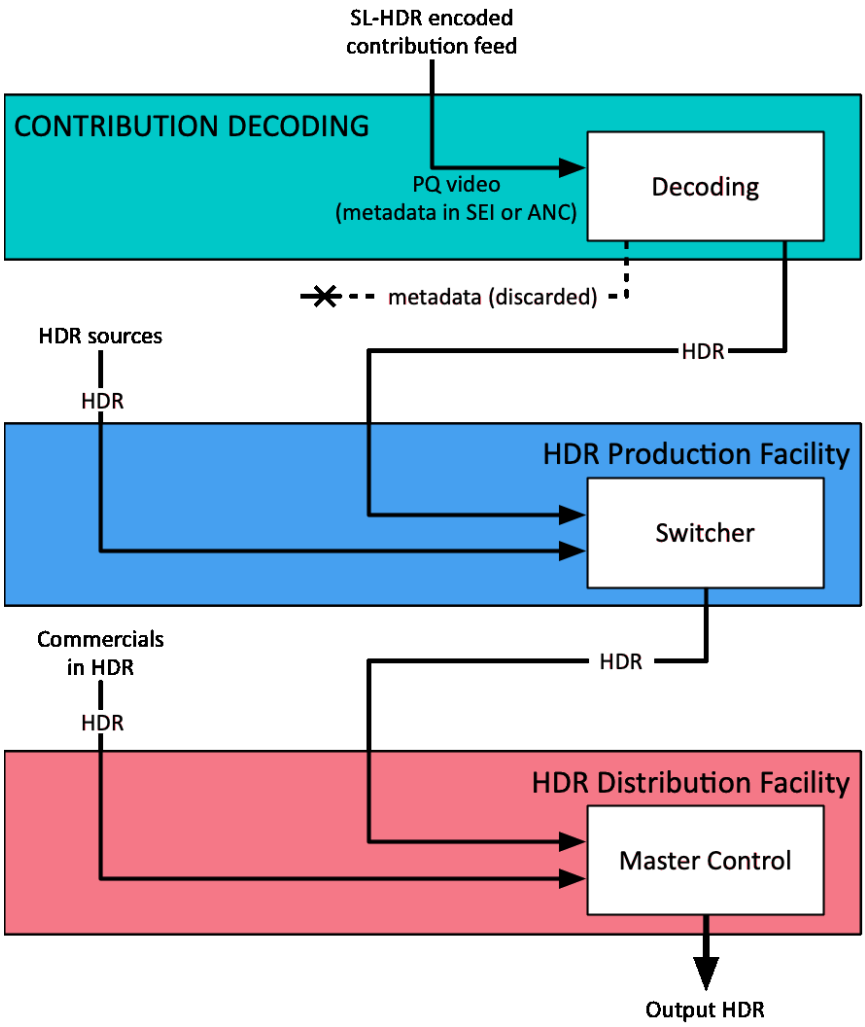

Figure 21. SL-HDR as a contribution feed to an HDR facility

In an HDR-based facility, the output HDR is complete immediately prior to the emission encode. As shown in Figure 15, this HDR signal (shown there as the “Input HDR”) is passed through the HDR formatting and encode processes. With this architecture, a distribution facility has available the signals to distribute to a channel that carries HDR video as PQ with SL-HDR metadata.

Figure 22 shows an SDR-based affiliate receiving an SL-HDR encoded contribution feed. Upon decode, only HDR video is available, though with the SL-HDR information carried in the contribution feed the SDR Reconstruction process will produce the SDR video. This mode of operation is considered suitable for those downstream affiliates or markets that will be late to convert to HDR operation. The decoding block and the HDR to SDR Reconstruction block

resemble the like-named blocks in Figure 16, with one potential exception: In Figure 22, the Gamut mapping block should use the forward gamut mapping described in Annex D of TS 103 433-1 [33].

In the case of distributions to an MVPD, distribution as HDR with SL-HDR information for HDR to SDR Reconstruction is well suited, because the HDR decomposition process shown in Figure 15 and detailed in Annex C of SMPTE ST 2094-1 [86] is performed only once, by professional equipment, and is not subject to variation in preferences that might be set on the receiving equipment. This can be used to ensure a consistent presentation to all affiliates receiving the contribution. Further, performance of such a down-conversion more consistently provides a quality presentation to the SDR customers.

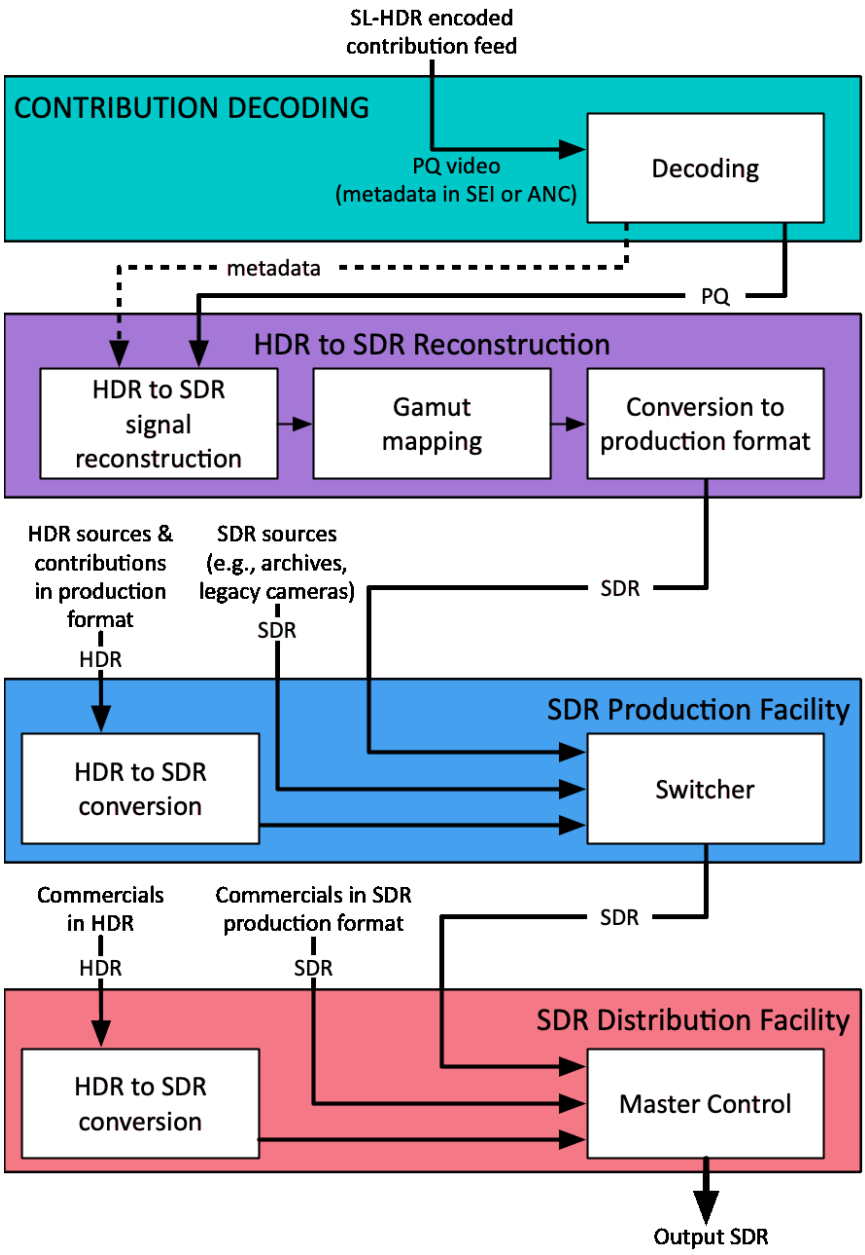

Figure 22. SL-HDR as a contribution feed to an SDR facility

9. Monographs on NGA

Complementing the visual enhancements that Ultra HD will bring to consumers, Next-Generation Audio (NGA) provides compelling new audio experiences. For an overview of NGA experiences and use cases, see the Yellow Book section 7.2 [Y02] on NGA. The monographs presented here detail each of three major NGA systems:

- Dolby AC (Section 9.1) is an audio system designed to address the current and future needs of next-generation video and audio entertainment services, including broadcast and Internet streaming.
- DTS-UHD (Section 9.2) is primarily designed to support audio objects that can represent a channel-based presentation, an Ambisonic sound field or audio objects used in Object-based Audio (OBA)
- MPEG-H Audio (Section 9.3) offers its object-based concept for delivering separate audio elements with metadata within one audio stream to enable personalization and universal delivery.

9.1. Dolby AC-4

AC-4 is an 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 [65] and adopted by Digital Video Broadcasting (DVB) in TS 101 154 [63] 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 [56]. AC-4 is specified in the ATSC 3.0 next-generation broadcast standard (A/342 [55]) and has been selected for use in North America (U.S., Canada and Mexico) as described in A/300 [51].

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 23, 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.

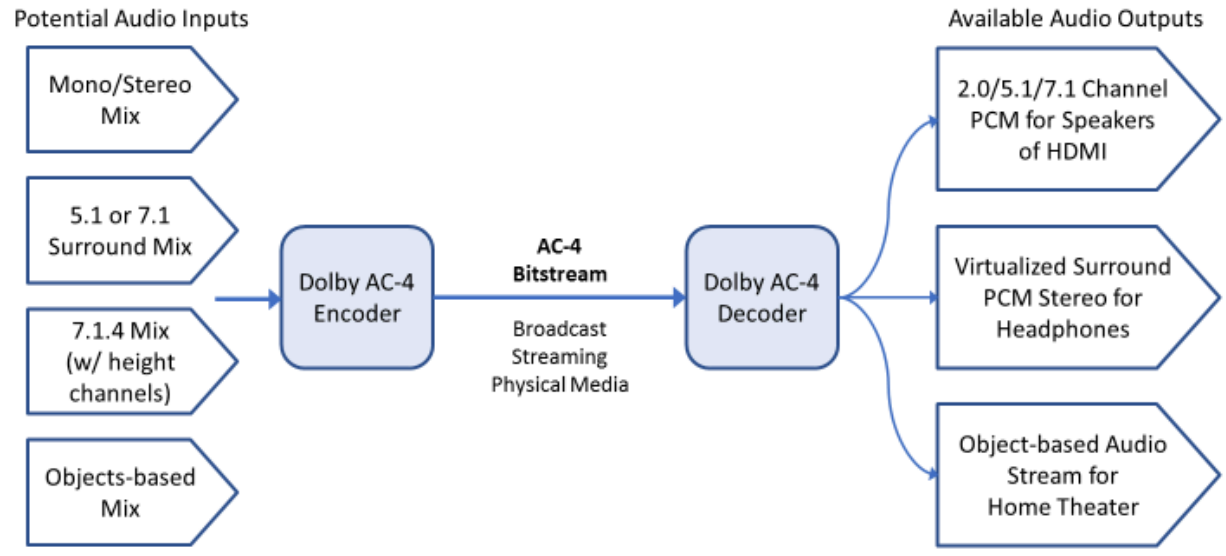

Figure 23. AC-4 Audio system chain

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 1 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 1. 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

9.1.1. 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 provides 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 24). 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.

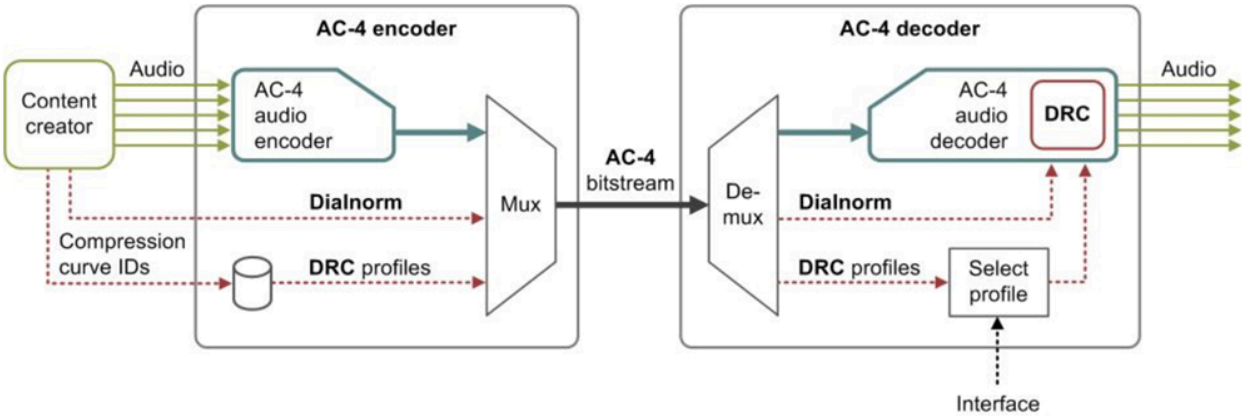

Figure 24. AC-4 DRC generation and application

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 2.

Table 2. 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

9.1.2. 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.

9.1.3. Backward Compatibility

Dolby Atmos audio programs can be encoded using the AC-4 [65] codec or the E-AC-3+JOC [35] codec. When Atmos is used with E-AC-3+JOC streams, backward compatibility is provided for existing non-Atmos E-AC-3 [29] decoders. See Section 11.5 of the Violet Book [V01] 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, it will correctly provide a downmix to those legacy formats.

9.1.4. 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.

9.1.5. Overview of Immersive Program Metadata and rendering

9.1.5.1. Object-based Audio Rendering

Object audio renderers also include control over the perceived object size (see “object width” metadata parameter in Section 9.1.7.3.), 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 25).

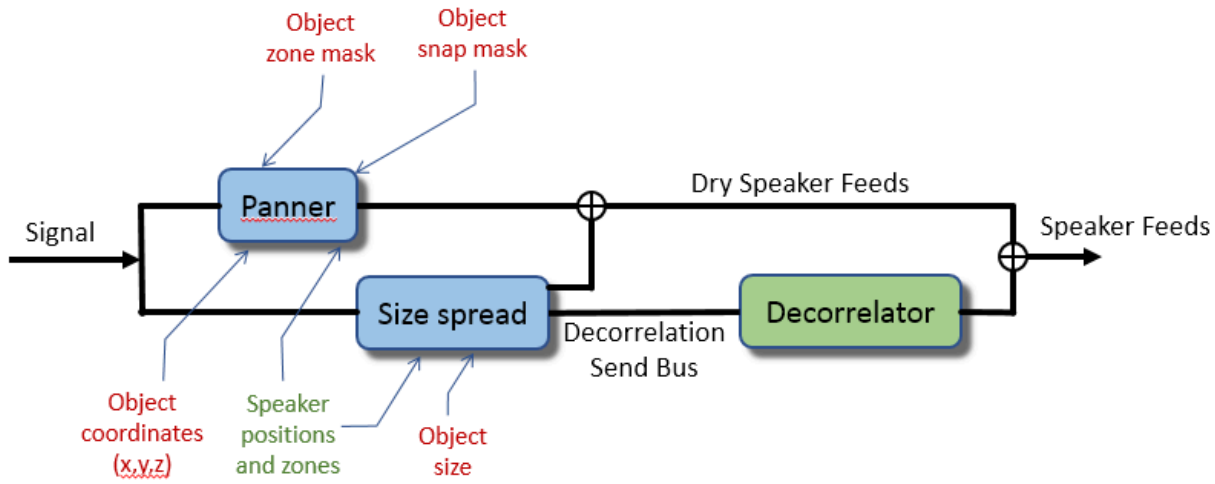

Figure 25. Object-based audio renderer

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.

At the core of rendering are pan and spread operators, each executing a panning algorithm (see Figure 25) 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 26 illustrates how different speakers may be utilized among various rendering (panning) algorithms to place an object's perceived position in the playback environment.

Directional pairwise panning (DIR) (see Figure 26-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 26-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.

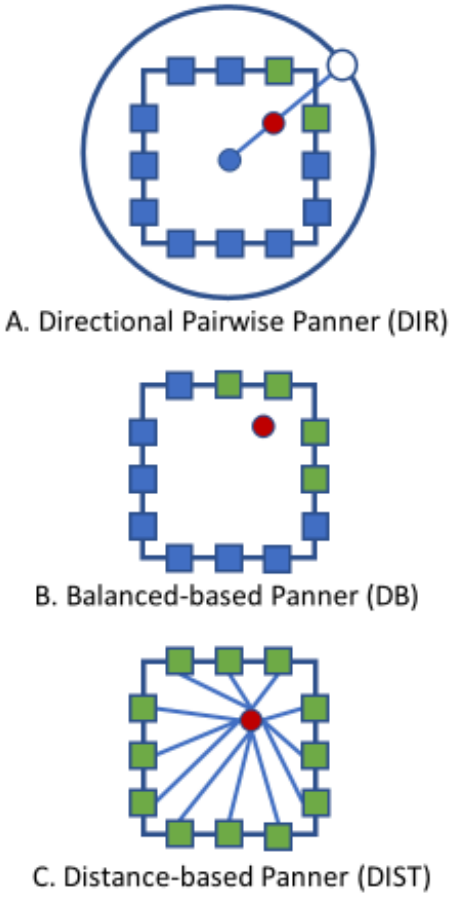

Figure 26. Common panning algorithms

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 26-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 artifacts, 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.

9.1.5.2. 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 25. 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 25). 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.

9.1.6. 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.

9.1.6.1. 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, at 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.

9.1.7. 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 [73]/ ADM formats [72]) and in serialized form for real-time workflows and interconnects utilizing SMPTE ST 337 [36] formatting/framing.

9.1.7.1. 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 [24]. A value of “0x2” indicates the author was adhering to EBU R 128 [62]. 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 [65].

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 [37]. In ATSC regions, this parameter would typically be -24.0 LKFS for short-form content as per ATSC A/85 [24]. 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.

max_loudstrm3s - This parameter indicates the maximum short-term loudness of the audio program measured per ITU-R BS.1771 [71].

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

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 [73] / ADM [72] 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 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 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.

9.1.7.2. 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.

9.1.7.3. 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 [65] are used for either an enhanced playback application 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.

9.1.8. 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.

9.1.8.1. 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 coexist to fully describe these experiences. The EBU Audio Definition Mode [61] has provided the foundation for the development of an international recommendation within ITU-R WP6B, which produced the ITU-R Audio Definition Model (ADM) [72].

The ITU-R ADM 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 [73] 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 [74]. 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.

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

While the transition to use ITU-R ADM[72]. for immersive audio file-based workflows is well underway, a similar need was identified for real-time audio production. A key requirement was for lossless conversion when transitioning between live and file-based assets. It was desirable to try to take advantage of the large amount of time invested in the development of ITU-R ADM. The result has been the creation of Serial ADM as specified in ITU-R BS.2125 [123].

Carriage of Serial ADM has been standardized in both legacy and IP based containers. SMPTE 2116 [124] defines Serial ADM as a SMPTE 337 [36] (see Figure 27) payload and enables carriage over AES3 and HD-SDI. Serial ADM can be used in a modern IP audio production facility today using SMPTE 2110-31 [126] which specifies the transparent transport of AES3 signals. This still relies on SMPTE 2116 and 337 for payload packaging, making it ideal for mixed environments where conversion between IP and legacy is common. Project work is underway in SMPTE to define a new IP native format for the carriage of metadata including Serial ADM. This will overcome the shortcomings of SMPTE ST 2110-31 which has its underpinnings in AES3 and SMPTE 337. It provides simpler packaging of the metadata and explicit signaling at the session layer.

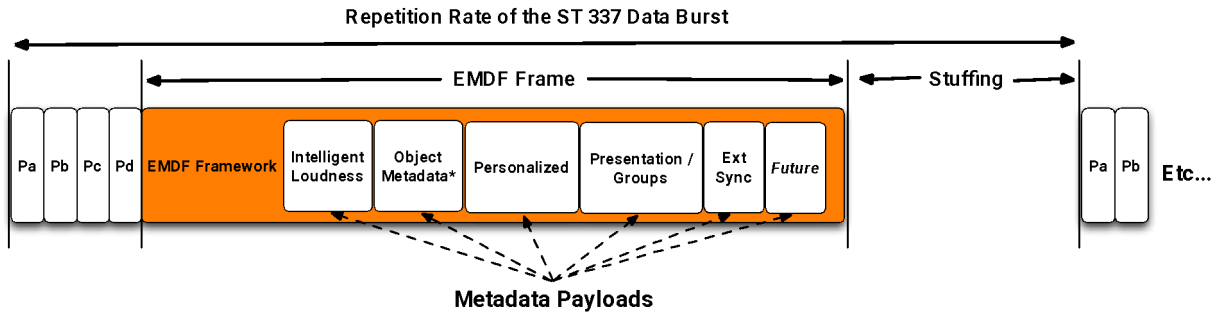

Figure 27. Serialized EMDF Frame formatted as per SMPTE ST 337 [36]

Also underway in SMPTE is the mapping of Serial ADM to the MXF file container. This will simplify file to linear conversion where MXF is commonly used in play-out or recording of live streams to file.

A repetition rate of once a video frame is recommended for most applications to allow for fast acquisition of the audio. When transmitting dynamic positional information for audio objects, higher update rates may be required.

Transmission of Serial ADM over HD-SDI or SMPTE 2110-31 may be subject to switching so it is recommended to align the metadata frame according to the SMPTE RP168 reference point. This avoids metadata loss due to a switch occurring during the metadata frame. The Dolby E timing specification in SMPTE RDD33 [125] can be used for this purpose.

The use of Serial ADM over highly constrained bandwidth channels may result in latencies that are too high for some applications. To reduce latency, ensure that the gzip compression method is used. To reduce latency further, SMPTE 2116 supports the use of multiple AES3 audio channels. This decreases latency at the expense of increased bandwidth.

In a dynamic switching scenario, accurate co-timing of the audio and Serial ADM metadata is critical. For legacy environments where both the audio and the Serial ADM occupy AES frames, maintaining this co-timing is relatively easy. In an SMPTE 2110 environment, where the audio and the metadata occupy different RTP streams, there is a greater risk that the co-timing will fail. To avoid this possibility the latencies of the audio and metadata paths must be carefully managed.

As interest in real-time workflows supporting immersive and personalized audio increases, Serial ADM is well positioned as an emission format agnostic solution for the media production industry. Its close relationship to ITU-R ADM [72], and an increasing portfolio of supporting standards, is expected to result in widespread adoption. In scenarios where switching or dynamic metadata is required careful management of timestamps and latencies is required to ensure accurate co-timing of audio and metadata throughout a facility.

9.2. DTS-UHD

9.2.1. Introduction

The DTS-UHD coding system is the third generation of DTS audio delivery formats. It is designed to both improve efficiency and deliver a richer set of features than the second generation DTS system.

The first two generations of DTS codecs were designed primarily for Channel-based Audio (CBA), whereas DTS-UHD is primarily designed to support audio objects, where a given object can represent a channel-based presentation, an Ambisonic sound field or audio objects used in Object-based Audio (OBA). It can support up to 224 discrete audio Objects for OBA and 32 Object Groups in one stream.

A primary advantage of CBA is a relatively light metadata burden, as a stream is constrained to a very limited number of playback options. OBA however, requires additional metadata to support the audio presentation and control but there are two major advantages to DTS-UHD Object-based Audio:

- Adaptability to the listening environment. Audio programs mixed using OBA do not need to assume a particular listening environment (e.g. speaker layout or dynamic range). This allows the playback system to render the best experience for the listener.
- The ability to adapt to the listener's preference. OBA allows efficient support for features like alternate speech tracks and listener customizations such as changing the speech volume (without affecting anything else).

DTS-UHD has been standardized with the European Telecommunications Standards Institute (ETSI) in TS 103 491 [91], and is included in the DVB Specification TS 101 154 [92] as well as also supported by the Society of Cable Television Engineers in SCTE 242-4 [93] and SCTE 243-4 [94]. DTS-UHD can be encapsulated in a number of transport formats including ISOBMFF, MPEG-2 Transport Stream and CMAF.

One of the challenges of OBA is the additional metadata necessary to support a presentation. DTS-UHD has provisions for reducing the frequency at which metadata is repeated, thus reducing this burden. OTT streaming methods such as DASH and HLS can utilize larger media in blocks of samples that have guaranteed entry points. DTS-UHD permits encoding options to only update metadata when necessary.

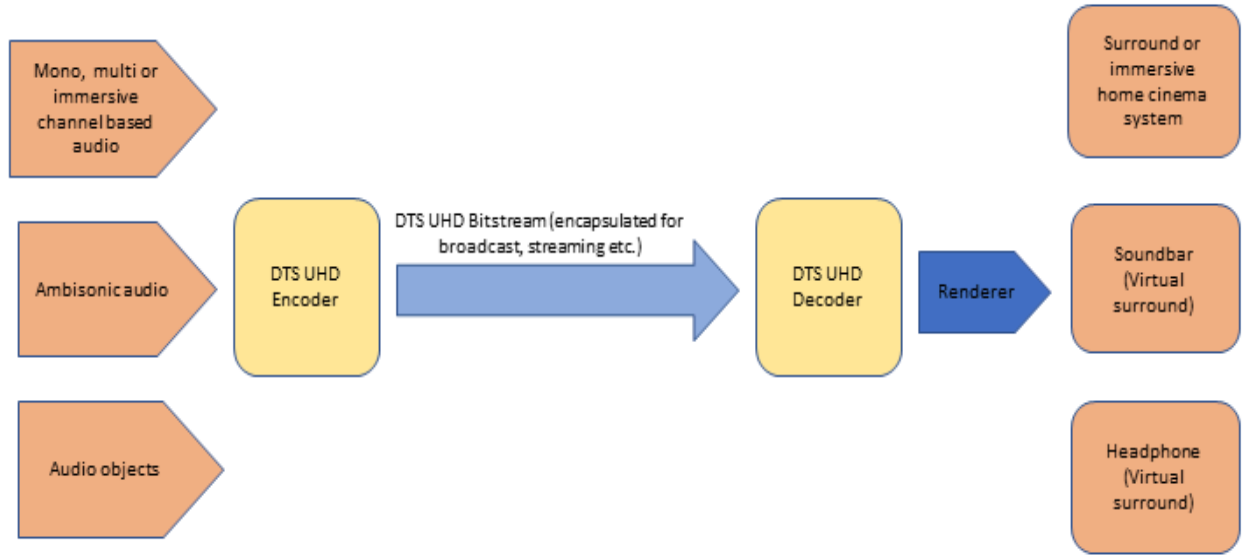

Figure 28. DTS-UHD System Overview

DTS-UHD allows users to interact with the content through controlling objects within the audio in order to personalize the experience. For the general user this would allow control of the relative level of the dialogue track in order to deliver a solution for clear speech. Additionally, it can allow the user to turn on or off additional aspect of the audio, either to deliver additional language tracks or alternative commentary tracks.

DTS-UHD provides additional support for accessibility services with the added interactivity. Two specific use cases are in the support of the visually and hearing impaired. For the hearing impaired the user may be able to interactively control audio tracks or with preset Dialog Enhancement settings. For the Visually impaired an additional ‘audio description’ service can be delivered as a separate object. This would allow not only control of the volume of audio description but could also allow the user to place the audio description in a position within the sound field.

DTS-UHD allows both the user and content author to manage the loudness of content. This ensures the end user receives uniform target loudness regardless of the incoming content loudness while maintaining as much as possible the original dynamic range of the content.

DTS-UHD allows hybrid delivery of different components of the audio content. This would allow the main audio and video service to be delivered via IP Multicast, with additional audio streams,

such as language services or audio description delivered via an additional distribution method such as DASH. A DTS-UHD supporting application can receive, decode, sync and render the content for the consumer.

9.2.3. DTS-UHD Bitstream

The DTS-UHD bitstream is a sequence of DTS-UHD audio frames comprising a Frame Table of Contents (FTOC), audio elements and metadata containing information on positioning as well as loudness.

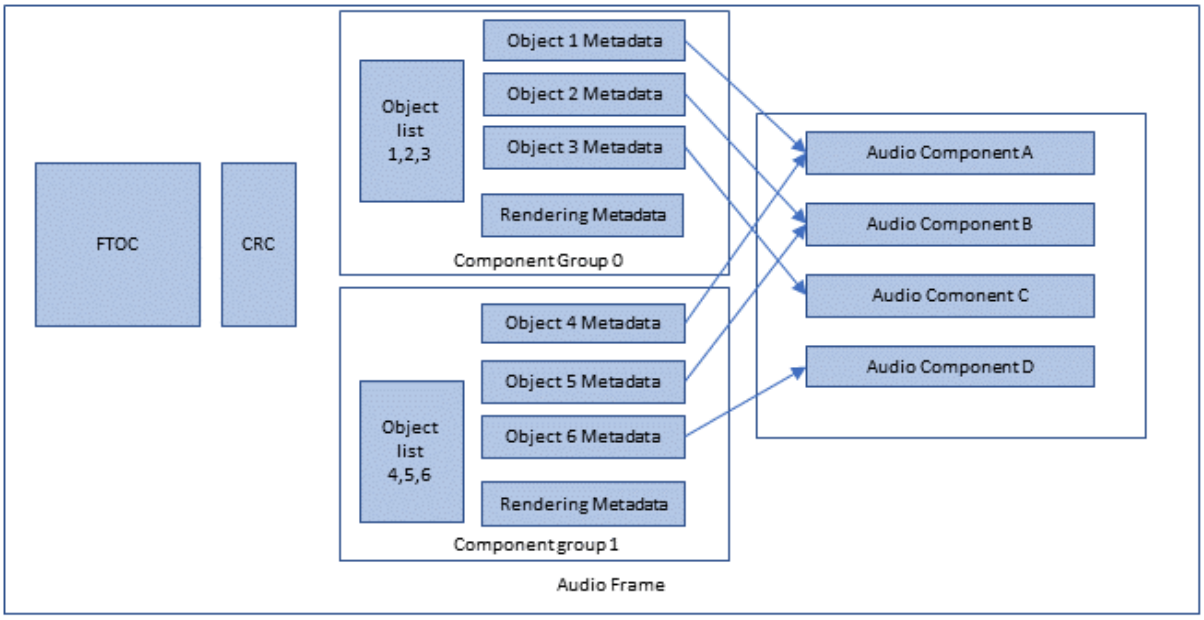

Figure 29. DTS-UHD Audio Frame Structure Example

The FTOC is the only element of an audio frame that is guaranteed to be present. The main components of the FTOC are the Sync Word, which indicates whether the frame is a sync frame or non-sync frame, default presentations and navigation information to the metadata chunks and audio chunks within the audio frame payload. Upon playback a device may invoke the default presentation, an alternate presentation (if present) or a custom playback presentation configuration.

9.2.3.1. Sync and Non-Sync Frames

The decoder does not need any information from previous or future frames to produce a frame of output Linear PCM samples from a sync frame. All parameters necessary to unpack metadata and audio chunks, describe audio chunks, render and process audio samples and generate a frame of Linear PCM samples can be found within the payload of a sync frame. A decoder establishes initial synchronization exclusively with a Sync frame. These frames represent the random access points for random navigation to a particular location in the bitstream.

A non-sync frame permits both metadata chunks and audio chunks to minimize payload size by only sending parameters that have changed in value since the previous frame or sync frame, as stated in the introduction. All parameters that are not re-transmitted are assumed to maintain their previous value. Any value or set of values may be updated in a non-sync frame.

A decoder cannot establish initial synchronization using non-sync frames, nor can these non-sync frames be used as random-access points.

9.2.4. Metadata

Metadata for multiple objects and object groups can be packed together within an associated metadata chunk. Each chunk may be associated with a particular audio presentation index.

Notice that two metadata chunks of the same type may have different audio presentations indexes.

Metadata chunks carry the full description of the audio data chunks and how decoded audio is rendered for a default audio presentation. Metadata may also lock out interactivity or limit the extent to which a user may personalize content. Additional types of metadata that may be useful for categorization of an audio presentation, support of some post-processing functionality, etc., may also be carried within the metadata chunks.

Each DTS-UHD stream decoder instance can be configured from the system layer in three different ways, depending on the type of information that is provided, in order to select desired audio playback presentation. In particular, within DTS-UHD metadata frame:

- Metadata describing different audio presentations may be present.
- The list of audio presentations/audio objects to be decoded within this stream is passed through a decoder instance API by one of three methods listed below. More detail of the object selection is shown in Section 9.2.6.4.

- Play default presentation. In this case, no parameters are presented to the decoder and the audio presentation with the lowest presentation index where bEnblDefaultAuPres is TRUE will be selected, and the default objects within that presentation will be played. This case is indicated by m_ucAudPresInterfaceType = API_PRES_SELECT_DEFAULT_AP.
- Play by presentation index. In this the desired audio presentation is indicated by a single parameter and the default objects within that presentation will be played. This API is aware only of the selectable audio presentations and non-selectable audio presentations are not counted in ucDesiredAuPresIndex. This case is indicated by m_ucAudPresInterfaceType = API_PRES_SELECT_SPECIFIC_AP.
- Play an explicit list of objects. In this case, an ordered list of object IDs are presented to the decoder and only the audio presentations containing objects from this list are unpacked and played. If some of the listed object IDs are that of an object group, then that group's object activity mask is respected and the corresponding referenced objects are played. This case is indicated by m_ucAudPresInterfaceType = API_PRES_SELECT_OBJECT_ID_LIST.

In every sync frame all active metadata is transmitted, and all previous states are reset with the exception of static metadata (pointed to by m_pCMFDStaticMD). When static metadata is distributed over multiple frames, it will be completely refreshed from one sync frame to the next. For example, if the interval between consecutive sync frames is 10 frame periods, then (conceivably) as little as 1/10 of the static metadata could be sent in each frame.

Other elements of metadata required for presentation are described below.

9.2.4.1. Loudness

The DTS-UHD elementary stream is capable of carrying multiple loudness parameter sets, some of which include (nominally) the complete presentation, the speech components only, and composition of all components excluding the speech. Loudness parameters are computed during encode, however the encoder does not modify the audio. Application of loudness parameters is either done in the decoder or as a post process, depending on the design of the system. The decoder can output any reasonable loudness level, e.g. from -31 to -16 LKFS. The system will apply DRC accordingly with reference to the output loudness. A field within the metadata provides an index of the long-term loudness measurement type for the audio, being either ATSC, EBU or ITU.

9.2.4.2. Dynamic Range Control and Personalization

Multiple selectable and custom dynamic range compression curves can be associated with an Audio Program to facilitate adaptation to various listening environments. The presence of a selectable DRC curve is indicated by the bitstream metadata parameter m_bCustomDRCCurveMDPresent as defined in ETSI TS 103 491 [91]. Different curves can be used to accommodate various playback environments. These curves are based on the DRC compression types and parameters based on the general symmetry between the amount of boost against attenuation. Specific slow/fast attack and release times are associated with each profile.

Table 3. Common DRC curves

DRC Curve	Compression type	Boost vs attenuation parameter (See note 1)
Common 1	Low	A
Common 2	Low	B
Common 3	Low	C
Common 4	Medium	A
Common 5	Medium	B
Common 6	Medium	C
Common 7	High	A
Common 8	High	B
Common 9	High	C

Table 3 Notes:

1. For the boost vs attenuation parameter:
    - A has less aggressive attenuation to loud content.
    - B has a less aggressive boost to quiet content.
    - C has equal amount of attenuation and boost

Other legacy DRC curves are also supported within the system for film, music and speech. Additionally, a fully customized curve can be included in the metadata, as described in ETSI TS 103 491.

9.2.4.3. Metadata Chunk CRC Word

To ensure error detection, if the CRC flag (transmitted within FTOC: Metadata and Audio Chunk Navigation Parameters) corresponding to particular metadata chunk is TRUE the metadata chunk CRC (16 bit) word is transmitted in order to allow verification of the metadata chunk data.

This CRC value is calculated over metadata fields, starting from and including the MD Chunk ID and up to and including the byte alignment field prior to the CRC word.

The decoder will:

- Calculate the CRC(16) value over the metadata chunk data fields.
- Extract the 16-bit MD chunk CRC field and compare it against the calculated CRC(16) value.
- If the two values match, reverse back to the beginning of the metadata chunk (return TRUE); otherwise, pronounce data corruption (return FALSE).

9.2.5. Audio Chunks

Audio chunks carry the compressed audio samples. Audio samples may represent speaker feeds, waveforms associated with a 3D audio object, waveforms associated with a sound field audio representation, or some other valid audio representation. The associated metadata chunk fully describes the way a particular audio chunk is presented and the type of audio carried within each audio chunk.

An audio chunk points to a minimum collection of compressed waveforms that can be decoded without dependency on any other audio chunks. All compressed waveforms within an audio chunk that has been selected for decoding shall be decoded and played together. In some cases an elementary stream may already have its own sub-division into individually decodable parts in which case all encoded objects within one DTS-UHD stream can be packed into a single audio chunk. In some cases an audio chunk does not point to any compressed waveforms but rather it points to header / metadata information within a compressed audio elementary stream.

For each audio chunk:

- The chunk ID, the payload size in bytes and the audio chunk index are all transmitted within the FTOC payload.
- There are no header parameters in addition to the chunk payload.
- An audio chunk type, as pointed by the chunk ID, identifies the type of data stored in a corresponding audio chunk.

9.2.6. Organization of Streams

9.2.6.1. Objects, Object Groups, Presentations

The fundamental unit of a DTS-UHD stream is the object. The simplest example of a DTS-UHD stream would be a stream containing one object. For example, one stereo audio presentation, or even a single 5.1 or 7.1 channel presentation, could be handled in such a manner.

Object Groups provide a mechanism to associate objects that should always be used together with a single identifier.

Presentations are composed of a selection of objects and / or object groups. Membership of an object or object group in a presentation is non-exclusive.

9.2.6.2. Properties of Objects

The **object** metadata carries parameters needed to:

- Uniquely identify an object within a DTS-UHD stream.
- Point to associated audio waveforms.
- Describe the audio object properties necessary to render associated audio waveforms.
- Assign whether the Default Playback status of the object is Active or Silent.
- Describe the type of audio content the object is associated with.
- Describe the object's loudness and dynamics properties.

9.2.6.3. Properties of Object Groups

The **object** group metadata carries parameters needed to:

- Uniquely identify an object group within one DTS-UHD stream by means of a unique object ID.
- Indicate which objects belong to the group by means of a list of object IDs.
- Assign whether the default status is to be played or to be silent.
- Indicate which objects within the group by default shall be rendered and which objects shall be silent; note that the object group setting can overwrite the individual object default activity flags.

Note that object groups do not directly point to any audio waveforms but only point to the specific object IDs. The definition of object groups is fairly generic and hence can be used for almost arbitrary object grouping.

9.2.6.4. Audio Presentations and Rendering

Multiple audio presentations may be defined within a single DTS-UHD bitstream. Each audio presentation has a unique audio presentation index within a stream.

Each DTS-UHD stream requires a dedicated DTS-UHD stream decoder instance. Each DTS-UHD stream decoder instance is configured with one of the three types of audio presentation selection APIs:

- Play the default presentation only
- Play a selected presentation
- Play a list of selected elements (object)

Figure 30 and Figure 31 show two playback examples; the first one using Default Playback and the second one using specific object and object group selection. In both examples the darker blocks indicate the active elements.

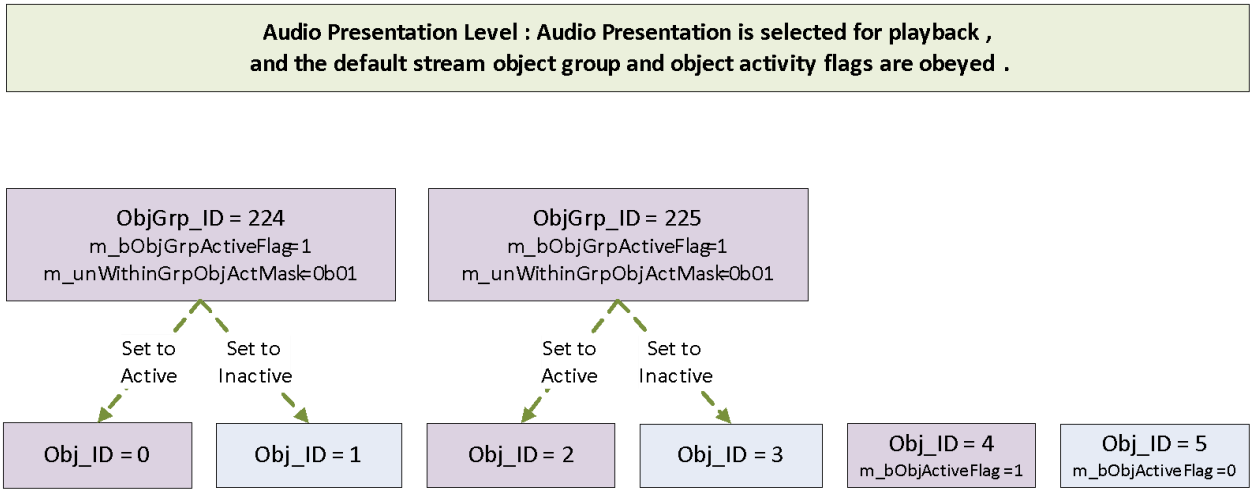

Figure 30. Default Playback

Once configured, the particular instantiation of a stream decoder cannot change the type of presentation selection API.

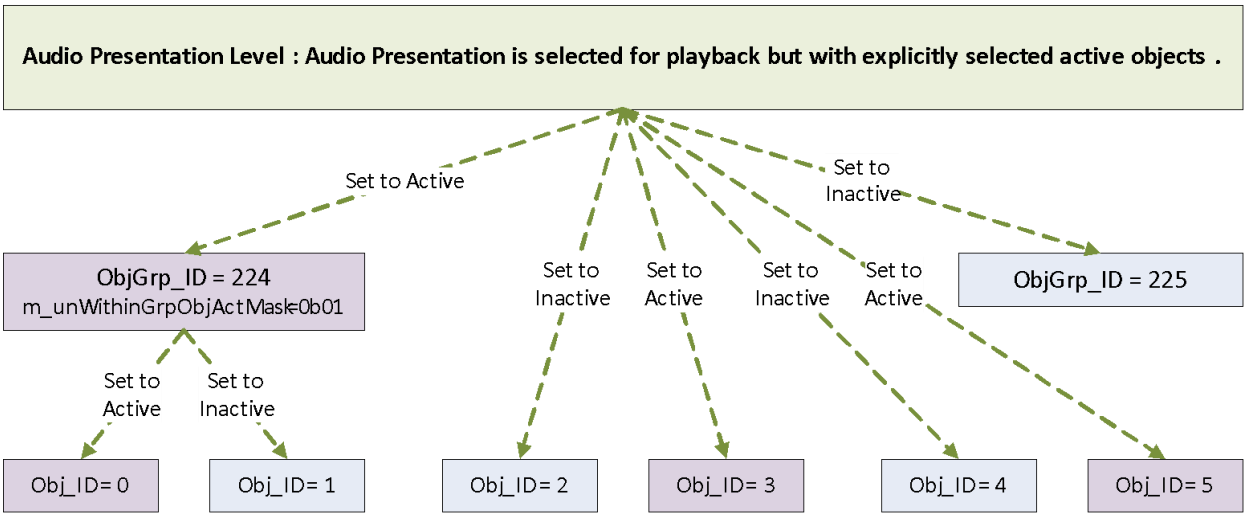

Figure 31. Specific Object and Group Selection

The following three diagrams illustrate examples of selecting desired objects to play from multiple presentations within a single stream. Purple blocks indicate active audio presentations and corresponding object groups and objects.

Figure 32 is an example of playback using the default audio presentation (*m_bEnblDefaultAuPres*=1), i.e. the lowest indexed selectable audio presentation (AP1). The default object group and object activity flags within AP1 are being obeyed. In addition external object groups are activated.

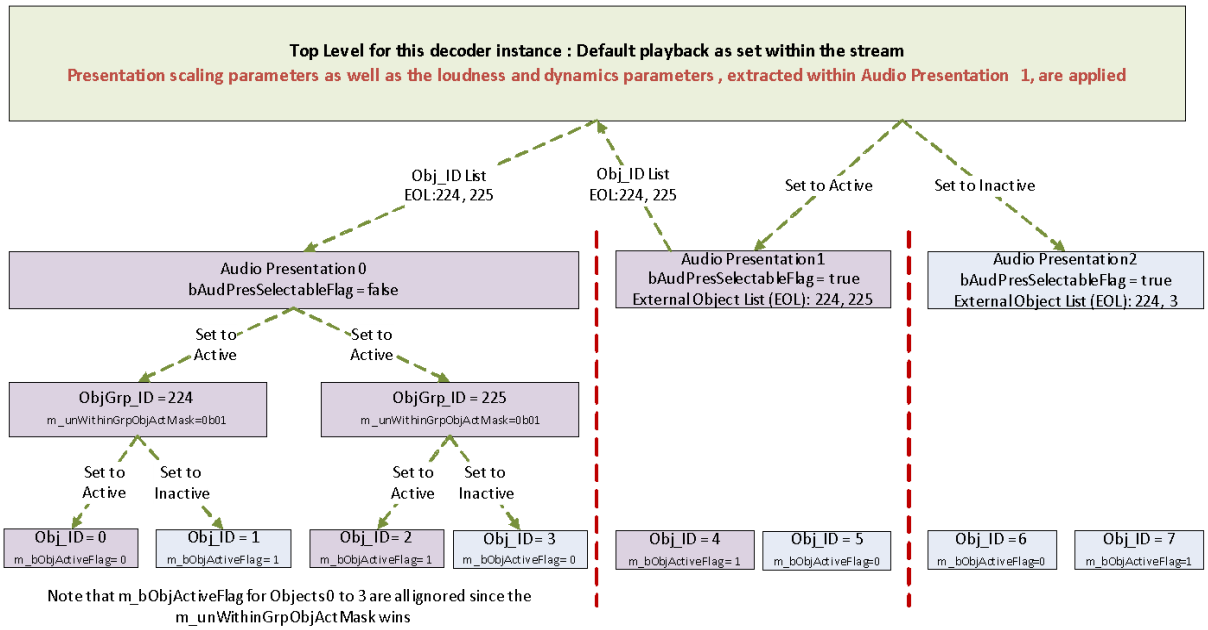

Figure 32. Playback using Default settings

The diagram, Figure 33, shows default object group and object activity flags within AP2 are being obeyed. In addition, external object groups are activated.

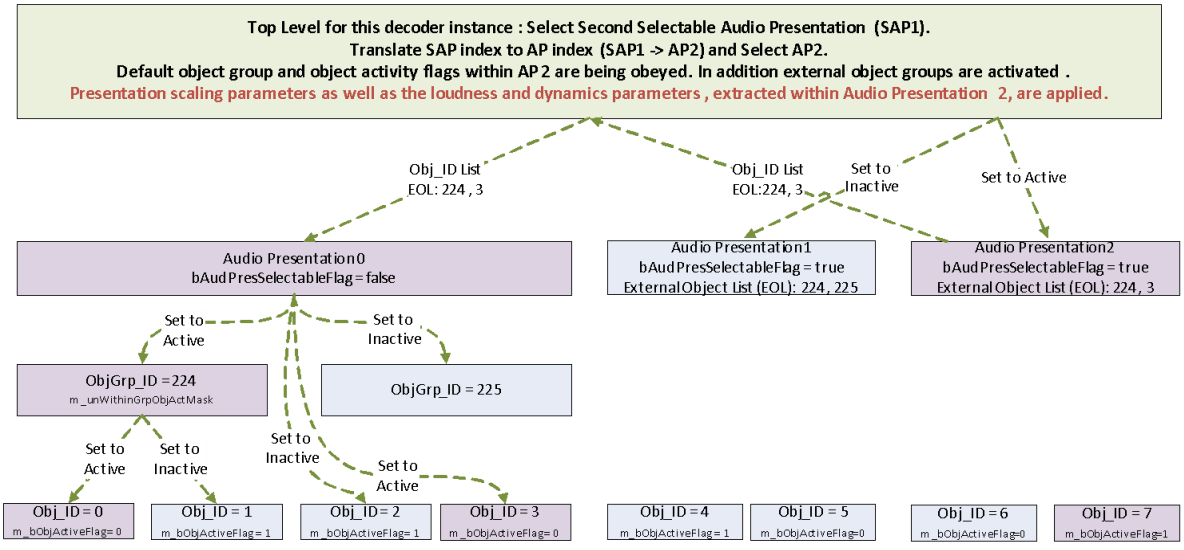

Figure 33. Example of Selecting Playback of Audio Presentation 2

The diagram in Figure 34 shows no default presentations being selected; rather an explicit playlist is selected which can override all defaults.

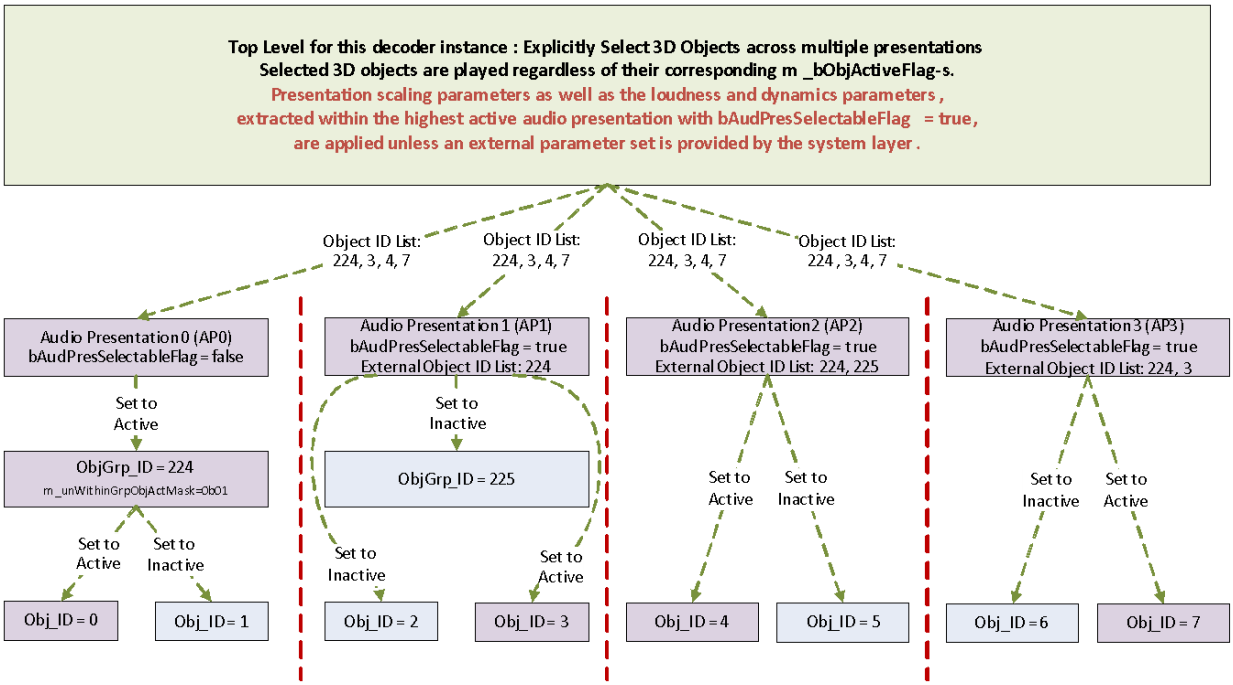

Figure 34. Example of Selecting Desired Objects to Play Within a Single Stream

The decoder processes the selected presentation from the DTS-UHD audio stream into a set of linear PCM waveforms, which are delivered to the rendered with the associated metadata for positioning, loudness etc, along with any additional information created through personalization. The renderer will then process the metadata to produce a mix of all of the objects and deliver these to the relevant output speakers.

9.2.7. Multi-Stream Playback

All of the above examples of playback have used a single DTS-UHD audio stream and as has been stated each audio stream requires a DTS-UHD decoding instance. However, this does not preclude the use of additional streams, as in the hybrid example in the previous section. The presentation may be made up of multiple streams, with a main stream and additional auxiliary streams. There are two options available in this instance to decode the streams in this case:

- A single decoder may process the audio frames from the various streams as required sequentially, then render all the waveforms from the given time interval together to generate the final output.
- Separate decoder instances can be used to decode each stream, with each stream passing the associated metadata to the renderer. In this case the final rendering metadata for scaling the output shall always be provided by the highest ordered elementary stream in the sequence that contains such metadata.

In the example of Figure 35, three elementary streams contribute to a particular preselection. Component #2 is from the highest ordered stream in a multi-stream preselection. The renderer will first look for metadata from Component #2 to perform the final scaling of the mix. If some metadata is missing, then the renderer looks at the metadata delivered with Component #1, and finally Component #0, in order, to fill in the missing metadata.

To illustrate this example, consider that the component from elementary stream #0 carries music and effects, the component from elementary stream #1 carries dialogue, and the component from elementary stream #2 adds spoken subtitles. Multiple dialogue objects might be able to use the same music and effects, so the mixing metadata with the dialogue will be preferred when only these two components are selected. Since the spoken subtitle is stored in stream #2, and was mastered with the M&E plus dialog, it was the only one mastered with the awareness of the other components. Therefore, the metadata in Component #2 can provide the best experience. In some scenarios, new mixing metadata may not be generated with the spoken subtitle, i.e. it was mastered in consideration of the stream #1 metadata. In this case, stream #1 metadata will be used for the final rendering.

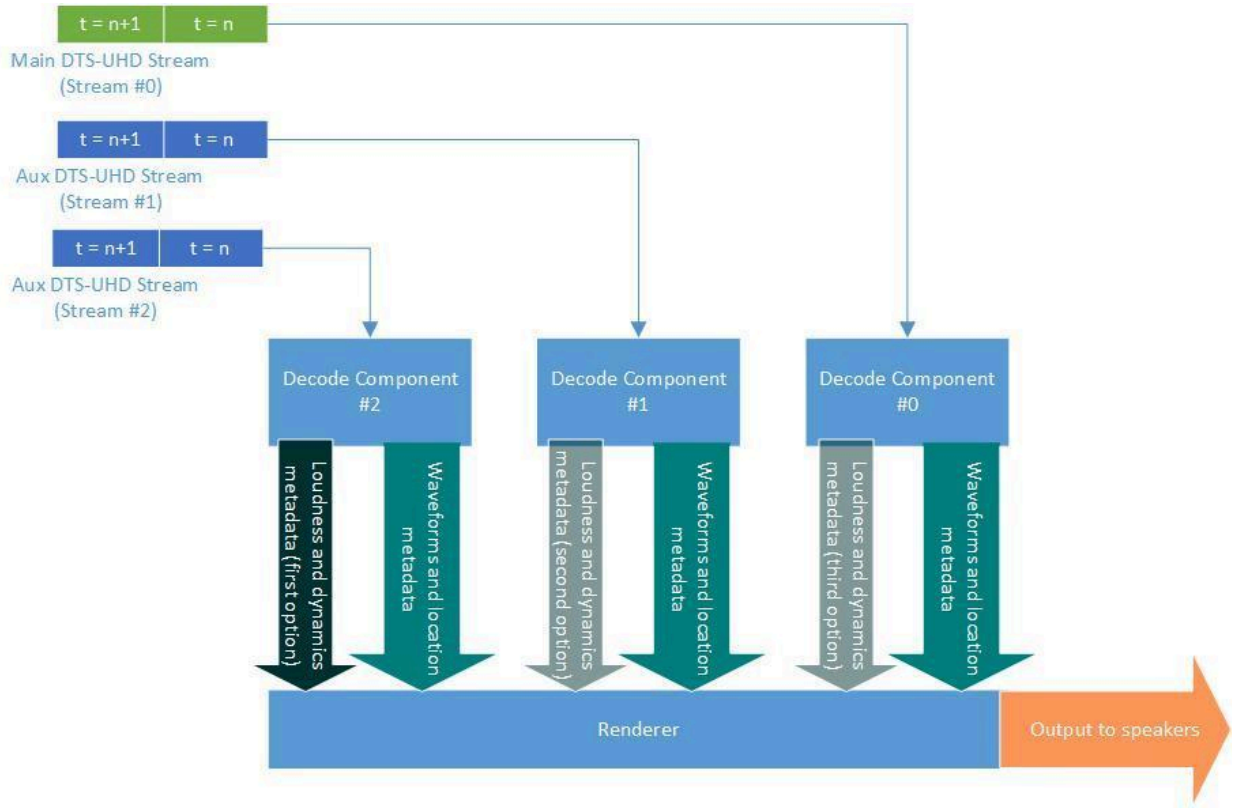

Figure 35. Example of multi-stream decoding

9.2.8. Rendering

For rendering DTS uses a point source renderer based on the ego-centric model. In Section 9.1.5 the differences between allocentric and ego-centric rendering has been explained, however in practice the production sound mixer will place objects within a soundfield or space, and the renderer itself uses either the allocentric or ego-centric method for audio object placement at reproduction. The renderer uses metadata as previously described in the DTS-UHD bitstream in order to manage the placement and gain of the objects.

For the DTS-UHD point source renderer all objects are placed on a point on the surface of a sphere, with each speaker within a system regarded as a point source.

The location of a point is specified by polar coordinates - azimuth (θ) elevation (φ) and radius (r). Only the points on the unit sphere are needed, which means the radius shall be equal to 1.

The listener is at the origin (θ = 0, φ = 0, r = 0), facing the location (θ = 0, φ = 0, r = 1) as shown below.

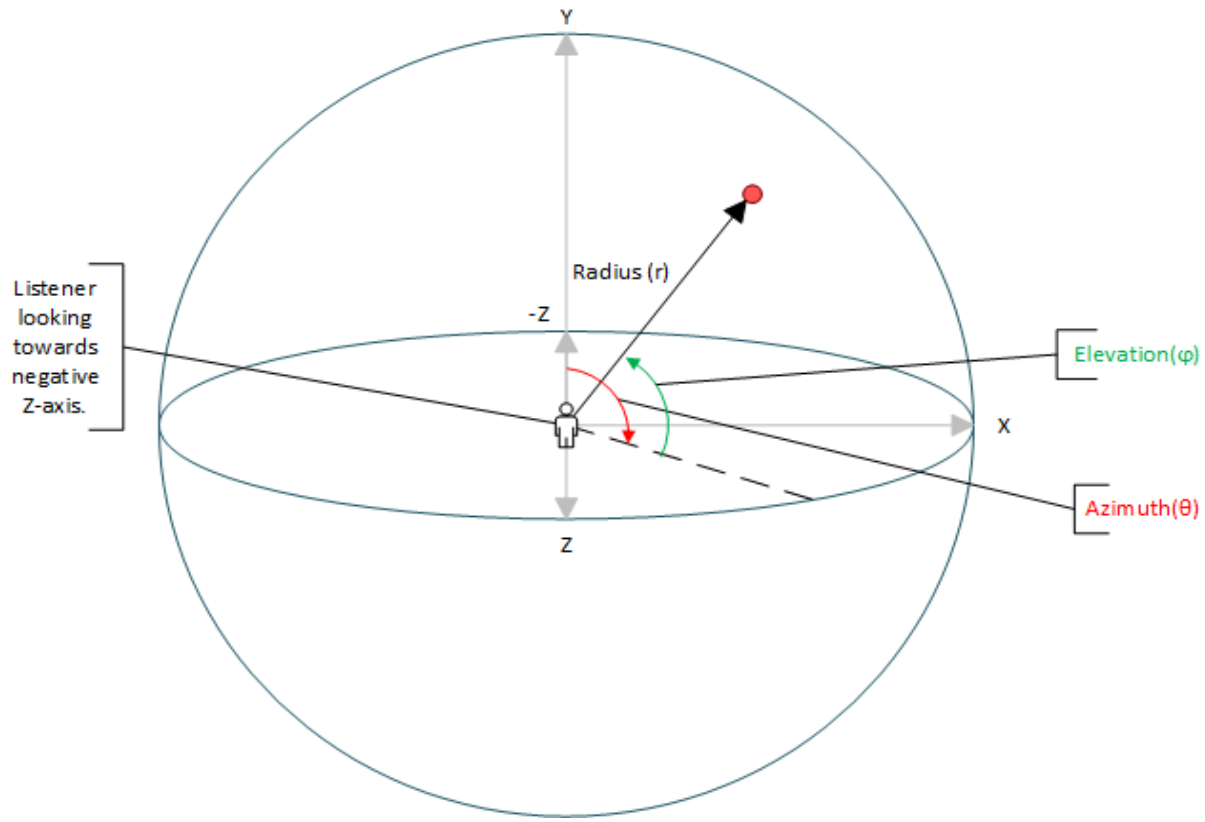

Figure 36. Point Source Object Renderer Coordinate System

As can be seen, this model places all objects on the surface of a sphere with Radius (r) =1. However, in order to place objects within the 3D sphere, the content creator can create a number of point sources associated with a single waveform, placed at different points on the edge of the sphere, and using vector based panning to calculate the gain contribution of each point source, the correct effect is produced. The DTS renderer has the ability to work within either a preset or arbitrary speaker layout. With a preset speaker layout using the previously noted API, the renderer is able to specifically target speaker point sources, however the system is equally able to reproduce 3D sound without an arbitrary layout.

In order to reproduce audio within the above sphere the speaker layout is used to create a mesh with a convex shape. As many speaker layouts have speakers at large angular spacing, this

prevents a convex mesh being created. The DTS renderer uses virtual speakers to create a complete 3-dimensional convex array of speakers within a given sound space. In this case, vector based panning rendering is done over the full set of both physical and virtual speakers, with the fold-down of the virtual speakers to the physical speakers then carried out as a post vector based panning process.

Figure 37. shows the virtual speaker setup for a standard 7.1 configuration. The figure displays virtual speakers in both the upper and lower hemispheres with yellow vertices.

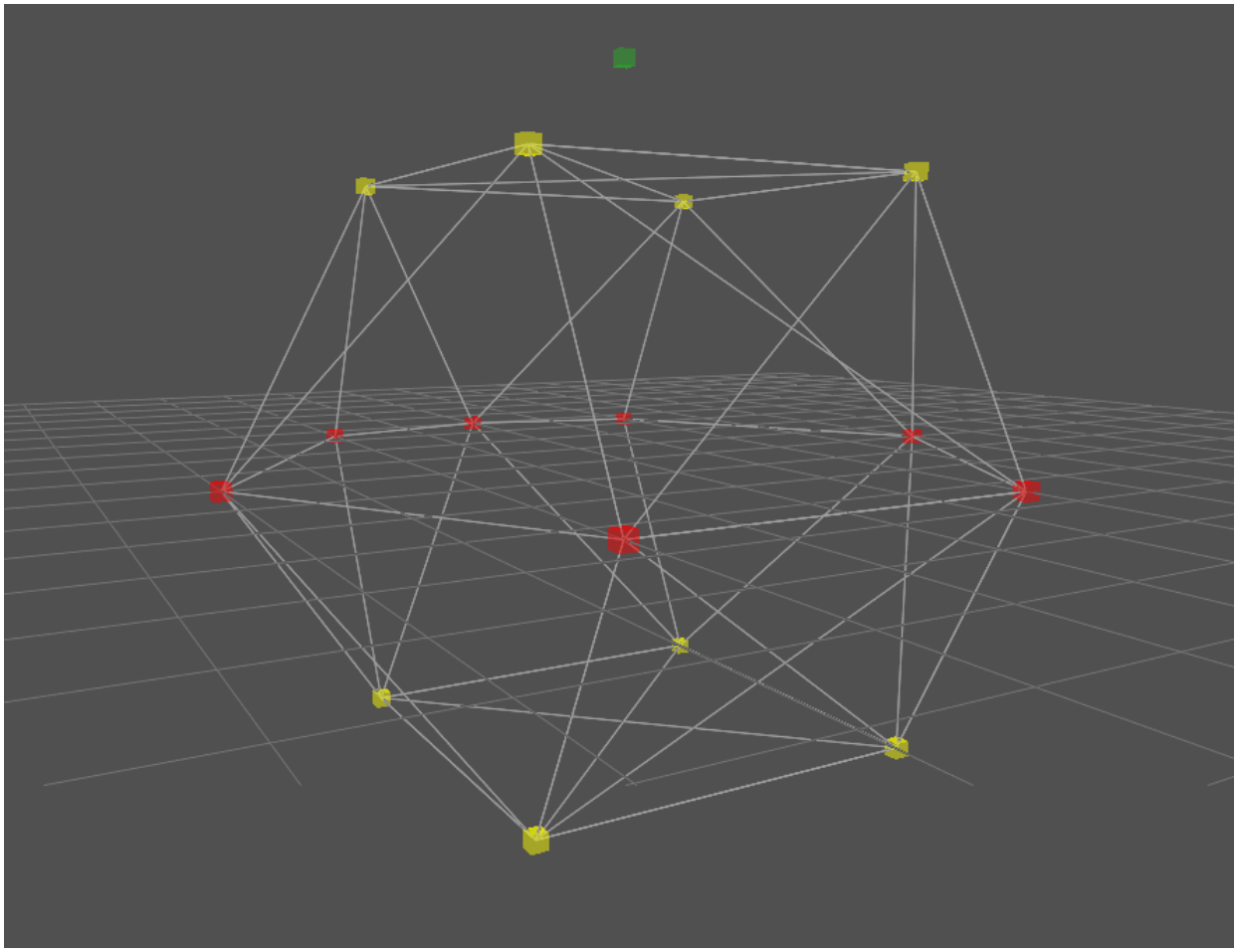

Figure 37. 7.x Output Configuration with Predefined Virtual Speakers

9.2.9. Personalization

As has been previously stated, object metadata from the DTS-UHD bitstream may be overridden by the user during playback for the purposes of dialogue enhancement for example. Metadata in the bitstream may also be set to limit or disable a user interaction. The object interactivity manager enforces these rules and applies any user changes to the metadata before calling the renderer. Figure 38 shows that the object interactivity manager sits just before the renderer, where it handles the user input and the limit rules specified by the bitstream creator.

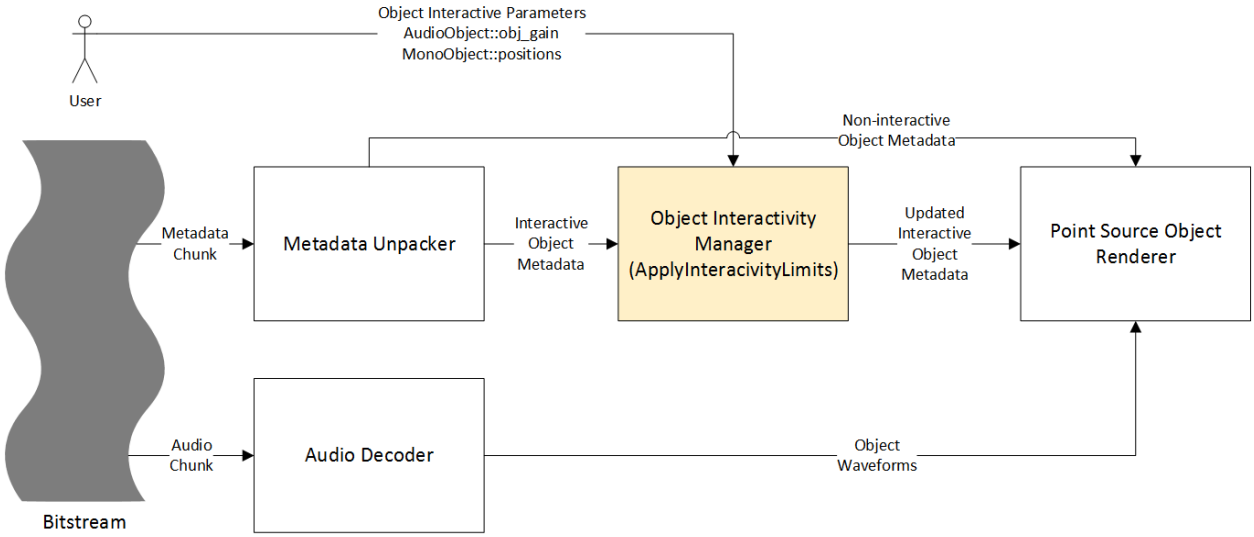

Figure 38. Object Interactivity Manager

9.3. MPEG-H Audio

9.3.1. 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 [70]. The MPEG-H 3D Audio Low Complexity Profile Level 3 is adopted by DVB in ETSI TS 101 154 v.2.3.1 [63] and is one of the audio systems standardized for use in ATSC 3.0 Systems as defined in A/342 Part 3 [57]. SCTE has included the MPEG-H Audio System into the suite of NGA standards for cable applications as specified in SCTE 242-3 [78].

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 [87] 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 39 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.

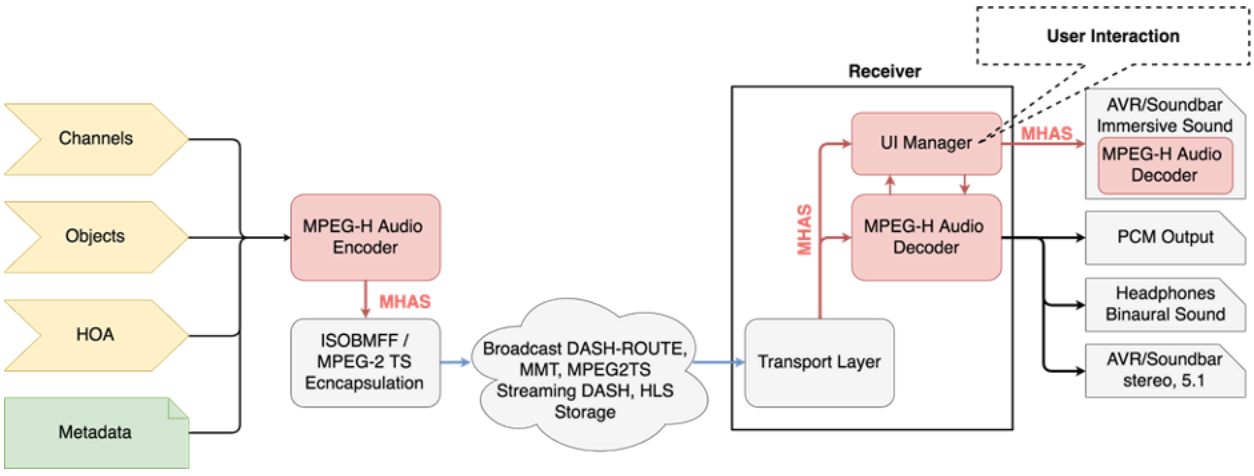

Figure 39. MPEG-H Audio system overview

The MPEG-H Audio Stream (MHAS), described in Section 9.3.3., contains the audio bitstream and various types of metadata packets and represents a 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 Section 9.3.1.4.

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

9.3.1.1. 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 40. (See also Section 9.3.2 MPEG-H Audio Metadata.)

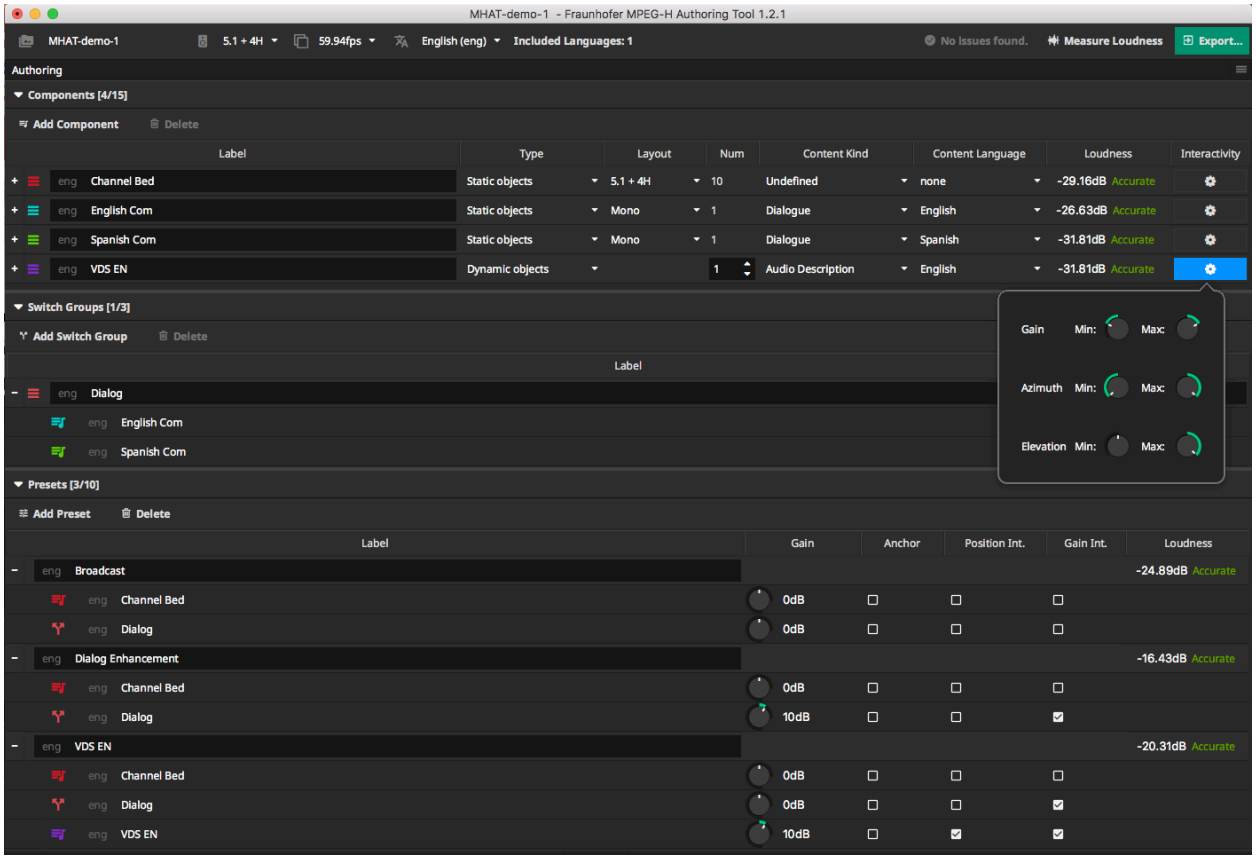

Figure 40. MPEG-H Authoring Tool example session

9.3.1.2. 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.

9.3.1.3. Immersive 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).

9.3.1.4. 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 41 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.

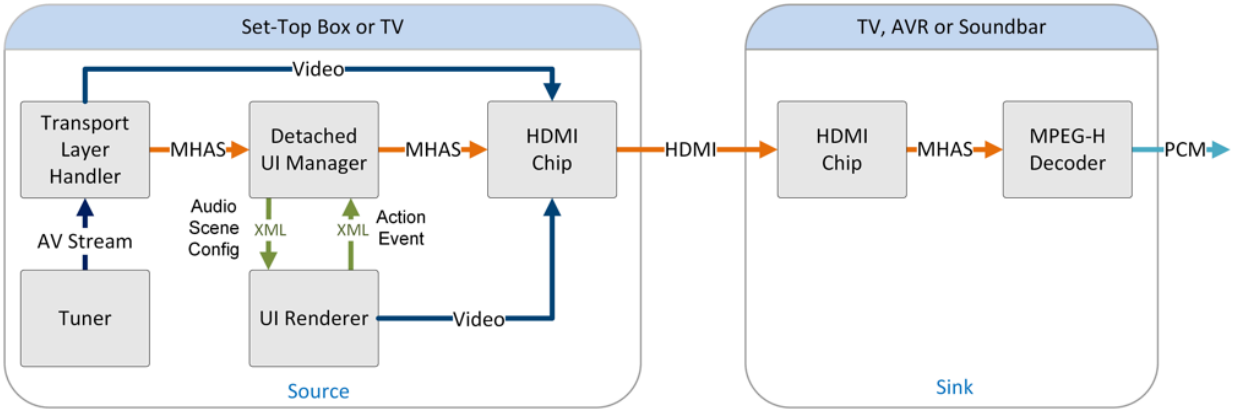

Figure 41. Distributed UI processing with transmission of user commands over HDMI

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.

9.3.2. 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.

9.3.2.1. 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 42, 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.

9.3.2.2. Metadata Example

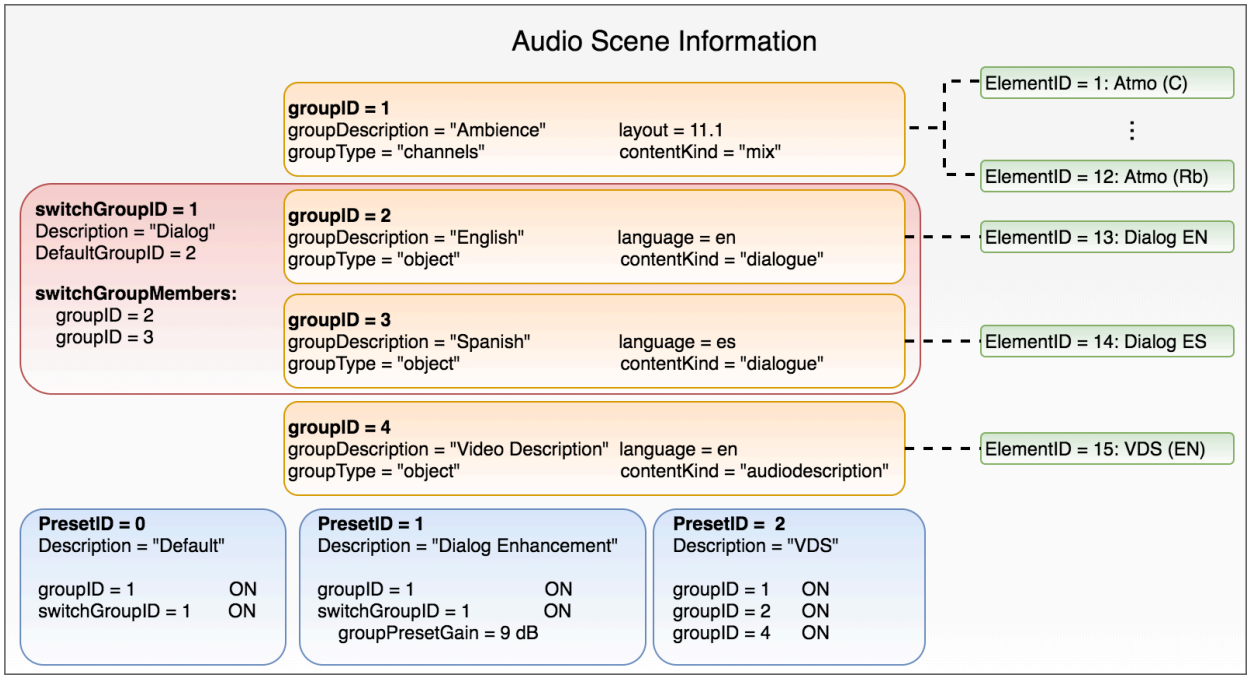

Figure 42. Example of an MPEG-H Audio Scene information

Figure 42 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 subsections.

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).

9.3.2.3. 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.

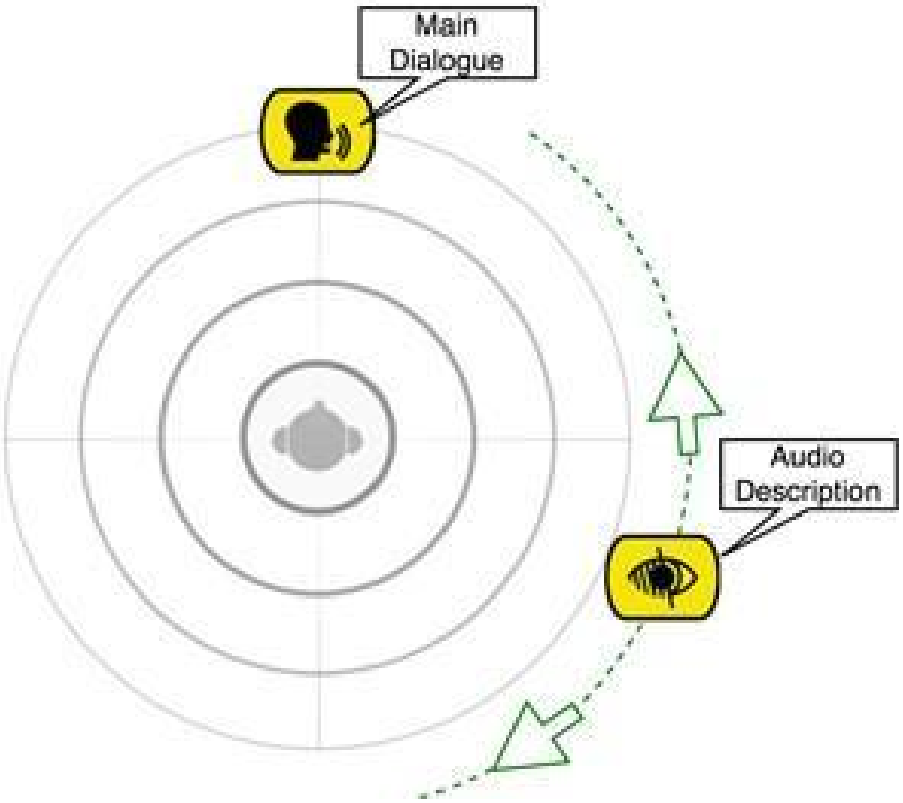

Figure 43. Audio description re-positioning example

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 43. 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).

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 42 using a broadcaster defined enhancement level for the dialog element (e.g., 10dB as shown in Figure 40).

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 40 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.

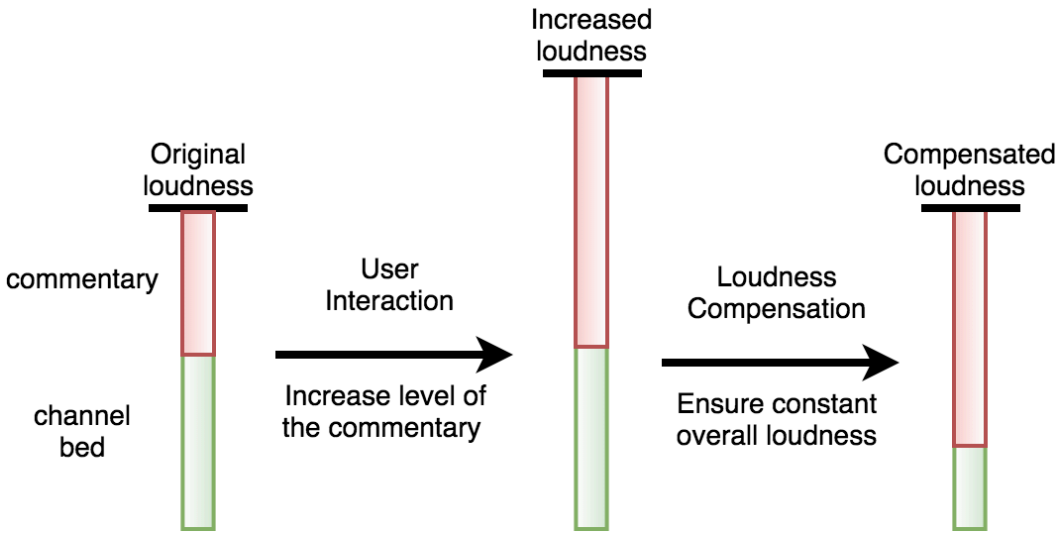

Figure 44. Loudness compensation after user interaction

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 44. This ensures constant loudness level not only across programs but also after user interactions.

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 signaling 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.

9.3.3. 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.

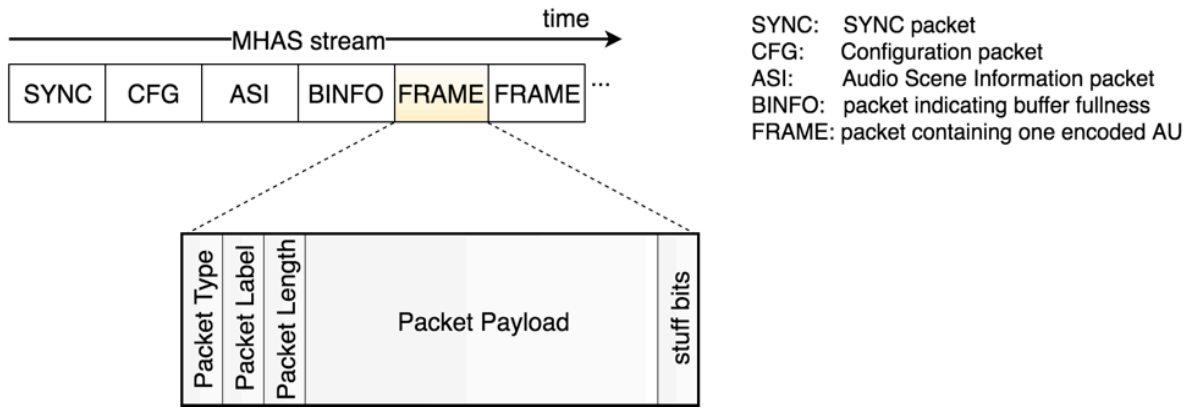

Figure 45. MHAS packet structure

Figure 45 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.

9.3.3.1. 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.

9.3.3.2. 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 signaling 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.

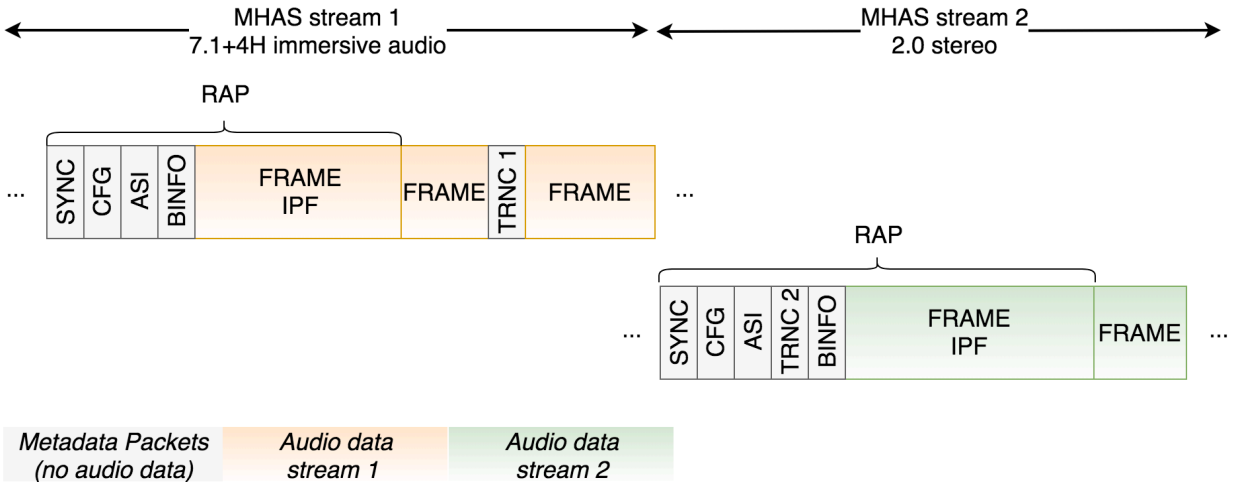

Figure 46. Example of a configuration change from 7.1+4H to 2.0 in the MHAS stream

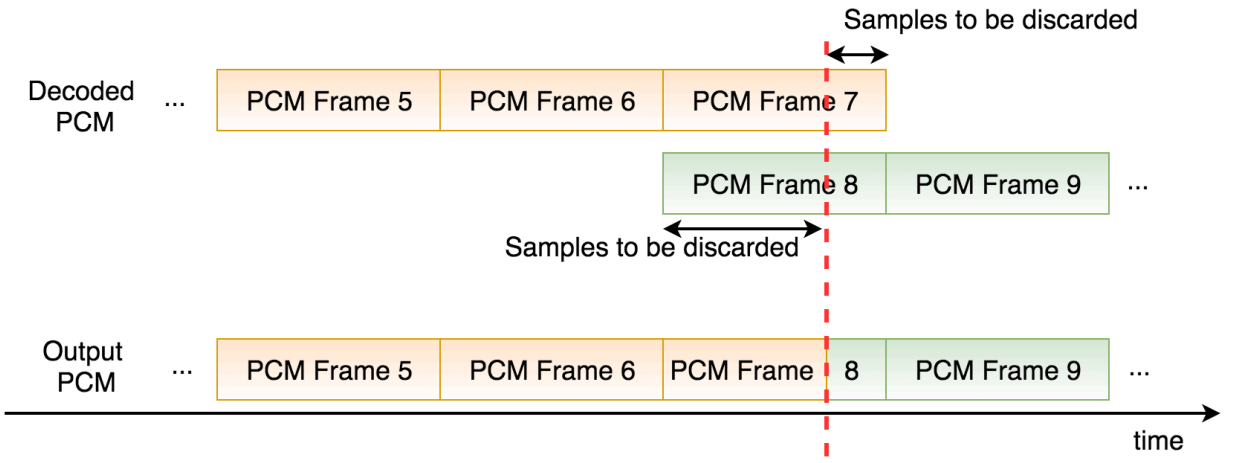

Figure 47. Example of a configuration change from 7.1+4H to 2.0 at the system output

Figure 46 and Figure 47 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.

10. Monographs on Workflow

Ultra HD workflows are discussed in many places throughout these Guidelines. Some workflows are standardized (e.g., see ITU-R BT.2408 [8]). Elsewhere, notable workflow “firsts” are often achieved and reported in conjunction with major spectacle events (see Section 13 [V02] in the Violet Book). As such cutting-edge work matures, still maturing techniques are more thoroughly documented here. Note that these techniques are not necessarily isolated from generic workflows, nor each other - hybridization and/or migration are possible. The monographs on workflow are:

- ACES Workflow for Color and Dynamic Range (Sec 10.1) models the processing from source to display in stages bounded by significant, purposeful transformations. Well-known for Ultra HD Blu-Ray authoring.
- IP-based Workflow - (Sec 10.2) supports modern media services with many alternate tracks of audio, alternative video views, etc. to be combined and distributed as different finished compositions.
- NBCU Single-Master HDR-SDR Workflow Recommendations (Sec 10.3) replace impractical dual-master live workflows to deliver to both emerging UHD/HDR and legacy HD/SDR platforms simultaneously.

10.1. ACES Workflow for Color and Dynamic Range

The ACES [50] project provides a thorough workflow that can be used to model the processing of HDR/WCG video signals from source to display in a series of stages that mark the boundaries of significant transformations, each with a specific purpose. Most, if not all, Ultra HD Blu-Ray (BD-ROM 3.1) content were authored in ACES workspace.

Source code and documentation for ACES is available at: https://github.com/ampas/aces-dev/

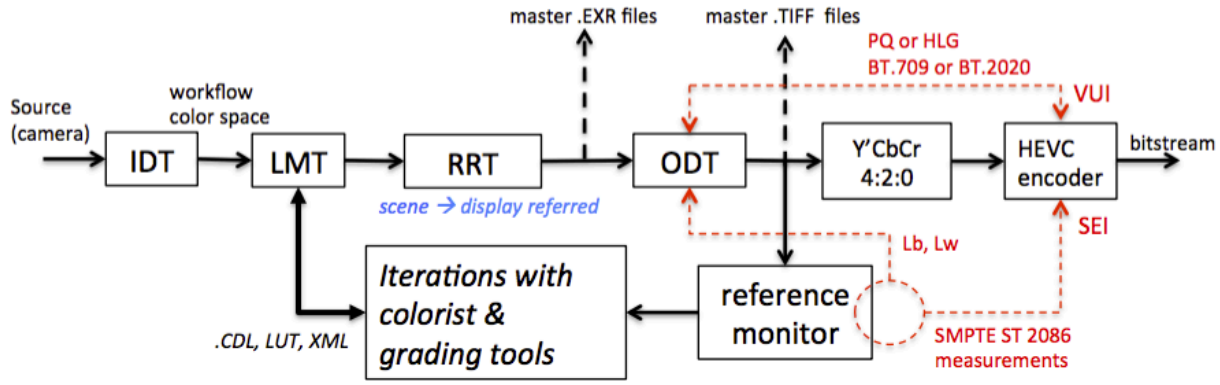

Figure 48. ACES Workflow Model

The basic ACES workflow model stages are described in the following Table 5.

Table 5. ACES Workflow Model

ACES Stage	Purpose
IDT	Input Device Transform: camera format (Bayer RAW, Slog3, etc.) to the ACES working color space
LMT	Look Modification Transform provides an appearance such as “dark night”, “indoor lighting”, etc. established by cinematographer.
RRT	Reference Rendering Transform: converts scene referred signal to display referred signal, with knowledge of reference viewing environments and limited display ranges.
ODT	Output Device Transform. Maps display referred to a specific display range (black and peak white levels), container color primaries (BT.709, BT.2020), and transfer function (gamma, PQ, HLG).

In Figure 48, the blue annotates the Reference Rendering Transform (RRT, e.g., “scene linear -> display referred”). The red detail indicates metadata input to the encoder for the system

colorimetry and transfer function (VUI = video usability information). This could include the Master Display Color Volume (MCDV) metadata (e.g., for HDR10). VUI can optionally populate an Alternative transfer characteristics SEI message to support backwards compatibility. (See also Blue Book, Section 7.3 [B01].)

The Ultra HD TV color grading process will start with ingesting digital camera rushes or scanned film at whatever the usable resolution, gamut and dynamic range is available from the source material. If the source content is in a format specific to the capture device, the source signal will undergo transformation to a more universal processing space in a stage such as the IDT depicted above.

Colorists working on Ultra HD TV projects are likely to continue the practice of first setting the overall mood of the film or program, then deciding how particular scenes fit that mood and finally how the viewer's interest is directed to characters, objects etc. on specific shots. A set of look-modification transforms, conceptualized in the LMT stage depicted above, reshape content according to the intent of directors, cinematographers, and colorists.

Source material will not necessarily be Rec. 2020 or Rec. 709, (unless it is a re - mastering or restoration project), because there are widely different capabilities in cameras, film stocks etc. Therefore, the common starting point for colorists or compositors, will be to bring in all what's available, as this allows more scope in post production.

The colorist or compositor will then make decisions about what to select from that available resolution, range and gamut and how to present it as "legal" PQ10 or HLG10 based content to Ultra HD TV consumer screens, which support Rec. 2100 containers. The display rendering stages (RRT and ODT) shape content to fit within the capabilities of a range of target displays, modeled by the mastering display monitor. The colorist will re-grade content and adjust the look based on the appearance of the rendered content on the mastering display reference monitor used to preview the final appearance. A more advanced workflow configuration could account for additional distortions added by reduced integer precision, $Y'C_BC_R$ signal format conversion, 4:2:0 chroma resampling, and video codec quantization by feeding the output of the decoder to the reference monitor.

Additional considerations would be needed if format interoperability (back compatibility) were being attempted, for example to 4K SDR / Rec. 709 or HD SDR / Rec. 709. Producing deliverables in a Rec. 709 / HDR rendered format is not recommended and it is not clear what the long-term market use case for this would be. HDR and WCG are intrinsically linked in the HSL (hue, saturation, and lightness) or RGB color representations and most importantly also in the way humans 'see'. They are different dimensions of the unified perceptual experience.

10.2. IP-based Workflow[1] – SMPTE ST 2110

When television networks and stations went digital several decades ago, the Serial Digital Interface (SDI) was the “digital analog” of the point-to-point analog cables in use up to that time. Support for higher bandwidth signals has been a progression of defining interfaces having additional cables, or higher bandwidth per cable, or both. New applications are frequently held, waiting for a standard to be developed, to define the needed modifications or augmentations to the related protocols, as critical information needed to process a signal must be encoded as a DID, SDID, embedded in certain lines, etc., and a standard is needed to tell you where.

10.2.1. Why IP?

In recent years, there has been a move to replace purpose-built media equipment with commodity IT equipment. The latter is typically manufactured in larger quantities with a commensurate cost savings and more rapid generational increases in available performance. IT equipment, in general, relies on the switched-packed Internet Protocol (IP) such that automatic routing algorithms determine how a signal gets from point A to point B without requiring, though not precluding, a direct point-to-point connection.

Initially, IP-based media transports were significant replacements for SDI cabling. Media originated at one point and was carried together to a destination. A number of IP-based media transports have become popular and have seen some professional use:

- Circa 2007, SMPTE presented the first standards of the ST 2022 suite, for transport of MPEG-2 transport streams over IP networks, though it wasn’t until ST 2022-6 [82] in 2012 that support for transport over high bit rate networks made the technology suitable for professional-grade media.
- Zixi is a proprietary transport introduced around 2008 by the eponymous Massachusetts -based company and has since developed is a codec agnostic platform able to accept most other streaming formats as input.
- Secure Reliable Transport (SRT), an open source project initiated circa 2012 by the Chicago-based Canadian company Haivision that carries live video in

[1] Since the topic of ST 2110 was first introduced as an Annex to these Guidelines in the April, 2020 edition, interest has expanded. The Ultra HD Interop Work Group has since founded the IP-base Production Users Group, to demonstrate and document IP-base workflows for the creation of content featuring UHD, HDR, WCG, and HFR media by promoting high-level test vectors that are more feature-complete and responsive to the industry’s needs. The protocols listed here are among those of particular interest to Ultra HD Forum members, an investigation is proceeding with some priority, again based on member interest, being given to ST 2110.

real-time without relying on TCP for its robustness. That project is presently maintained by the SRT Alliance. SRT is codec agnostic and offers low latency.

- Network Device Interface (NDI) is a proprietary standard introduced by Texas-based NewTek in 2015 that originally relied on TCP but has since offered options for UDP. NDI features a proprietary codec, said to be lossless, which carries HD video at about 100 Mbit/s.

These technologies, along with others not listed here, remain popular to date, particularly for remote news gathering, cloud-based workflows, including real-time cloud ingress/egress over unmanaged networks.

10.2.2. Why SMPTE ST 2110?

Beyond mere signal transport for pre-synchronized audio, video, and metadata, modern media services contemplate many alternate tracks of audio, alternative video views, etc. To support such separate media essence streams, making them available to be combined and distributed as multiple, different finished compositions downstream, other services are necessary: Timing and synchronization, discovery and connection, configuration and monitoring, and security.

In 2013, the Video Service Forum (VSF), the Society of Motion Picture and Television Engineers (SMPTE), the European Broadcasting Union (EBU), and the Advanced Media Work Flow Association (AMWA) created the Joint Task Force on Networked Media (JT-NM), whose vision was to enable new business opportunities through the exchange of professional media across networks. The JT-NM collected business-driven use cases & requirements, and defined a reference architecture[2] for implementing media networking an IP-based media facility.

Based on the JT-NM requirements & architecture, in 2015, the VSF published a recommendation for carrying elementary streams over IP[3], with the expectation that flexibility and agility of a production facility would be enhanced, in part because IP networks are agnostic to resolution, bit depth, and frame rate, being limited instead only by bandwidth which, if oversized, can often be shared among different connections. Additionally, unlike the earlier SMPTE ST 2022-6 standard for carrying professional media over IP, the VSF recommendation did not bind the video, audio and ancillary data into an SDI multiplex, but instead allowed the elementary media streams to flow independently, though linked synchronously via timestamps,

[2] EBU TECH 3371, “The Technology Pyramid for Media Nodes”

[3] Technical Recommendation TR-03: Transport of Uncompressed Elementary Stream Media over IP, Video Services Forum, http://www.videoservicesforum.org/download/technical_recommendations/VSF_TR-03_2015-11-12.pdf

and the streams could be flexibly composed together by end devices into the final media product.

To transform its recommendation into a complete set of standards around which industry could first rally and then interoperate, the VSF approached SMPTE. SMPTE ST 2110 [43/44/45/46/47], published in 2017, is the resulting suite of standards directed to the support of Professional Media over Managed IP Networks. While the VSF recommendation and SMPTE's initial standards documents were directed to uncompressed elementary streams, more recent work (SMPTE ST 2110-22 [132]) considers application of constant bit rate compression to support certain workflows.

10.2.3. Deployment

The Alliance for IP Media Solutions (AIMS) was formed in early 2016. AIMS, along with the IABM and support from AES, AMWA, EBU, SMPTE and VSF, organized the first IP Interoperability Zone at the IBC conference in September, 2016, to demonstrate interoperability of draft ST 2110. This was followed in March, 2017 by a JT-NM Interop held at the Fox Woodlands, Texas facility in February, 2017. Since then, every NAB and IBC show have held multi-vendor ST 2110 interop demos.

As early as 2014, Game Creek built an outside broadcast (OB) truck unit called "Encore" based on uncompressed media over IP, although at first using a proprietary transport format from Evertz. In 2016, Arena Television built OB trucks that used a mixture of ST 2022-6 and VSF TR-03. One of the first ST 2110 OB trucks was the Swiss tpc production vehicle in 2017, followed by many more including a Tencent Video OB truck in late 2017, a NEP truck in 2018, and in 2019 trucks by Mediaset and Mobile TV Group. Two UHD IP OB trucks from NEP were used for the BBC s 2018 Royal Wedding coverage described in Annex A 16.4. Two UHD IP OB trucks from Arena Television were used for the BBC s 2019 Football Association Challenge Cup coverage described in Annex A Section 16.2. Two UHD IP OB trucks from NEP were used for the BBC's 2018 Royal Wedding coverage described in Annex A 16.4. Two UHD IP OB trucks from Arena Television were used for the BBC's 2019 Football Association Challenge Cup coverage described in Annex A 16.5.

By 2018, compression encoder vendors such as Rohde & Schwarz were adding ST 2110 inputs. Solutions now exist from MediaKind, Synamedia, Harmonic, Vitech, and others.

Some efforts have been undertaken to make available Open Source tools for application development and testing of ST 2110 systems.[4]

The Telemundo Enterprises headquarters facility in Miami, Florida, now known as Telemundo Center, at the time of launch, was the world's largest SMPTE ST 2110 environment. NBCUniversal began the project in 2016 with a deadline of broadcast production of the FIFA World Cup in 2018. Telemundo Center features 13 production studios and seven control rooms. The facility produces scripted episodic content, daily live news, and sports programming. Supporting 12,000 unique HD streams and 150,000 multicast streams across audio and video offered some unique challenges.[5]

Part of the advantage of carrying each essence as a separate stream is that content can more easily be recombined. A different set of language tracks can be selected to accompany a particular video, and the device where it is all drawn together, the encoder, receives what it needs to receive, in sync, and then produces the emission stream, where the right picture, audio (in the right language and format), captions, ratings data, Nielsen codes, SCTE-104 ad insertion codes, etc. are all combined for distribution. A slight change, and the same materials can be ready for a different region.

10.2.4. Technology

The first challenge, when switching to an IP network, is timing, and this is addressed in SMPTE ST 2110-10 [43]. An SDI cable from one point to another had a latency defined by its length. In a packet-switched network, contention for a network connection or bus of a switch, and buffering at any transmitter or receiver along a route could cause the transit time to vary, in some circumstances, packets routed along different paths arrive out of order. For use in supplying two or more media sources, delivering them in sync and without missing deadlines requires careful systems timing relationships be specified with respect to a common clock. Prescribing how distinct devices achieve synchronization to a common master clock over an IP network, and how such master clocks can be distributed and made sufficiently reliable for use in broadcasting, is described in SMPTE ST 2059 [133], based on the IEEE 1588 Precision Time Protocol (PTP) [134].

[4] Ievgen Kostiukevych, Willem Vermost, Pedro Ferreira; "An Open-source Software Toolkit for Professional Media over IP (ST 2110 and more)", https://tech.ebu.ch/docs/groups/list/Live_IP_Software_Toolkit-paper.pdf

[5] Steve Sneddon, Chris Swisher, Jeff Mayzurk; SMPTE Conference – Large Scale Deployment of SMPTE 2110: The IP Live Production Facility

Transport of uncompressed video is described in SMPTE ST 2110-20 [44]. Video carriage is more recently specified for constant bit-rate compression, SMPTE ST 2110-22 [132], which offers opportunities to handle larger or higher frame rate images within the limits of a facility not otherwise equipped for them[6]. Unlike video carriage over SDI, there is no ancillary data space, neither horizontal nor vertical, being carried in these video streams. Instead, such information is carried independently, according to SMPTE ST 2110-40 [47], as a separate element, but the ability to maintain synchronization is maintained.

Audio is likewise not carried with the picture: Standards define audio carriage as PCM audio (ST 2110-30 [46]) and AES3 Transport (ST 2110-31 [126]).

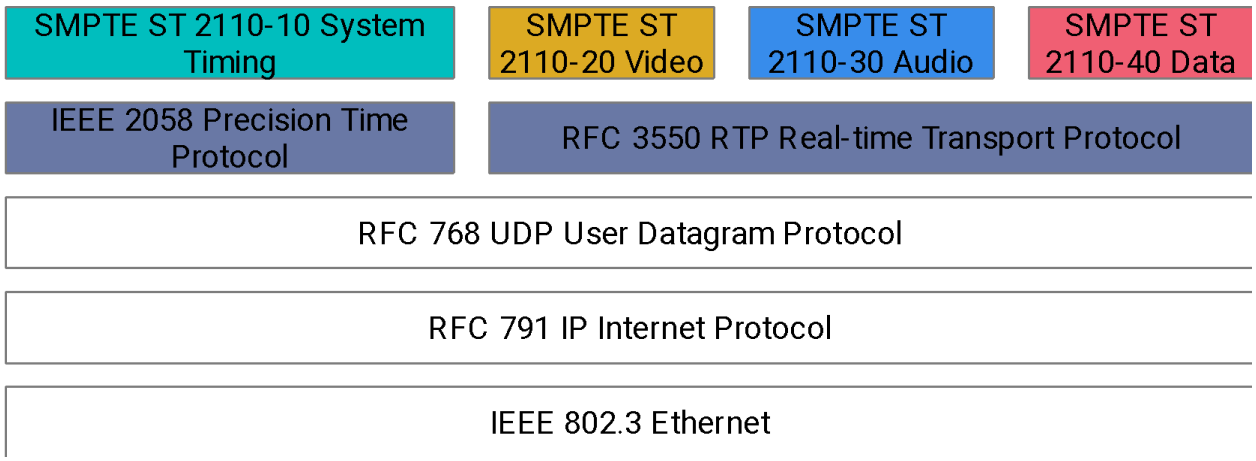

Figure 49. SMPTE ST 2110 protocol stack

The ultimate convergence of the ST 2110 streams and synchronization occurs at the encoder or at any production switch/mix point (i.e., vision mixer, audio console). The fundamental protocol stack of IP-based, separate-stream, professional media is shown in Figure 49. A schematic illustration of an IP-based media plan, though without any redundancy being shown, appears in Figure 50.

[6] Jean-Baptiste Lorent, Antonin Descampe, Charles Buysschaert; SMPTE Conference - Creating Bandwidth-Efficient Workflows with JPEG XS and ST 2110

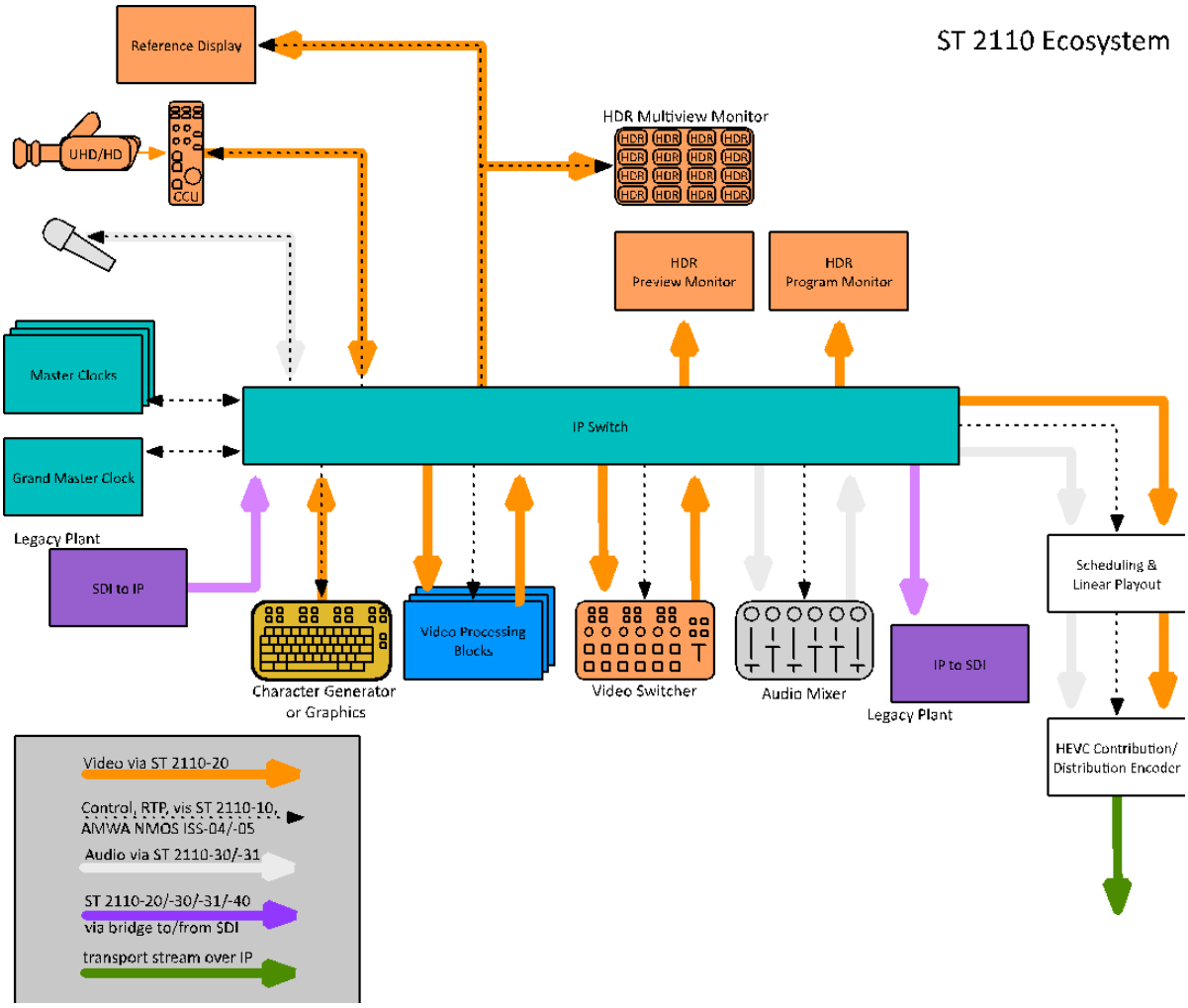

Figure 50. SMPTE ST 2110 live production workflow

Historically, flexibility within a facility was represented by a well-labeled patch panel and a slew of patch cables. AMWA specifies an even more flexible analog for IP-based systems: A suite of technologies able to register a service, facilitate its discovery[7], and create connections among such services[8]. AMWA described registration and discovery with APIs for nodes to register their services and queries for discovering them. Connection management relies on client and server

[7] AMWA IS-04, “NMOS Discovery and Registration Specification”, v1.3

[8] AMWA IS-05, “NMOS Device Connection Management Specification”, v1.1

implementations built to use protocols such as Real-time Transport Protocol (RTP), MQTT,[9*] or WebSockets.

The need for additional support is recognized, relating to establishing a configuration for a facility, monitoring the operational status of each device, and rapidly reconfiguring around a failed element or as a particular need arises, and these technologies are only just emerging.

Whereas traditional media systems could rely on perimeter control of its facilities for much of its security, the transition to an IP-based media system requires reconsidering security requirements. Beyond just requiring secure-HTTP calls, which is a solid start, the EBU recommends a suite of security tests[10] and safeguards[11] for media devices and systems.

The transition of professional media facilities to IP-based systems is relatively new and significant parts of the solution are still in development.

The Alliance for IP Media Solutions (AIMS), a sister organization to the Ultra HD Forum, works to foster a ubiquitous, single, interoperable, standards-based, IP approach for professional media production. AIMS members collaborate to validate, test, and demonstrate equipment and systems.

[9*] A former acronym for “MQ Telemetry Transport”, a communication protocol for the IBM ‘MQ’ series of devices. Now, MQTT is just the name of the protocol.

[10] EBU R 148, “Cybersecurity Recommendation on minimum security tests for networked media equipment”, April 2018

[11] EBU R 143 “Cybersecurity Recommendation for media vendors’ systems, software & services”, April 2016

10.3. NBCU Single-Master Dual-Focused HDR-SDR Workflow Recommendations

10.3.1. Introduction

For movies and early broadcast trials, creating HDR/SDR content versions has required separate grading/shading and mastering processes. As dual live production workflows are impractical, this encouraged the development of a single-stream approach that can deliver both emerging UHD/HDR and legacy HD/SDR platforms simultaneously. The process utilizes the available technology to maximize the dynamic range and color volume in HDR, without compromising the core legacy HD/SDR broadcasts.

In collaboration with Cromorama, and building on ITU working group discussions involving Dolby, BBC and Philips, NBCU has developed single-stream production and distribution techniques. These techniques and supporting conversions have been developed using objective color science as well as traditional real-world testing.

This document provides a potential best practice for broadcast production and distribution processes including conversion. We are sharing our experiences to continue a dialog with our colleagues so that consistent creation, delivery and media exchange of HDR and SDR content is possible. It is also well documented for international exchange in Report ITU-R BT.2408-7 [8].

10.3.2. NBCUniversal Single-Master Dual-Focused HDR-SDR Workflow Guide

Workflow Guide PDF downloadable at this link:
https://github.com/digitaltvguy/NBCUniversal-UHD-HDR-SDR-Single-Master-Production-Workflow-Recommendation-LUTs/blob/main/NBCU%20Single-Master%20UHD-HDR-SDR%20Production-Distribution%20and%20LUTs.pdf

Here is the full repository with LUTs:

https://github.com/digitaltvguy/NBCUniversal-UHD-HDR-SDR-Single-Master-Production-Workflow-Recommendation-LUTs

All the resources for this workflow are available under this table of contents:
https://github.com/digitaltvguy/NBCU-UHD-HDR-SDR-Resources-Table-of-Contents

10.3.3. Reference Files, Tools and Test Patterns

Table 7. NBCU References

NBCU LUTS	Download of the NBCU LUT files as referenced in this annex. NBCU Single-Stream Recommendations
Vooya Video Player	Video player software from Vooya
Color Metric Plug-in for Vooya	Plug-in for the Vooya Video Player that provides the Objective Color Metrics used in this annex. https://www.offminor.de/plugins.html

11. Monographs on Encoding

(reserved for future encoding technologies)

12. Monographs on Distribution

Terrestrial broadcast, satellite delivery, fiber and cable utilities, cellular data, and over-the-top (OTT) internet delivery are how billions receive video media. Transitioning such systems to support Ultra HD technologies is not a quick process, but the motivation exists for higher quality, improved experience. These monographs present transitions taking place today, as more services are becoming available (see a list of known services at ultrahdforum.org/uhd-service-tracker/). The monographs here represent distribution systems being standardized by international and national organizations. To date, we have not had contributions for OTT systems, as these are often proprietary and thus not subject to standardization. The monographs on distribution are:

- ATSC3.0 (Section 12.1) .
- Brazilian Roadmap to UHD (Section 12.2) .
- DVB-T2 UHD (Section 12.3) .

12.1. ATSC 3.0

The Advanced Television Systems Committee, Inc. (ATSC) is an international, non-profit organization developing voluntary standards for digital television. Member organizations represent broadcast, broadcast equipment, motion picture, consumer electronics, cable, satellite, and semiconductor industries. The first suite of standards for digital television from the ATSC was adopted beginning in 1996 throughout North America and Korea, plus a few other countries in South America, the Caribbean, and South Pacific.

ATSC 3.0 is a new suite of standards and recommended practices that emerged beginning in 2016. The goal of the ATSC was to develop an advanced and future proof digital television standard, supporting new television formats as well as enabling new functions and features. The decision was taken to not support backward compatibility with current ATSC standard, thus opening the door to the best available technology and thus due to fundamental technology differences, is incompatible with prior ATSC systems. In the discussion below, the ATSC standards referenced can be found here: https://www.atsc.org/standards/.

12.1.1. Why ATSC 3.0?

ATSC 3.0 diverges from earlier designs to allow substantial improvements in performance, functionality, and efficiency sufficient to warrant implementation of a non-backwards-compatible system. With higher capacity to deliver Ultra HD services, robust reception on a wide range of

devices (including mobile), improved efficiency, IP transport, advanced emergency alerting, personalization features, and interactive capability, the ATSC 3.0 Standard provides much more capability than previous generations of terrestrial broadcasting.

Elements of the ATSC 3.0 System are defined in separate standards, to facilitate flexibility and extensibility. In some cases, there's more than one way to do things and a broadcaster can choose whichever best suits their preference or operations.

12.1.2. Deployment

Korean broadcasters have been on-air with ATSC 3.0, starting in the Seoul metropolitan area since 2017, largely motivated by the timing of the 2018 Winter Olympics in Pyeongchang and simplified by the Korean government allocating 30MHz of spectrum dedicated for five new channels for advanced broadcasting services. Notably, while Korean broadcasters are actively producing and supplying 4K broadcasts, they are not yet produced and delivered with high dynamic range. The Korean standard includes the use of MPEG-H 3D audio compression standard. For additional details, see Sec 5.1 on Digital Terrestrial Transmission.

In the United States, the situation has been more complex, as reallocation of broadcast spectrum was undertaken in a multi-phase incentive auction and repacking. With that substantially complete, large-scale broadcast trials ran through 2019, primarily in Baltimore, Indianapolis, and Phoenix, with commencement of commercial ATSC 3.0 services expected in 2020. Accordingly, the Consumer Electronics Show in January 2020, which dubbed ATSC 3.0 as "Next Gen TV", ATSC 3.0-ready products for the North American market were first presented en masse.

By the end of 2020, the rollout ATSC 3.0 services supporting full-time Ultra HD formats had expanded to eleven markets with over twice as many more expected in 2021. Note that interactive information about Ultra HD services, including the ATSC 3.0 rollout, is available from the Ultra HD Forum Service Tracker, available here:
https://ultrahdforum.org/uhd-service-tracker/

12.1.3. Technology

Fundamentally, ATSC 3.0 is a layered architecture, as shown in Figure 89. The physical layer describes how bits are sent in a 6 MHz channel[12] with which any ATSC 3.0 receiver can discover available services by detecting a predetermined pattern indicating the "bootstrap signal", which describes parameters of the immediately following physical layer frame in the

[12] Doc. A/322:2020, ATSC Standard: A/322, Physical Layer Protocol [52]

digital transmission signal, e.g., robustness and bandwidth, essential for decoding the rest of the frame.[13]

Among the protocols of the Organizational layer are the ATSC Link-Layer Protocol (ALP)[14], two methods of broadcast Service delivery[15]: MPEG Media Transport[16] (MMT) and the DASH-IF profile[17], a Studio-to-Transmitter Link (STL) and corresponding tunneling protocol[18], and many others.

ALP defines the path into and out of the physical layer for IP packets, link-layer signals, and MPEG-2 Transport Stream packets, and optimizes for useful data by overhead reduction mechanisms. The STL subsystem maps APL packets to Physical Layer Pipes (PLPs) as dictated in detail by a Scheduler, allowing separate parallel streams of content to be multiplexed into the same bitstream.

[13] Doc. A/321:2016, ATSC Standard: A/321, System Discovery and Signaling.

[14] Doc. A/330:2019, ATSC Standard: A/330, Link-Layer Protocol

[15] Doc. A/331:2019 [53], ATSC Standard: A/331, Signaling, Delivery, Synchronization, and Error Protection

[16] ISO/IEC 23008-1:2017, Information technology - High efficiency coding and media delivery in heterogeneous environments - Part 1: MPEG media transport (MMT)

[17] “Guidelines for Implementation: DASH-IF Interoperability Points for ATSC 3.0” DASH Interoperability Forum, https://dashif.org/guidelines

[18] Doc. A/324:2018, ATSC Standard: A/324, Scheduler/Studio to Transmitter Link

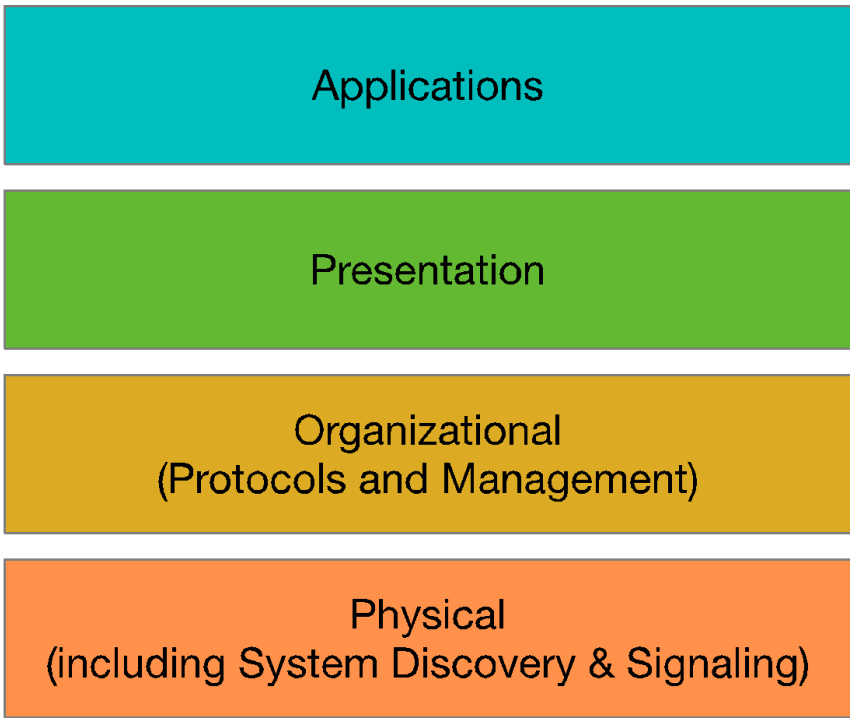

Figure 89. ATSC 3.0 architecture layers

ATSC 3.0 specifies two methods of delivery of real time objects (video, audio, captions); MMT (MPEG Multimedia Transport) and DASH (Dynamic Adaptive Streams over HTTP). In addition, there are two specified transport protocols, MMTP and ROUTE (ATSC A/331 [53]) (Real time Object Uni-directional Transport). While either protocol can be used for the broadcast service, the transport protocol for broadband is restricted to DASH-IF over HTTP/TCP/IP. See Figure 90.

In the presentation layer, video is encoded as HEVC, ATSC A/341 [54] and may include certain advanced aspects including SHVC and temporal sub-layering, each representing a base layer image plus an optionally received and decoded enhancement layer providing either higher resolution or higher frame rate. ATSC has provided certain constraints with respect to these features, so as to avoid implementations having a level of complexity deemed unnecessary. Audio is encoded in either of two formats: AC-4, ATSC A/342-2 [56] and MPEG-H, ATSC A/342-3 [57], which have so far been adopted regionally, with Korea having chosen MPEG-H and North America opting for AC-4. Carriage for captions and subtitles[19] and interactive content[20] is also included.

[19] Doc. A/343:2018, ATSC Standard: Captions and Subtitles

[20] Doc. A/344:2019, ATSC Standard: Interactive Content

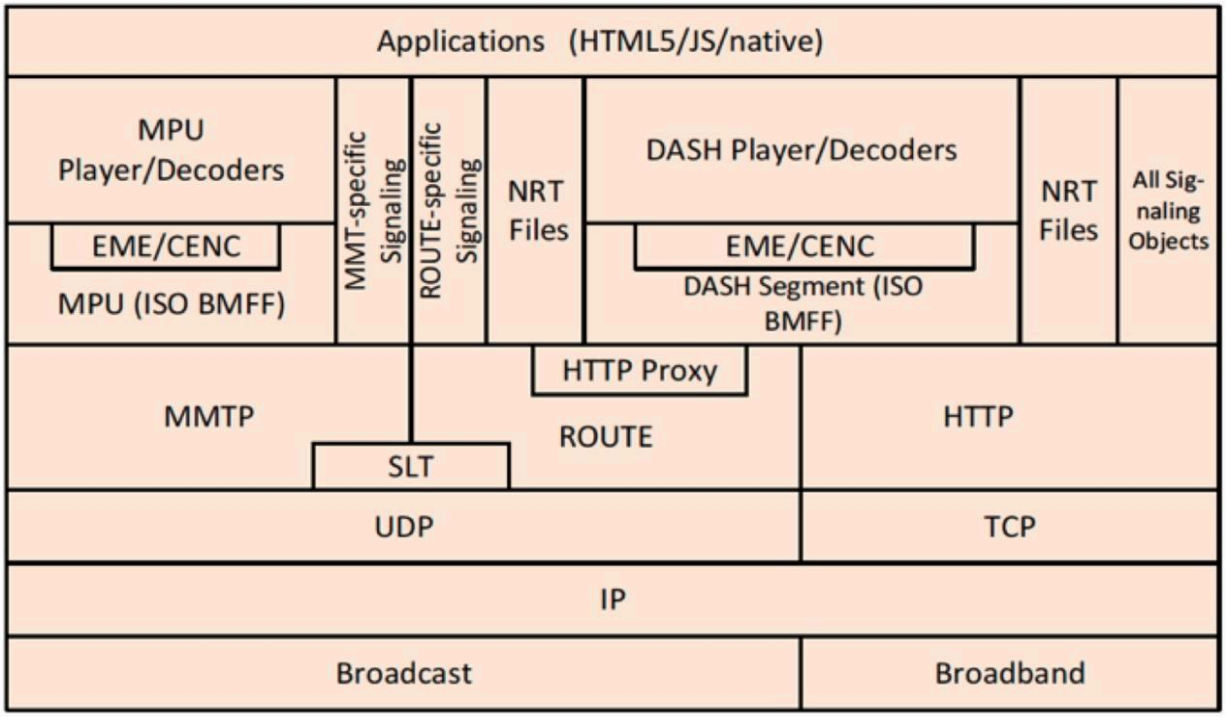

Figure 90. ATSC 3.0 Conceptual Protocol Stack

Options are also included for securing such signaling and contents[21] by both authenticating the contents themselves and the recipient's rights to access them. Security is based on a chain of trust based on provisioning of X.509 certificates. Separately, watermarking of audio and video are also specified. A set of best practices for implementers is also available.[22]

Applications are built in a standardized runtime environment, and broadcasters may author interactive applications that can render supplemental content delivered via broadcast or broadband. As an example, this could display additional content regarding an emergency service message. (Support for emergency messages is spread over multiple standards in the ATSC 3.0 suite.) Applications can make use of companion devices[23], e.g., smart phones or tablets, too.

[21] Doc. A/360:2019, ATSC Standard: ATSC 3.0 Security and Service Protection

[22] Doc. A/362:2020, ATSC Recommended Practice: Digital Rights Management

[23] Doc. A/338:2019, ATSC Standard: Companion Device

12.1.4. ATSC 3.0 and OTT services

From the ground up, the design of ATSC 3.0 was developed with a plan to align with IP streaming technologies and standards. While the broadcast component is a ‘one to many’ protocol, the underlying use of DASH protocols allows ATSC 3.0 to be delivered via the Internet, with or without a broadcast component. There have been tests of this concept by at least one major broadcast station operator with help from streaming software development companies. In these tests an ATSC 3.0 enabled television demonstrated the ability to tune to an IP delivered ATSC 3.0 stream delivering the advanced services and formats that ATSC 3.0 supports. This ‘OTT’ only approach may enable coverage to areas where the 3.0 broadcast signal is not available, and thus extend the coverage of the broadcaster.

ATSC 3.0 can also be operated in a ‘hybrid’ mode, with both the OTA and OTT delivery of content and features. ATSC 3.0 provides an ability to synchronize the two streams so that additional program components such as alternate languages, can be delivered over the Internet and reproduced in alignment with the OTA delivered components (main video and audio). With the addition of an IP connection to the TV, the reverse channel protocols can be enabled, thus facilitating interactivity and program enhancement features as well as the possibility of targeted advertising and audience measurement data.

12.2. Brazilian Roadmap to UHD

Following Brazil's lead in 2007, a total of fifteen South American countries have adopted Sistema Brasileiro de Televisão Digital (SBTVD)[24] for terrestrial broadcasting. Also known as Integrated Services Digital Broadcasting, Terrestrial, Brazilian version (ISDB-Tb) and less formally as "TV 2.0", it is based upon the Japanese digital television standard, ISDB-T, plus some additional technical features and improvements.

The SBTVD Forum[25] is a non-profit organization with the mission to "guarantee every Brazilian the right to receive a broadcast with high quality images and sound". The Forum acts as a technical advisory arm of the Brazilian government in matters related to Open Digital TV. Its activities focus on the implementation of digital television throughout the country, in the development of technical standards for Open Digital TV, and in the continuing improvement of the Open TV service offered to the population.

In Brazil, terrestrial TV promotes cultural diversity and national integration through its primary characteristic – it's free of charge, hence the moniker "Open TV". Even as Brazil completed the transition of its national mass-communication platform from analog to the digital SBTVD/TV 2.0, the home of Carnival, frequent FIFA champions, and their avid fans found the allure of high dynamic range images and immersive audio to be irresistible! With the ultimate goal of attaining Ultra HD features, Brazil envisioned a sustainable, pragmatic, two-step evolution.

12.2.1. Sustainable, Pragmatic Progression: TV 2.0/2.5/3.0

In 2019, the SBTVD Forum issued a call for proposals for an HDR broadcast system providing full video backward compatibility with their recently deployed digital broadcast system. This imposed the TV 2.0 constraints of AVC/H.264 operating with 8-bit samples. After a competitive evaluation of the candidate technologies, most represented within these Guidelines, those selected for "TV 2.5" were standardized in the Associação Brasileira de Normas Técnicas

[24] Normas Técnicas – Sistema Brasileiro de TV Digital Terrestre, ABNT NBR 15601-15608, 15610, https://forumsbtvd.org.br/legislacao-e-normas-tecnicas/normas-tecnicas-da-tv-digital/english/

[25] https://forumsbtvd.org.br/legislacao-e-normas-tecnicas/normas-tecnicas-da-tv-digital/english/

(ABNT) as of May, 2020, amending half a dozen of Brazil's Open TV standards [26] [27] [28] in little more than a year.

The TV 2.5 step is both sustainable and practical because some regions of the country have only just transitioned to the digital broadcast of TV 2.0, and both broadcasters and consumers have essentially new equipment. At the time TV 2.5 was ratified, the equipment of the earliest adopters might have been only 13 years old, with a remaining service life usable in through the transition to TV 2.5, broadcasters and consumers both being able to self-select their individual transitions to HDR programming.

SBTVD immediately issued a request for proposal for TV 3.0 technologies, and for the technologies proposed, undertook a detailed test, evaluation, and selection process, concluding in July, 2024. The resulting system allows distribution of current and future formats of up to 8K resolution to broadcast and OTT networks. Deployment could begin as early as 2025.

The TV 3.0 system embodies sustainability features such as best-in-class efficiency, with the ATSC 3.0 physical layer having bit coding and modulation that operate near the Shannon limit and VVC/H.266 video coding representing about a 78% bitrate savings over TV 2.0 for video of similar quality. Additionally, because TV 2.5 is in-place with an ability to support HDR, programming originated for TV 3.0 still has a pathway to legacy devices. Meanwhile, devices newly coming into market could be required to support reception of TV 3.0, so as to ease an eventual phase out of TV 2.0/TV 2.5 system in a decade or more.

12.2.2. Deployment – TV 2.5

The component technologies selected for TV 2.5 were already supported by chipsets from multiple manufacturers, for both set-top boxes and televisions. The unique feature of TV 2.5 is the combination of these technologies. Further, video decoder IP was available for manufacturers wanting to design their own SoC. On the production side, encoders were likewise available from multiple manufacturers.

ISDB-Tb had already seen widespread deployment, being used widely in Brazil, Argentina, Uruguay, Peru, Chile, Venezuela, Ecuador, Costa Rica, Paraguay, Bolivia, and Honduras. Further, SBTVD has been officially adopted, but yet to be broadcast, in these countries:

[26] Televisão digital terrestre - Codificação de vídeo, áudio e multiplexação (Video, Audio, and Multiplex Encoding), ABNT NBR 15602-1, -2, -3:2007 Emenda 1:2020.

[27] Televisão digital terrestre — Multiplexação e serviços de informação (Multiplexing and Information Services), ABNT NBR 15603-1:2007, -2:2017 Emenda 1:2020

[28] Televisão digital terrestre — Receptores (Receivers), ABNT NBR 15604-1:2018, Emenda 1:2020

Nicaragua, Guatemala, Honduras, and El Salvador, and in Africa, Botswana and Angola. The backward compatible nature of TV 2.5 ensures that these systems will remain supported as TV 2.5 rolls out.

Production-wise, in July 2016, TV Globo, Brazil and Latin America's largest network, had already offered a popular 10-episode miniseries "Ligações Perigosas" entirely shot and produced in 4K HDR through its OTT service, witnessing the availability of native HDR production in the region.

12.2.2.1. Rede Amazônica started TV 2.5 regular broadcast with MPEG-H Audio

In December 2021, Grupo Rede Amazônica became the first broadcaster in Latin America to provide a 24/7 MPEG-H Audio service on one of their regular terrestrial broadcast channels using ISDB-Tb TV 2.5[29].

Brazil's TV 2.5 standards include next-generation audio (NGA) technologies, including immersive and personalized audio. A major advantage of NGA is that its personalization features benefit immersive as well as 5.1 surround and stereo productions. MPEG-H Audio has been included in the TV 2.5 standard, particularly due to its capabilities to deliver immersive sound and, even more important, its advanced personalization and accessibility options within a single production.

[29] Grupo Rede Amazônica rolls out MPEG-H Audio on their free-to-air Amazon Sat broadcast channel using ISDB-Tb TV2.5, https://www.iis.fraunhofer.de/en/pr/2021/20211203_mpeg_h_amazon_sat.html

Figure 91. Rede Amazonica program with MPEG-H Audio personalization

The system enables the audience to interact with the content and adapt the audio experience to their preferences. The option to personalize the balance of dialogue and background sounds, choose between different languages, and position the audio description at different locations in the room, makes for a highly individualized experience for everyone while at the same time ensuring an unmatched degree of accessibility for people with visual or hearing impairments.

Sound producers and editors are already creating new, exciting NGA content using MPEG-H while using their existing tools and workflows with minor modifications to produce stereo, 5.1, and immersive content with the full range of personalization options.

12.2.2.2. Globo and TV 2.5 Readiness-Paris Olympic Demo

During the 2024 Paris Olympics, Globo showcased content from the Paris Olympics, featuring HDR and NGA enhancements, playing on Hisense TV models complying with the country's TV

2.5 standard, to illustrate how a backward-compatible emission can deliver a premium experience for live sports.[30]

12.2.3. Technology

For the TV 2.5 solution, some elements of SBTVD, such as broadcast A/V encoding (ABNT NBR 15602)[31], service information (ABNT NBR 15603)[32], receiver specifications (ABNT NBR 15604) [33], and operational guidelines (ABNT NBR 15608), are amended to remain fully compatible with Brazil's just-built, Open Digital TV infrastructure. HLG, as a transfer function to address both SDR and HDR receivers simultaneously, is allowed. However, those more concerned with creative intent, wanting native BT.709 SDR programming and receivers to experience an uncompromised image, will instead use the signal flow shown in Figure 92 (compared to Figure 8 SL-HDR processing, distribution, reconstruction, and presentation). This provides SL-HDR1 dynamic metadata atop the AVC SDR transmission, for a robust, high-quality HDR reconstruction of both native HDR programming and real-time ITM up-conversions, supporting receivers that enable the new, more immersive experience.

Brazil's TV 2.5 demonstrates how component technologies, promoted by the UHDF ~~Ultra HD Forum~~ during its first five years supporting the industry, have been adapted to create a new, unique platform that is fully backwards compatible with Brazil's Open TV, as laid out by SBTVD

[30] See: https://www.tvtechnology.com/news/globo-shows-off-tv-25-enhancements-on-hisense-tvs-during-paris-olympics

[31] Multiplexação e serviços de informação (Service Information) ABNT NBR 15603

[32] Televisão digital terrestre - Receptores, (Receivers) ABNT NBR 15604

[33] Guia de operação (Operational Guidelines) ABNT NBR 15608

and standardized by ABNT.

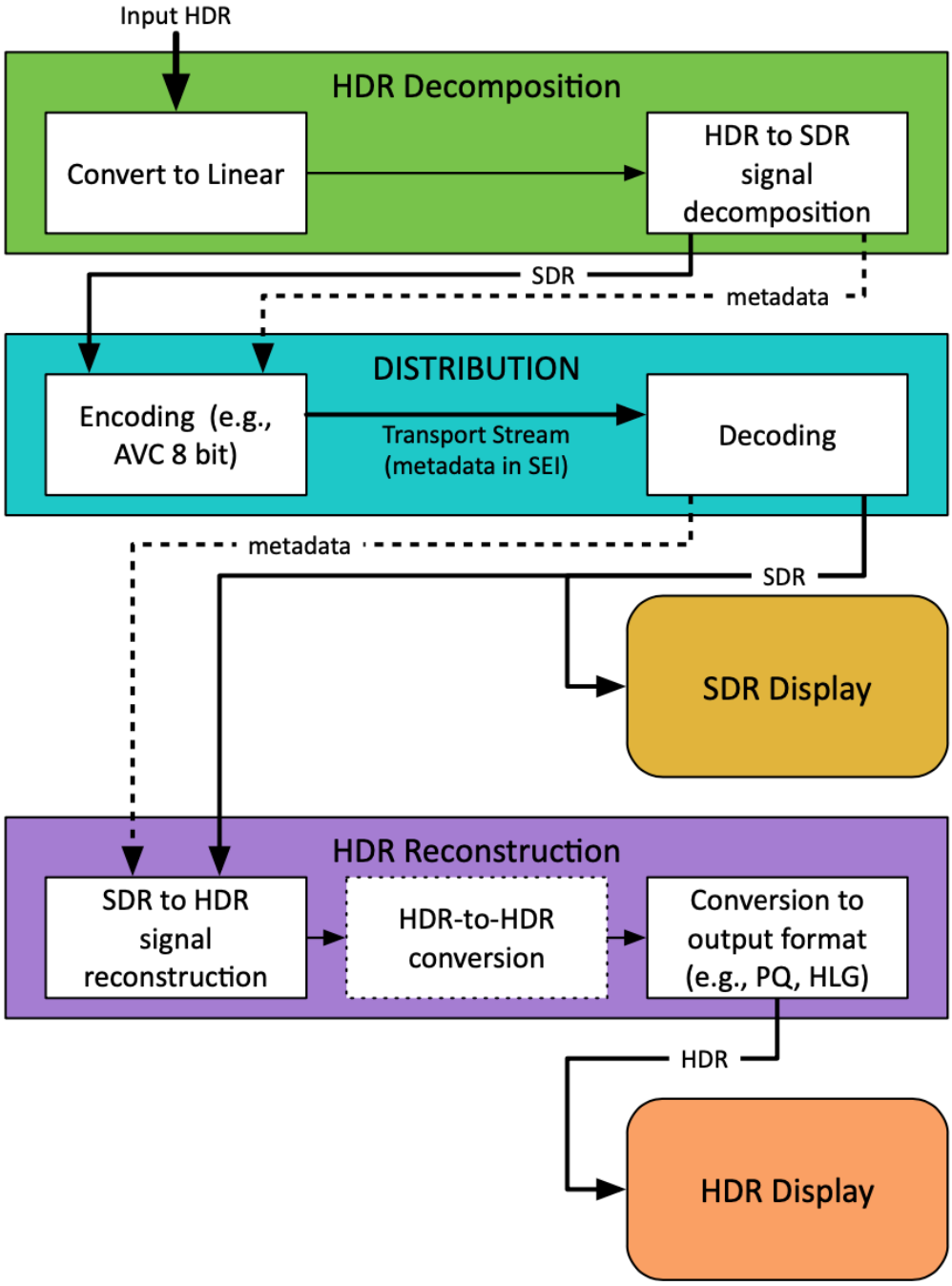

Figure 92. Signal flow for Brazil's TV 2.5

12.2.4. Timeline for TV 3.0

After a lengthy RFP process where Brazil's TV 3.0 project has tested different technologies for the distribution of current and future formats up to 8K resolution for both broadcast and OTT networks, the results of their testing and selection of technologies has been announced in their blog.[34]

TV 3.0 is scheduled to be on-air by the end of 2025, as the standardization and frequency allocations finalize before the end of 2024. A summary of the different technologies selected is summarized in the Figure 93 below.

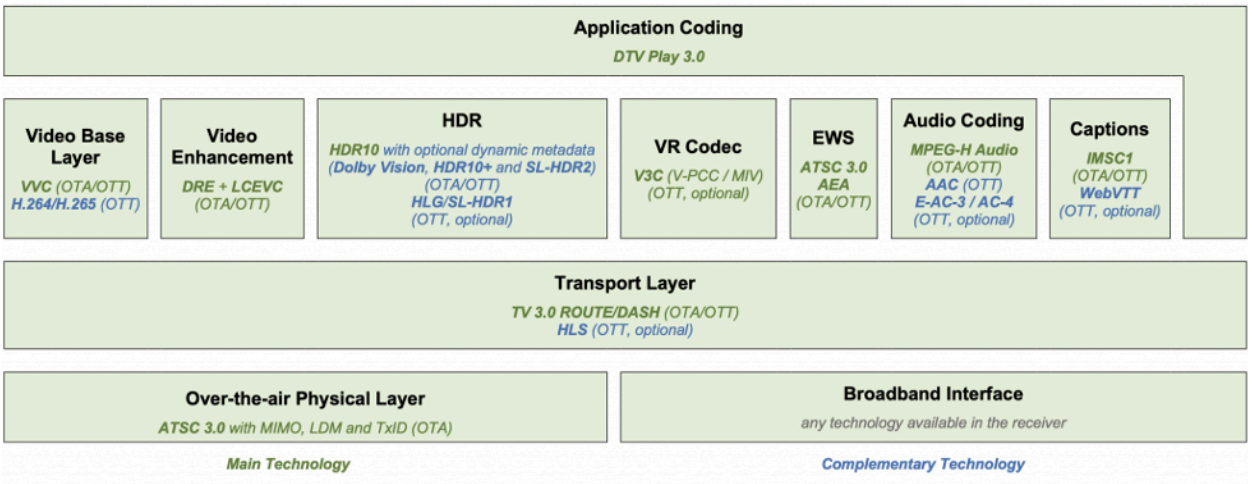

Figure 93. Overview of the TV 3.0 selected technologies for each layer

A comparison of the different technologies selected vs. what currently exists in the Ultra HD Forum Guidelines is shown in Table 7.

[34] Other references for Brazil TV 3.0:
1. SBTVD Forum - TV 3.0 - CfP Phase 2 / Testing and Evaluation, https://forumsbtvd.org.br/tv3_0/
2. The "TV 3.0 CfP Phase 2 / Testing and Evaluation", https://forumsbtvd.org.br/wp-content/uploads/2021/03/SBTVD-TV_3_0-P2_TE_2021-03-15.pdf

Table 7. Ultra HD Forum Guidelines vs. Brazil 3.0 Specification

Parameter	Ultra HD Forum Guidelines	TV 3.0 specification	Note
Resolution	2160p120 max	up to 8K	frame rates (by resolution) TBD
Color space	BT 2020	BT 2020	
HDR	PQ10 + optional: dynamic metadata (DV, SL-HDR2); HLG; SL-HDR1	**OTA/OTT**: PQ10 + optional: dynamic metadata (DV, SL-HDR2, HDR10+) **OTT** optional : SL-HDR1/HLG	
Video codec	HEVC	**OTA/OTT**: VVC **OTT:** AVC/HEVC	
Video enhancement	**OTT**: CAE	**OTA/OTT**: DRE + LC EVC	
Audio	ATMOS / MPEG-H AAC, E-AC-3+JoC, AC4	**OTA/OTT**: MPEG-H **OTT:** AAC/(E) AC-3, optional AC-4	
Captions	CTA 708/608, SCTE 27, IMSC1, DVB Subtitles	**OTA/OTT**: IMSC1 **OTT**: optional WebVTT	
Transport	Not specified for OTA OTT: HLS/DASH	**OTA/OTT**: ROUTE/DASH **OTT:** optional HLS	

Here is a brief analysis the Brazil 3.0 specification:

- HDR: PQ10 is the only HDR format for the OTA workflow, with dynamic metadata variants (DV, HDR10+, SL-HDR2) allowed as an option. OTT further allows, as options, HLG and SL-HDR1.
- Video Codec: VVC is the only video codec adopted for TV 3.0. Note that VVC is not yet part of the Forum's Guidelines at the time of writing, but is anticipated once documented in international broadcast and streaming standards. Encoding tools, such as DRE (Dynamic Resolution Encoding) and LC EVC are expected to provide additional gains (TBD) when combined with VVC. OTT further requires support for AVC and HEVC, which are part of the Guidelines.
- Audio: MPEG-H is the immersive audio/codec selected for OTA. OTT further requires support for other audio codecs: AAC, (E)AC-3, and allows AC-4 as an option.
- Transport layer: DASH-ROUTE transport layer is adopted and aligned with the ATSC 3.0 specifications for OTA. OTT further allows an HLS option.

The next phase of the TV 3.0 project is to write a detailed specification for all system elements. Once the specification is ratified as an international standard, the Ultra HD Forum will include any new technologies in future Guidelines.

The following Ultra HD Forum companies were involved in the TV 3.0 project: ATEME Dolby, Fraunhofer, Harmonic, and Interdigital. Table 36 below indicates the parts of the TV 3.0 contributed by the Ultra HD Forum members.

Table 8. Ultra HD Forum Member Contribution to Brazil TV 3.0 Project

Members	TV 3.0 contribution
ATEME	VVC
Fraunhofer	VVC MPEG-H HDR Dynamic Mapping
Interdigital	VVC MPEG-H HDR Dynamic Mapping
Harmonic	DRE LC EVC
Dolby	Dynamic Mapping AC-4

12.3. DVB-T2 UHD

12.3.1. What is DVB

Founded in 1994, DVB is an international industry consortium that produces technical specifications for digital television. Published by ETSI as open standards, DVB's specifications that define the physical and data link layers for the transmission of broadcast digital television – over satellite, cable, and terrestrial networks – serve more than 1.5 billion receivers around the world. DVB specifications also cover broadband delivery, targeting both hybrid and broadband-only devices.

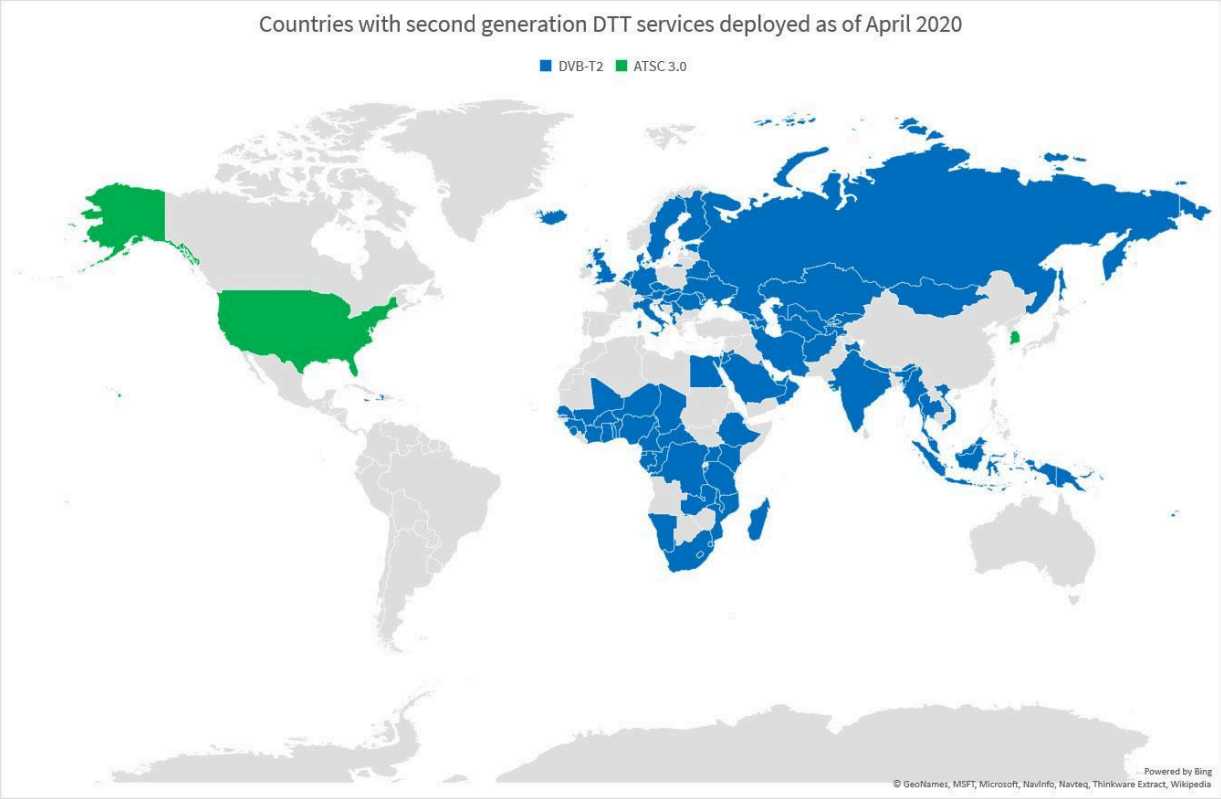

Figure 94. Global Deployment of 2nd Gen DTT Services

DVB's second-generation terrestrial delivery specification is DVB-T2. First published in 2008, the most recent version[35] was published in July 2015.

12.3.2. Global Deployment of Second Generation DTT

DVB-T2 services are on air in at least 93 countries around the world, with a combined population of about 3.5 billion people. Coverage comprises large parts of Europe, Africa and Asia.

12.3.3. Technology in Use

DVB-T2 primarily defines the physical layer of the related terrestrial broadcasting system. Following the principle of layer independence, there is a choice regarding the layers above the physical layer and DVB offers several options.

Implementations today, in June 2020, are based on the MPEG-2 Transport Stream [1] – using the protocol stack outlined below:

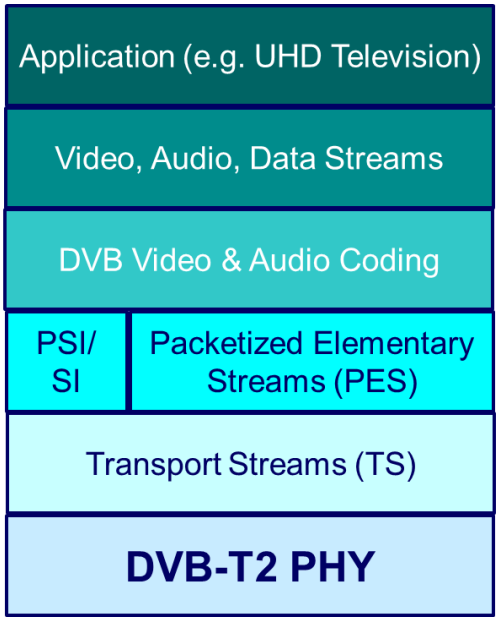

Figure 95. DVB-T2 Protocol Stack

[35] ETSI EN 302 755 V1.4.1: "Digital Video Broadcasting (DVB); Frame structure channel coding and modulation for a second generation digital terrestrial television broadcasting system (DVB-T2)"

The relevant standards making up the layers of this TS-based protocol stack are listed in clause 1.6 References, namely DVB Video and Audio Coding [63]; MPEG-2 Transport Streams, Packetized Elementary Streams and MPEG Program-specific Information (PSI); DVB Service Information[36]; and DVB-T2.

The performance of the DVB-T2 physical layer is set out in the table below (taken from[37]), outlining the Signal-to-Noise Ratios required under different channel conditions, depending on the chosen constellations and LDPC code rates. All constellation and code rate options are listed. Note that this table covers only implementations using the long codeword length.

A typical DVB-T2 set-up, addressing stationary rooftop reception with directional aerials in a Multi-Frequency Network (MFN), deploys a 256-QAM constellation and an LDPC code rate of 2/3. This provides a net throughput of about 40 Mbits/s in an 8 MHz RF channel. For the estimation of the number of UHD services that can be provided per multiplex, this throughput figure is taken as the starting point.

12.3.4. Suitability for UHD

A typical DVB-T2 set-up addressing stationary rooftop reception enables a net data throughput of about 40 Mbit/s in an 8 MHz channel – see the previous chapter. For UHD services, DVB's Video and Audio Coding specification[38] currently recommends the use of the HEVC codec. One or more next generation video codecs, which will deliver bit-rate savings of at least 35%, will be added to the DVB specification in due course. The published work plan targeted mid-2021 for the completion of this work.

The average bitrate for a UHD HEVC service using statistical multiplexing is in the range of 10 - 13 Mbit/s.

On that basis, the number of UHD services per multiplex with the settings outlined above would be in the order of three to four (noting that audio needs to be provided as well).

Meanwhile as not all services would have UHD content available all the time, and as spectrum capacity is often limited, countries planning such deployments plan envisage service switching

[36] EN 300 468 V1.16.1: "Digital Video Broadcasting (DVB); Specification for Service Information (SI) in DVB systems"

[37] ETSI TS 102 831 V1.2.1: "Digital Video Broadcasting (DVB); Implementation guidelines for a second generation digital terrestrial television broadcasting system (DVB-T2)"

[38] TS 101 154 V2.6.1 [63]: "Digital Video Broadcasting (DVB); Specification for the use of Video and Audio Coding in Broadcast and Broadband Applications"

seamlessly between 1080p50 and 2160p50 content in an integrated service-capacity-sharing scheme.

Outlined below are a few possible combination scenarios with different numbers of UHD and HD services in a DVB-T2/HEVC multiplex at any given time assuming:

- 40 Mbit/s net throughput per 8 MHz multiplex
- 10 Mbit/s per VBR encoded 2160p50 service
- 3.25 Mbit/s per a VBR encoded HD 1080p50 service (see implementations in the field)

LDPC block length: 64 800 bits

Constel-lation	Code rate	Spectral Efficiency (see note 2)	Required $(C/N)_0$ (dB) for BER = 1×10^{-7} after LDPC decoding			
			Gaussian Channel (AWGN)	Ricean channel (F_1)	Rayleigh channel (P_1)	0 dB echo channel @ 90 % GI
QPSK	1/2	*0,99*	*1,0*	*1,2*	*2,0*	*1,7*
QPSK	3/5	*1,19*	*2,3*	*2,5*	*3,6*	*3,2*
QPSK	2/3	*1,33*	*3,1*	*3,4*	*4,9*	*4,5*
QPSK	3/4	*1,49*	*4,1*	*4,4*	*6,2*	*5,7*
QPSK	4/5	*1,59*	*4,7*	*5,1*	*7,1*	*6,6*
QPSK	5/6	*1,66*	*5,2*	*5,6*	*7,9*	*7,5*
16-QAM	1/2	*1,99*	*6,0*	*6,2*	*7,5*	*7,2*
16-QAM	3/5	*2,39*	*7,6*	*7,8*	*9,3*	*9,0*
16-QAM	2/3	*2,66*	*8,9*	*9,1*	*10,8*	*10,4*
16-QAM	3/4	*2,99*	*10,0*	*10,4*	*12,4*	*12,1*
16-QAM	4/5	*3,19*	*10,8*	*11,2*	*13,6*	*13,4*
16-QAM	5/6	*3,32*	*11,4*	*11,8*	*14,5*	*14,4*
64-QAM	1/2	*2,98*	*9,9*	*10,2*	*11,9*	*11,8*
64-QAM	3/5	*3,58*	*12,0*	*12,3*	*14,0*	*13,9*
64-QAM	2/3	*3,99*	*13,5*	*13,8*	*15,6*	*15,5*
64-QAM	3/4	*4,48*	*15,1*	*15,4*	*17,7*	*17,6*
64-QAM	4/5	*4,78*	*16,1*	*16,6*	*19,2*	*19,2*
64-QAM	5/6	*4,99*	*16,8*	*17,2*	*20,2*	*20,4*
256-QAM	1/2	*3,98*	*13,2*	*13,6*	*15,6*	*15,7*
256-QAM	3/5	*4,78*	*16,1*	*16,3*	*18,3*	*18,4*
256-QAM	2/3	*5,31*	*17,8*	*18,1*	*20,1*	*20,3*
256-QAM	3/4	*5,98*	*20,0*	*20,3*	*22,6*	*22,7*
256-QAM	4/5	*6,38*	*21,3*	*21,7*	*24,3*	*24,5*
256-QAM	5/6	*6,65*	*22,0*	*22,4*	*25,4*	*25,8*

NOTE 1: Figures in italics are approximate values.
NOTE 2: Spectral efficiency does not take into account loss due to signalling / synchronization / sounding and Guard interval.
NOTE 3: The BER targets are discussed above.
NOTE 4: The expected implementation loss due to real channel estimation needs to be added to the above figures (see clause 14.5). This value will be significantly less than the corresponding figure for DVB-T in some cases, due to better optimisation of the boosting and pattern densities for DVB-T2.
NOTE 5: Entries shaded blue are results from a single implementation. All other results are confirmed by multiple implementations.

Figure 96. Performance of DVB-T2 PHY Layer

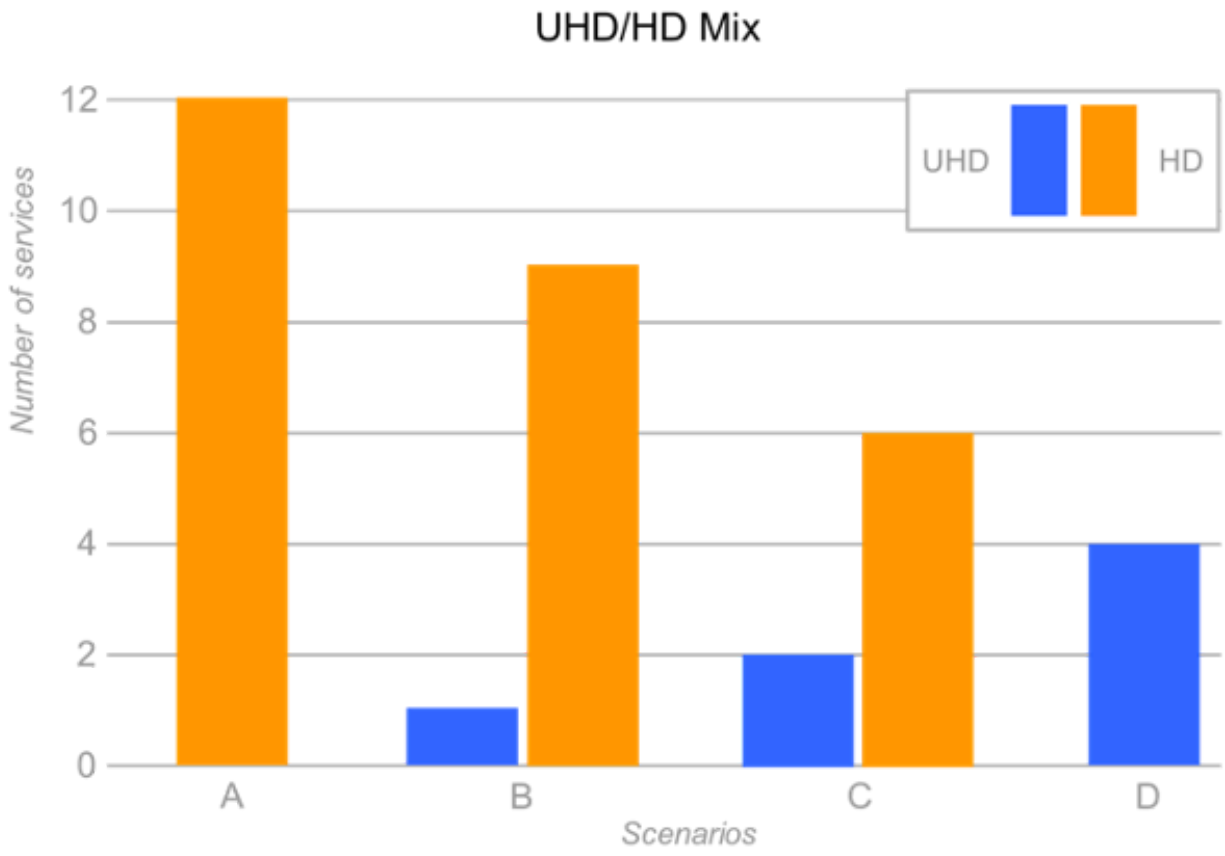

Figure 97. Possible UHD/HD Service Multiplexes using DVB-T2

12.3.5. Conclusions

Providing UHD services via DVB-T2 networks is undoubtedly a realistic option, with all required components in place. Most UHD TV sets offered in Europe in 2020 are equipped to reproduce HEVC-encoded UHD services received via DVB-T2 (and this has been the case for several years already). It is, therefore, rather a matter of content and service providers taking advantage of this opportunity.

With regard to UHD services on DVB-T2 networks, planning is underway in some countries and regions. Publicly available information at the time of writing (June 2020) includes the following:

- The NorDig group, which publishes guidelines for terrestrial television services in the Nordic countries and Ireland, has defined the specifications to be used for a UHD terrestrial service[39].
- France is in the process of finalizing the specification of the DVB-T2 UHD service that should be launched at the Paris Olympics in 2024. The roadmap is being reviewed by the CSA (French regulator for television)[40].

12.3.6. Additional DVB-T2 References

1. UHD / HDR SERVICE PARAMETERS TECH 3372: https://tech.ebu.ch/docs/tech/tech3372.pdf
2. ETSI EN 300 468 v1.17.1: DVB Blue Book A038 (June 2019) Digital Video Broadcasting (DVB); Specification for Service Information (SI) in DVB systems https://www.dvb.org/resources/public/standards/a038_tm1217r37_en300468v1_17_1_-_rev-134_-_si_specification.pdf

[39] NorDig Unified Requirements for Integrated Receiver Decoder s for use in cable, satellite, terrestrial and managed IPTV based networks Date: 27 October 2018, version 3.1. https://nordig.org/wp-content/uploads/2018/10/NorDig-Unified-Requirements-ver.-3.1.pdf

[40] Modernizing the digital terrestrial television (DTT) platform: http://www.csa.fr/content/download/246706/651929/file/CSA_AvenirTNT%20Feuille%20de%20route%20(vEN.3).pdf

13. References

[1] Recommendation ITU-T H.222.0 | ISO/IEC 13818-1:2021, “Information Technology—Generic coding of moving pictures and associated audio information - Part 1: Systems”, June 2021, https://www.itu.int/rec/T-REC-H.222.0-202106-I/en

[2] Recommendation ITU-R BT.709-6:2015, “Parameter values for the HDTV standards for production and international programme exchange”, July 2015, https://www.itu.int/rec/R-REC-BT.709-6-201506-I/en

[3] Recommendation ITU-R BT.2020:2015, “Parameter values for ultra-high definition television systems for production and international programme exchange”, Oct 2015, https://www.itu.int/rec/R-REC-BT.2020-2-201510-I/en

[5] Recommendation ITU-R BT.2100, “Image parameter values for high dynamic range television for use in production and international programme exchange”, July 2018, http://www.itu.int/rec/R-REC-BT.2100

[6] Report ITU-R BT.2390-2, “High dynamic range television for production and international programme exchange”, https://www.itu.int/pub/R-REP-BT.2390-2016 (companion report to ITU-R Recommendation BT.2100)

[8] Recommendation ITU-R BT.2408-7:2023. “Guidance for operational practices in HDR television production”, Sept 2023., https://www.itu.int/pub/R-REP-BT.2408-7-2023

[9] SMPTE ST 2084:2014, “High Dynamic Range Electro-Optical Transfer Function of Mastering Reference Displays”, Aug 2014, https://my.smpte.org/s/product-details?id=a1BVR0000008kaj2AA

[10] SMPTE ST ST 2086:2018, “Mastering Display Color Volume Metadata Supporting High Luminance and Wide Color Gamut Images”, April 2018, https://my.smpte.org/s/product-details?id=a1BVR0000008kal2AA

[24] ATSC: "Techniques for Establishing and Maintaining Audio Loudness for Digital Television," Doc. A/85, Advanced Television Systems Committee, Washington, D.C., 12 March 2013, Corrigendum No 1, "SPL", Feb 2021, https://www.atsc.org/atsc-documents/a85-techniques-for-establishing-and-maintaining-audio-loudness-for-digital-television/

[29] ETSI 102 366 v1.4.1 (2017-09), "Digital Audio Compression (AC-3, Enhanced AC-3) Standard", Sept 2009, https://www.etsi.org/deliver/etsi_ts/102300_102399/102366/01.04.01_60/ts_102366v010401p.pdf

[31] ANSI/CTA 861-I, "A DTV Profile for Uncompressed High Speed Digital Interfaces", Feb 2023, https://shop.cta.tech/products/a-dtv-profile-for-uncompressed-high-speed-digital-interfaces-ansi-cta-861-i?_ga=2.162090229.872823572.1725386735-468313473.1725386735&_gl=1%2A1s6163t%2A_gcl_au%2AMTAyODg0ODg0Ny4xNzI1Mzg2NzM1%2A_ga%2ANDY4MzEzNDczLjE3MjUzODY3MzU.%2A_ga_5P7N8TBME7%2AMTcyNTM4NjczNS4xLjEuMTcyNTM4Njc4Ni45LjAuMA..

[33] 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)", http://www.etsi.org/deliver/etsi_ts/103400_103499/10343301/01.02.01_60/ts_10343301v010201p.pdf

[34] ETSI TS 103 433-2 v1.1.1 (2018-01) "High-Performance Single Layer High Dynamic Range (HDR) System for use in Consumer Electronics devices;Part 2: Enhancements for Perceptual Quantization (PQ) transfer function based High Dynamic Range (HDR) Systems (SL-HDR2)", https://www.etsi.org/deliver/etsi_ts/103400_103499/10343302/01.01.01_60/ts_10343302v010101p.pdf

[35] ETSI TS 103 420 v1.2.1 (2018-10), "Object-based audio coding for Enhanced AC-3 (E-AC-3)".

https://www.etsi.org/deliver/etsi_ts/103400_103499/103420/01.02.01_60/ts_103420v010201p.pdf

[36] SMPTE ST 337:2015, “Format for Non-PCM Audio and Data in AES 3 Serial Digital Audio Interface”
https://my.smpte.org/s/product-details?id=a1BVR0000008kc22AA

[37] Recommendation ITU-R BS.1770-5, “Algorithms to measure audio programme loudness and true-peak audio level”, Nov 2023.
https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.1770-5-202311-I!!PDF-E.pdf

[43] SMPTE ST 2110-10:2022, “Professional Media over IP Networks: System Timing and Definitions”
https://pub.smpte.org/doc/st2110-10/20220328-pub/

[44] SMPTE ST 2110-20:2022, “Professional Media over IP Networks: Uncompressed Active Video”
https://pub.smpte.org/doc/st2110-20/20221214-pub/

[45] SMPTE ST 2110-21:2022, “Professional Media over IP Networks: Traffic Shaping and Delivery Timing for Video”
https://pub.smpte.org/doc/st2110-21/20221214-pub/

[46] SMPTE ST 2110-30:2017, “Professional Media over IP Networks: PCM Digital Audio”
https://pub.smpte.org/doc/st2110-30/

[47] SMPTE ST 2110-40:2023, “Professional Media over IP Networks: SMPTE ST 291-1 Ancillary Data”
https://pub.smpte.org/doc/st2110-40/20231231-pub

[48] SMPTE ST 2108-1:2018, “HDR/WCG Metadata Packing and Signaling in the Vertical Ancillary Data Space”
https://pub.smpte.org/doc/st2108-1/

[50] SMPTE ST 2065-1:2020, “Academy Color Encoding Specification (ACES)” https://pub.smpte.org/doc/st2065-1/20200909-pub/

[51] ATSC: A/300:2024-4, “ATSC 3.0 System”, April 2024, https://www.atsc.org/wp-content/uploads/2024/04/A300-2024-04-ATSC-3-System-Standard.pdf

[52] ATSC: A/322:2024, “Physical Layer Protocol”, April 3, 2024, https://www.atsc.org/wp-content/uploads/2024/04/A322-2024-04-Physical-Layer-Protocol.pdf

[53] ATSC: A/331:2023-02, “, “Signaling, Delivery, Synchronization, and Error Protection”, February 2023, https://prdatsc.wpenginepowered.com/wp-content/uploads/2023/02/A331-2023-02-Signaling-Delivery-Sync-FEC.pdf

[54] ATSC: A/341:2024-04, “Video-HEVC”, April 3, 2024, https://www.atsc.org/wp-content/uploads/2024/04/A341-2024-04-Video-HEVC.pdf

[55] ATSC: A/331:2024-04, “, “Signaling, Delivery, Synchronization, and Error Protection”, April 3, 2024, https://www.atsc.org/wp-content/uploads/2024/04/A331-2024-04-Signaling-Delivery-Sync-FEC.pdf

[56] ATSC: A/342-2:2024-0, “AC-4 System”, April 3, 2024, https://www.atsc.org/wp-content/uploads/2024/04/A342-2-2024-04-AC4-System.pdf

[57] ATSC: A/342-3:2024-4, “MPEG-H System”, April 3, 2024 https://www.atsc.org/wp-content/uploads/2024/04/A342-3-2024-04-MPEG-System.pdff

[61] EBU Tech 3364, “Audio Definition Model Metadata Specification Ver. 2.0”, June 2018, https://tech.ebu.ch/docs/tech/tech3364.pdf

[62] EBU R 128, “Loudness Normalisation and Permitted Maximum Level of Audio Signals”, June 2023, https://tech.ebu.ch/docs/r/r128.pdf

[63] ETSI TS 101 154 v2.4.1 (2018-02), “Digital Video Broadcasting (DVB); Specification for the use of Video and Audio Coding in Broadcasting”, February 14, 2018, https://www.etsi.org/deliver/etsi_ts/101100_101199/101154/02.04.01_60/ts_101154v020401p.pdf

[65] 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

[69] ISO/IEC: 23008-2:2023, “Information technology -- High efficiency coding and media delivery in heterogeneous environments -- Part 2: High efficiency video coding”, Oct 2023, https://www.iso.org/standard/85457.html[41]

[70] ISO/IEC: 23008-3, "Information technology – High efficiency coding and media delivery in heterogeneous environments – Part 3: 3D audio", Aug 2022, https://www.iso.org/standard/83525.html

[71] 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

[72] ITU-R BS.2076-2, “Audio Definition Model”, Oct 2019, https://www.itu.int/rec/R-REC-BS.2076-2-201910-I/en

[73] ITU-R BS.2088-1, “Long-form file format for the international exchange of audio programme materials with metadata”, October 2019, https://www.itu.int/rec/R-REC-BS.2088/en

[41] Also published by ITU as ITU-T Recommendation H.265:2024, https://www.itu.int/rec/T-REC-H.265-202407-P/en

[74] 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

[78] SCTE 242-3:2022, “Next Generation Audio Coding Constraints for Cable Systems: Part 3 –MPEG-H Audio Coding Constraints”, 2022, https://account.scte.org/standards/library/catalog/scte-242-3-next-generation-audio-coding-constraints-for-cable-systems-part3/

[82] SMPTE ST 2022-6:2012, “Transport of High Bit Rate Media Signals over IP Networks (HBRMT)”, October 9, 2012, https://pub.smpte.org/doc/st2022-6/20121009-pub/

[86] SMPTE ST 2094-10:2021, “Dynamic Metadata for Color Volume Transform – Application #1”, Dec 2, 1020, https://pub.smpte.org/doc/st2094-10/20201202-pub/

[87] TTA: TTAK-KO-07.0127R5:2020, “Transmission and Reception for Terrestrial UHDTV Broadcasting Service”, December 10, 2020, https://www.tta.or.kr/tta/ttaSearchView.do?key=77&rep=1&searchStandardNo=TTAK.KO-07.0127/R1&searchCate=TTAS

[90] Dolby Vision Profiles and Levels, v1.4, October, 2023, https://professionalsupport.dolby.com/s/article/What-is-Dolby-Vision-Profile?language=en_US

[91] ETSI TS 103 491 v1.2.1 (2019-05) DTS-UHD Audio Format: Delivery of Channels, Objects and Ambisonic Sound Fields, May 2019, https://www.etsi.org/deliver/etsi_ts/103400_103499/103491/01.02.01_60/ts_103491v010201p.pdf

[92] ETSI TS 101 154 v2.6.1 (2019-09), “Digital Video Broadcasting (DVB); Specification for the use of Video and Audio Coding in Broadcasting Application based on the MPEG-2 Transport Stream”, September 2019,

https://www.etsi.org/deliver/etsi_ts/101100_101199/101154/02.06.01_60/ts_101154v020601p.pdf

[93] ANSI/SCTE 242-4 2022, Next Generation Audio Coding Constraints for Cable Systems: Part 4 – DTS-UHD Audio Coding Constraints, https://account.scte.org/standards/library/catalog/scte-242-4-next-generation-audio-coding-constraints-for-cable-systems-part4/

[94] ANSI/SCTE 243-4 2022, Next Generation Audio Carriage Constraints for Cable Systems: Part 4 – DTS-UHD Audio Carriage Constraints, https://account.scte.org/standards/library/catalog/scte-243-4-next-generation-audio-carriage-for-cable-systems-part4/

[106] EBU R 103, “Video signal tolerance in Digital Television Systems”, Version 3.0, June 2020, https://tech.ebu.ch/docs/r/r103.pdf

[107] ITU-R BT.2124, “Objective metric for the assessment of the potential visibility of colour differences in television”, January 2019, https://www.itu.int/rec/R-REC-BT.2124-0-201901-I/en

[123] Recommendation ITU-R BS.2125 “A serial representation of the Audio Definition Model”, April 2020, https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.2125-1-202205-I!!PDF-E.pdf

[124] SMPTE ST 2116:2019, “Format for Non-PCM Audio and Data in AES3 — Carriage of Metadata of Serial ADM (Audio Definition Model)”, Jan 2020, https://pub.smpte.org/doc/st2116/20191018-pub/

[125] SMPTE RDD 33:2015, “Format for Non-PCM Audio and Data in AES3 — Dolby-E ® Data Type”, May 2015, https://pub.smpte.org/doc/rdd33/20150327-pub/

[126] SMPTE ST 2110-31:2022, ,SMPTE Standard – Professional Media Over Managed IP Networks: AES3 Transparent Transport, Nov 2022, https://pub.smpte.org/doc/st2110-31/20220624-pub/

[132] SMPTE ST 2110-22:2022, SMPTE Standard-Professional Media Over Managed IP Networks: Constant Bit-Rate Compressed Video, March 2022, https://pub.smpte.org/doc/st2110-22/20220331-pub/

[133] SMPTE ST 2059-2, SMPTE Standard-SMPTE Profile for the use of IEEE-1588 Precision Time Protocol in Professional Broadcast Applications, June 2021, https://pub.smpte.org/doc/st2059-2/20201209-pub/

[134] IEEE 1588-2019, IEEE Standard for a Precision Clock Synchronization Protocol for Networked Measurement and Control Systems, June 2020, https://standards.ieee.org/ieee/1588/6825/

[Y] **Yellow Book** – Beyond Foundational Technologies

[Y01] Section 7.1.2, Dolby VIsion Cross Compatibility

[Y02] Section 7.2, Next Generation Audio (NGA)

[B] **Blue Boo**k – Ultra HD Production and Post Production

[B01] Section 7.3, Signaling Transfer Function, System Colorimetry and Matrix Coefficients

[V] **Violet Book** – Real World Ultra HD

[V01] Section 11.5, Interoperability of Atmos Immersive Audio

[V02] Section 13, Real World Foundation Ultra HD Deployments

(End of Indigo Book)