Ultra HD Forum Guidelines

Yellow Book: Beyond Foundational Technologies

Ultra HD Forum

8377 Fremont Blvd., Suite 117

Fremont, CA 94538

UNITED STATES

VERSION 3.0.0

Apr 7, 2023

1. Foreword

This new version v3 of 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, characterized in large part by the video resolutions and the dynamic range offered for media in “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.

This work represents over eight years of collaborative effort. These new books would not have been possible without the leadership of Jim DeFilippis, who represents Fraunhofer and chairs 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.

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), Andrew Cotton (BBC), Jean Louis Diascorn (Harmonic), Richard Doherty (Dolby), Chris Johns (Sky UK), Katy Noland (BBC), Bill Redmann (InterDigital), Chris Seeger (Comcast/NBCUniversal), and Alessandro Travaglini (Fraunhofer).

This document, *Beyond Foundational Technologies* (Yellow Book), 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.

Nandhu Nandhakumar, President, Ultra HD Forum
April 2023

1.1 Changes from version 2.6 to 3.0

What’s new in the Spring 2023 version of the UHDF Guidelines Yellow Book, *Beyond Foundational Technologies* (v3.0), edited by Jean-Louis Diascorn.

The *Beyond Foundational Technologies* is the third of the series of Rainbow Books on the Guidelines for Ultra HD. The scope and purpose of this book is to describe the Ultra HD technologies that augment the Ultra HD Foundational technologies (Orange Book), including a discussion of HDR dynamic metadata (Dolby Vision, SL-HDR1 and 2), Next Generation Audio (NGA), High Frame Rate video (HFR), advanced encoding (AVS2 and AVC 3, CAE).
While most of the information in this edition is material from the previous version of the Guidelines (v2.6), the information has been updated. There is new information in the Annex on Dynamic Resolution Encoding, a novel technique that promises improved bandwidth use as well as QoE in OTT delivery of UHD content but has yet to be deployed widely. Each section includes references to more detailed information contained in the companion Rainbow Books.

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

Jim DeFilippis and Pete Sellar,

Guidelines Working Group Co-Chairs, Ultra HD Forum, April 2023

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. List of Figures

List Of Images

6. List of Tables

List Of Tables

7. Beyond Foundational Technologies

Beyond foundational technologies are the improvements over what is called foundation Ultra HD in the Orange book.

Ultra HD is in fact a vast range of improvements and here we present additional characteristics on top of the main improvements.

These additional characteristics include

- Dynamic metadata for HDR: providing even more precise luminance levels.
- Next Generation Audio, providing an even more immersive experience and precise location of objects in the audio space.
- High Frame: for a better portrayal of fast movement video with details such as sports.
- Encoding improvements.

This book also proposes recommendations on these technologies and also recommendations for HDR, colorimetry and conversions between HDR and SDR.

7.1. HDR w/Dynamic Metadata

For PQ HDR content[1], as described in Section 7.2.4 in Orange Book, HDR10 provides the static metadata elements in a PQ10-based HDR format as specified by SMPTE ST 2086 [10], MaxFALL, and MaxCLL. That section identifies a number of limitations with these particular HDR metadata values, notably the difficulty with setting these values in a live environment, real-world experience suggesting that these values have been set to artificial numbers to force certain looks on consumer displays, and the inability to correctly set these values given limitations of mastering displays.

[1] HLG10 does not specify any display metadata as it is based on normalized scene-light, rather than the absolute luminance of the signal seen on the mastering display, as described in Orange Book, section 7.2.2. As such, the headroom (measured in f-stops) for HLG highlights above HDR Reference White, is approximately constant regardless of the display's nominal peak luminance. Moreover the HLG10 display EOTF, which is fully specified by the ITU-R BT.2100 [5], includes a variable display gamma to provide adjustment for a specific display's peak brightness capabilities, along with eye adaptation; thereby allowing HLG to function in brighter viewing environments. Thus static or dynamic metadata is not required for HDR productions using HLG10.

In addition to these limitations, the values of MaxFALL and MaxCLL are also very limited in that they are only currently specified to provide single values for the entirety of the program. The dynamic range of both narrative and live content can vary dramatically from scene to scene. As a result the static, program-wide metadata values, as strictly defined, are of limited use for a great deal of content that does not have a static, unchanging dynamic range. Interoperability tests show that receivers can recognize changes in the static metadata within the duration of a program; however, it is yet unknown how frequently or quickly such changes can be recognized. For example, it is not expected that static metadata would change on a frame-by-frame basis.

Finally, there is no standardized way of utilizing these values in the final consumer display, so displays differ significantly in reproduction of the image. In practice some displays may ignore the values altogether. This is not consistent with the goal of displaying the image as close to the creative intent as possible on the target display.

A number of Dynamic Metadata methodologies have been developed to address the limitations of PQ10 and HDR10. Dynamic Metadata refers to metadata that describes the image at a much finer temporal granularity, scene-by-scene or even frame-by-frame and produces significantly more information about the mastering and creative intent of the scenes. In addition, most of these methodologies provide detailed information about tone mapping in the consumer display with the goal of consistent images across different manufacturers' displays. The methodologies also are designed to preserve creative intent, with the final displayed image being as close to the mastered image as the consumer display has the ability to reproduce.

Some of these methodologies go further by capturing the metadata during the color grading session and passing that metadata to consumer displays to better reproduce the creative intent. Most metadata schemes also provide for automatic metadata creation, which is useful in workflows for live content.

In general, several of these dynamic metadata schemes are additive, in that they provide additional information about the carried PQ10 image, and the HDR10 static metadata remains intact alongside the dynamic metadata. In some cases, this can provide a simple backwards compatibility to an HDR10-only display - the dynamic metadata is simply ignored.

Finally, many of these methodologies have considered how the signal can be backward compatible with SDR displays and have built-in methods for conversion. See Dolby Vision™ described in Section 7.1.2 and SL-HDR2 described in Section 7.1.3.

SL-HDR1 is another HDR dynamic metadata technology, which serves a different purpose than Dolby Vision or SL-HDR2. SL-HDR1 is intended to enable the service provider to emit an HDR/2020 service in an SDR/709 format that can be "reconstructed" to HDR/2020 by the receiver. HDR/2020 receivers that can interpret the SL-HDR1 metadata can present the HDR/2020 format to the viewer. The SDR/709 content can be displayed by receivers that cannot display HDR/2020. In this way SL-HDR1 provides a measure of backward compatibility for both HLG and PQ-based HDR content. It should be noted that SL-HDR1 requires 10-bit encoding, and so may not help address legacy SDR/709 receivers that are only capable of 8-bit decoding. See Section 7.3 of the Indigo Book.

7.1.1. 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 861.4 [104]
- Dynamic metadata: SMPTE 2094-10 [86] and CTA 861.4
- MaxFall/MaxCLL: CTA 861-I [31]

7.1.1.1. 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).

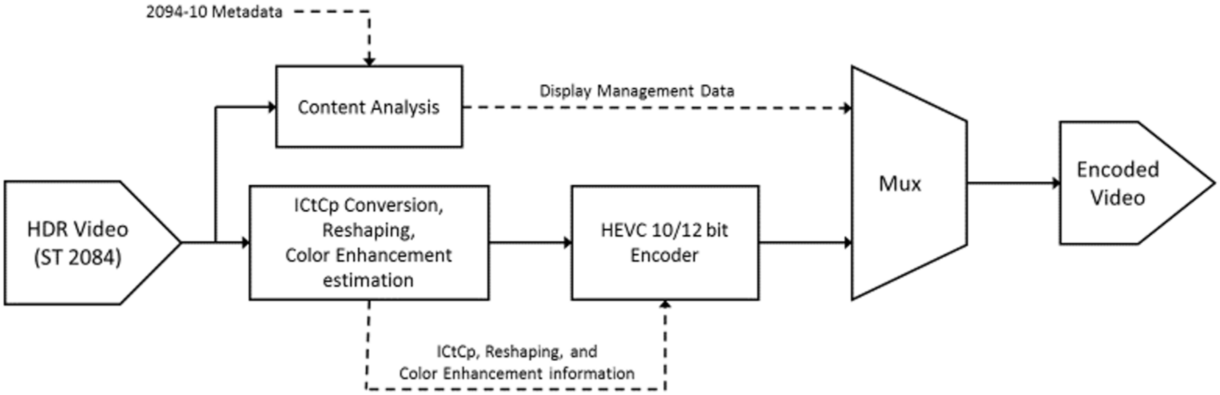


Figure 1. Encoder functional block diagram

If not natively in IC$_T$C$_P$ signal format, it may be advantageous to convert the HDR video into IC$_T$C$_P$ 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, un-modified HEVC decoder is used, thereby preserving the investment made by hardware vendors and owners.

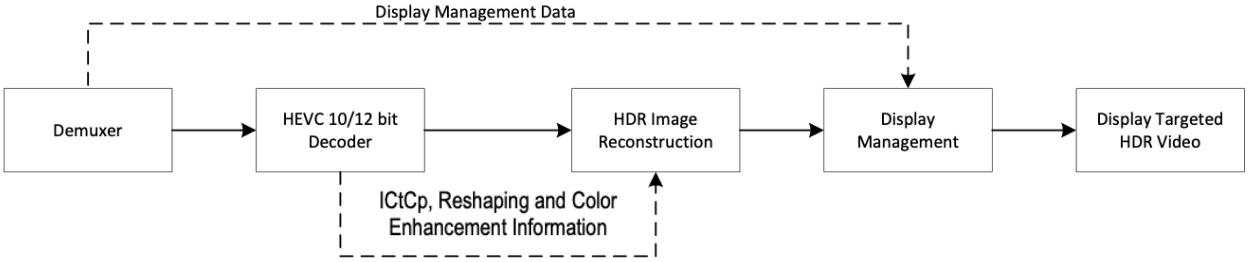


Figure 2. Decoder function block diagram

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.

7.1.2. Dolby Vision Cross Compatibility

Dolby Vision constrained as described in these Guidelines is based on SMPTE 2094-10 [86] metadata contained in SEI messages as described in Section 7.1.2.2. 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).

7.1.2.1. 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.

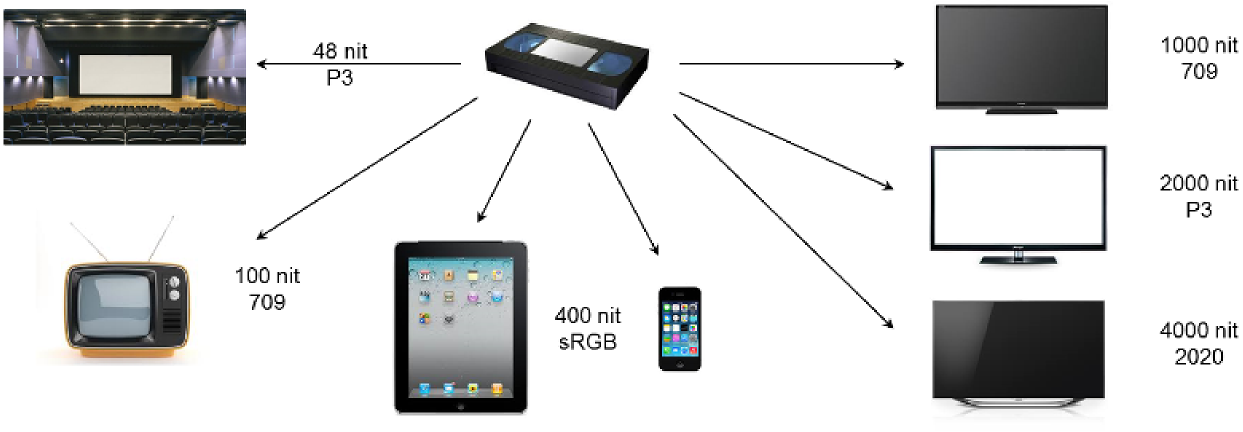


Figure 3. Example display device color volumes

While Dolby Vision works with the Y’C’$_B$C’$_R$ signal format model, in light of the limitations of Y’C’$_B$C’$_R$, especially at higher dynamic range, Dolby Vision also supports the use of IC$_T$C$_P$ signal format model as defined in BT.2100 [5]. IC$_T$C$_P$ isolates intensity from the color difference channels and may be a superior format in which to perform color volume mapping.

7.1.2.2. Dolby Vision in Broadcast

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

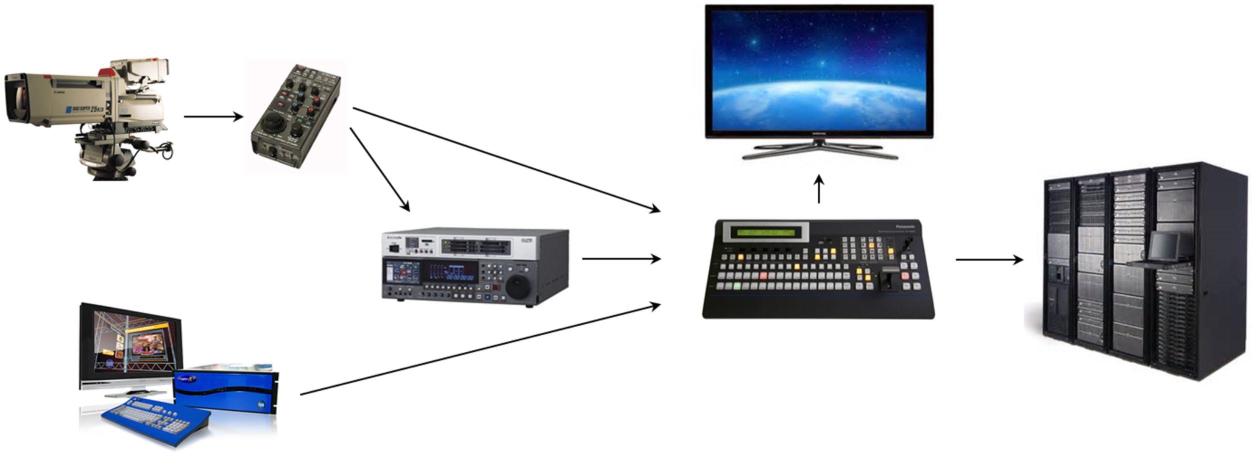


Figure 4. Example broadcast production facility components

Figure 5 shows a block diagram of the workflow in an HDR Broadcast facility using BT.2100 [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 SMPTE 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 5). In the case of generation at the encoder, the display management metadata can be inserted directly into the bitstream using standardized SEI messages by the HPU. Each payload of the display management metadata message is about 500 bits. It may be sent once per scene, per GOP, or per frame. 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 SMPTE ST 2094-10 metadata) and Dolby Vision devices that correctly utilize the included metadata.

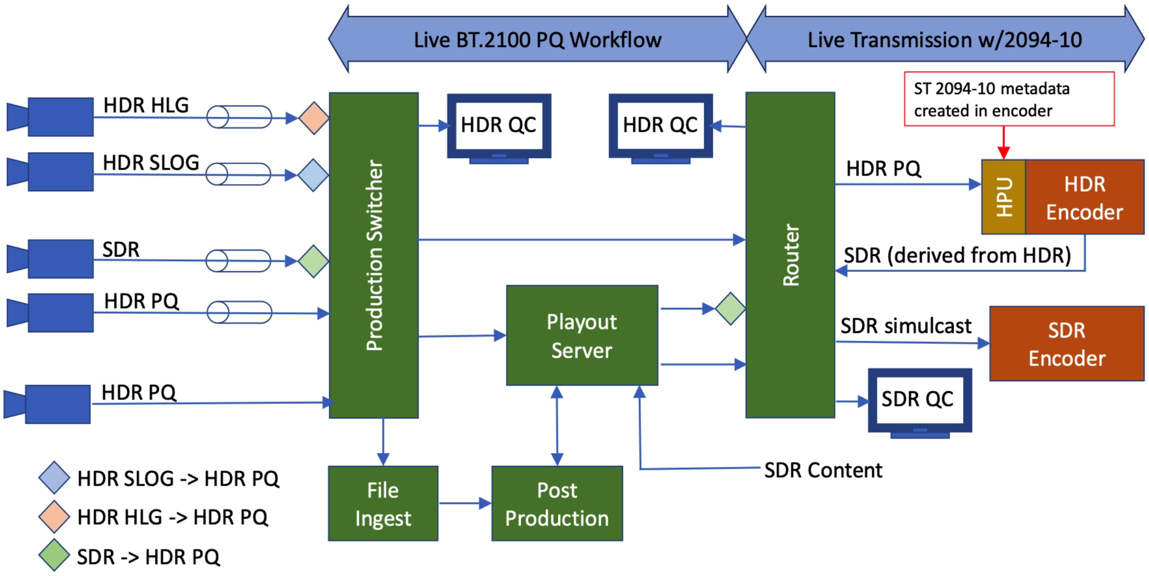


Figure 5. HDR broadcast production facility with BT.2100 PQ workflow- transition phase

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.

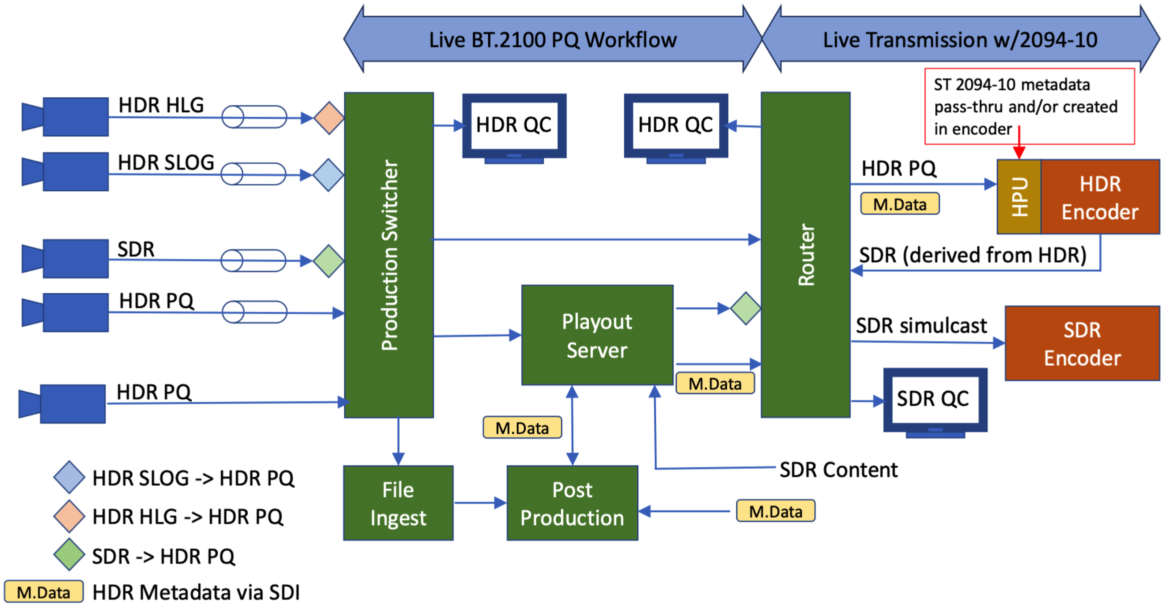


Figure 6. HDR broadcast production facility with BT.2100 PQ workflow- SDI metadata

7.1.3. Dual Layer (SL-HDR 1 and SL-HDR 2)

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 ATSC A/300 [51] section 5.1.6, wherein one or more program elements may be transported over a broadband path, as might be used for an EL). The EL may be delivered over a less resilient channel, since if lost, the image decoded from the BL is likely to remain available. The ability to tradeoff capacity and robustness is a significant feature of the physical layer protocols in ATSC 3.0, as discussed in Section 4.1 of 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 ETSI 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.

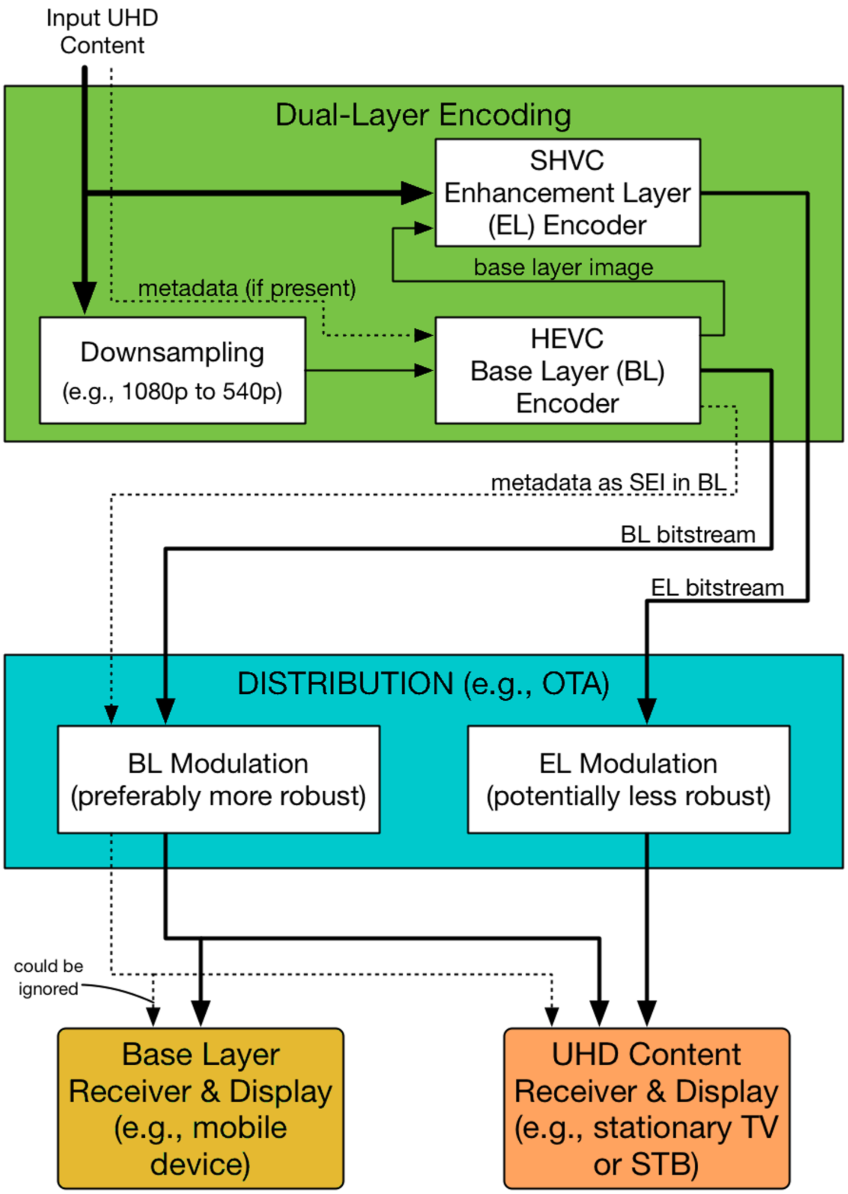

Input UHD Content
Dual-Layer Encoding
SHVC Enhancement Layer (EL) Encoder
base layer image
metadata (if present)
Downsampling (e.g., 1080p to 540p)
HEVC Base Layer (BL) Encoder
metadata as SEI in BL
BL bitstream
EL bitstream
DISTRIBUTION (e.g., OTA)
BL Modulation (preferably more robust)
EL Modulation (potentially less robust)
could be ignored
Base Layer Receiver & Display (e.g., mobile device)
UHD Content Receiver & Display (e.g., stationary TV or STB)

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 ETSI TS 103 433-1 SL-HDR1 [33]) 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. While SHVC is also supported by DASH, so that when connection bandwidth is limited, a DASH client may select only the BL, but as the connection bandwidth increases, the DASH client may additionally select the EL, so while not specifically noted herein, dual layer distribution is suitable for OTT distribution as well, both for VOD and linear programs.

7.1.3.1. SL-HDR1

As pointed out in Section 8.4 of 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.222.0 [1]) and may optionally deliver SDR/BT.2020 as the down-conversion target.

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

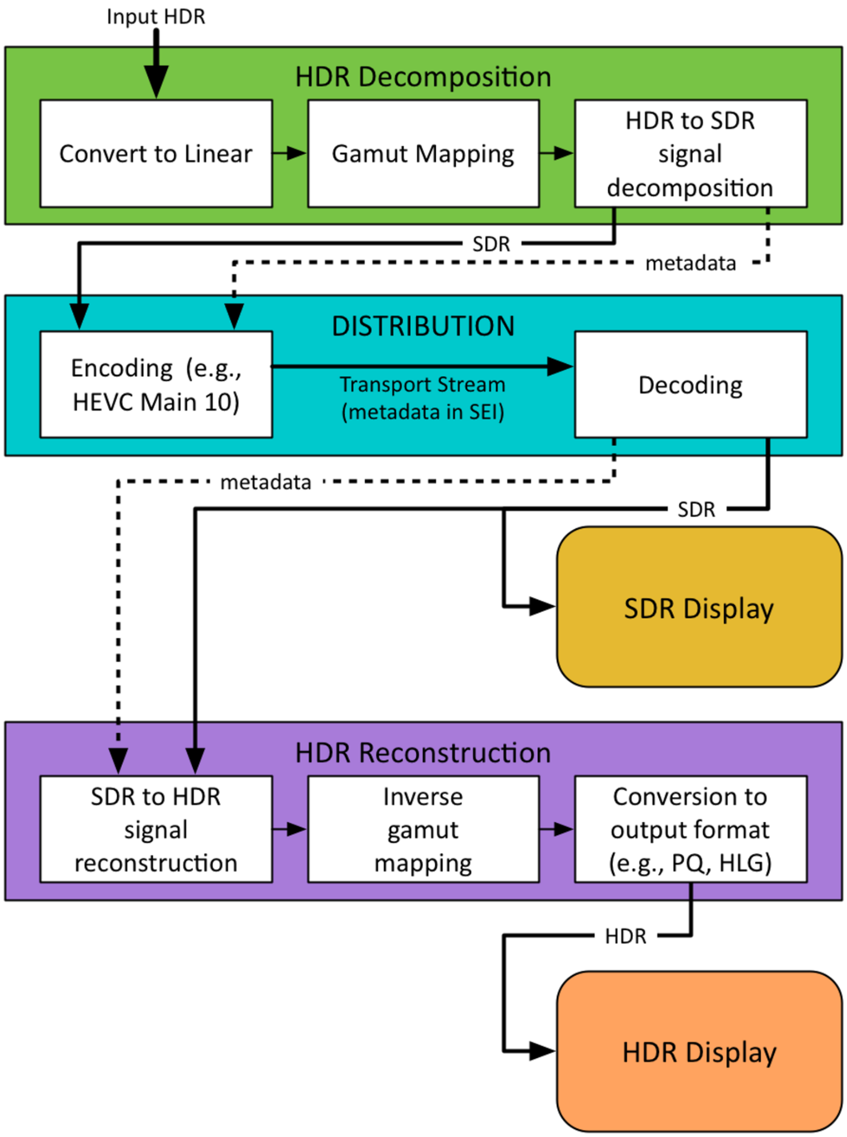

Input HDR
HDR Decomposition
Convert to Linear
Gamut Mapping
HDR to SDR signal decomposition
SDR
metadata
DISTRIBUTION
Encoding (e.g., HEVC Main 10)
Transport Stream (metadata in SEI)
Decoding
metadata
SDR
SDR Display
HDR Reconstruction
SDR to HDR signal reconstruction
Inverse gamut mapping
Conversion to output format (e.g., PQ, HLG)
HDR
HDR Display

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

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

For distribution, the metadata is embedded in the HEVC bitstream as SL-HDR information SEI messages, defined in ETSI TS 103 433-1 [33], 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.

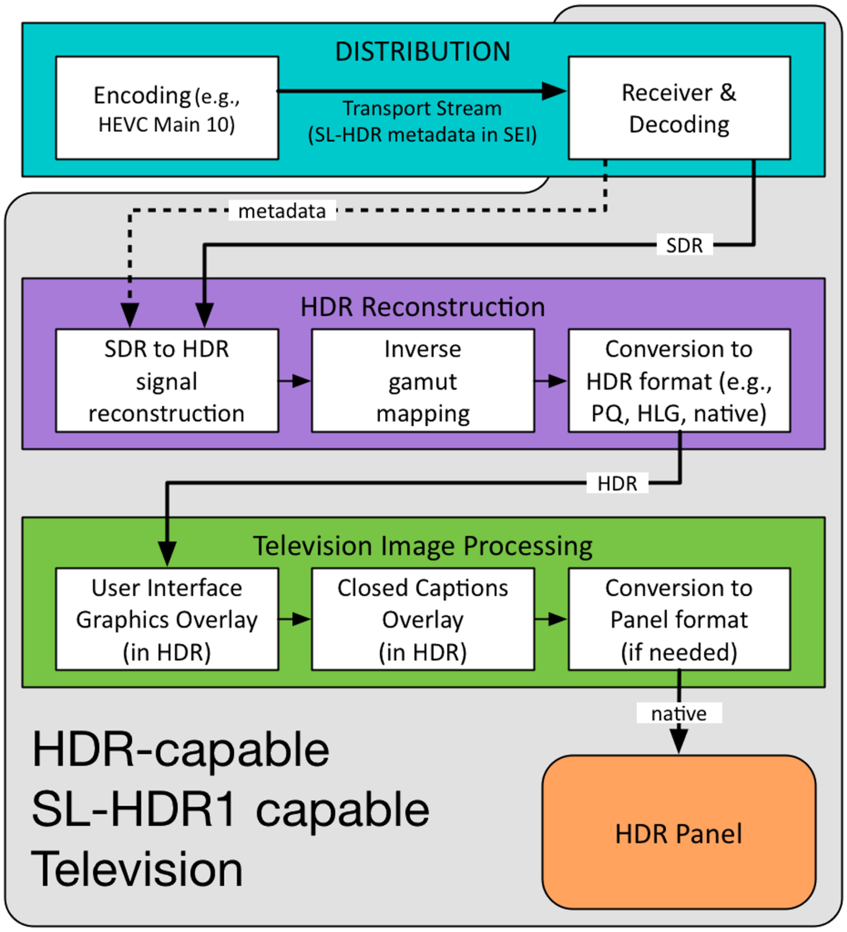

DISTRIBUTION
Encoding (e.g., HEVC Main 10)
Transport Stream (SL-HDR metadata in SEI)
Receiver & Decoding
metadata
SDR
HDR Reconstruction
SDR to HDR signal reconstruction
Inverse gamut mapping
Conversion to HDR format (e.g., PQ, HLG, native)
HDR
Television Image Processing
User Interface Graphics Overlay (in HDR)
Closed Captions Overlay (in HDR)
Conversion to Panel format (if needed)
native
HDR-capable SL-HDR1 capable Television
HDR Panel

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

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 9 if received by a decoder that recognizes the metadata and is connected to an HDR/WCG display, the metadata may be used by the decoder to reconstruct the HDR/WCG image, with the reconstruction taking place as shown by the HDR reconstruction block of Figure 8.

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

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

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

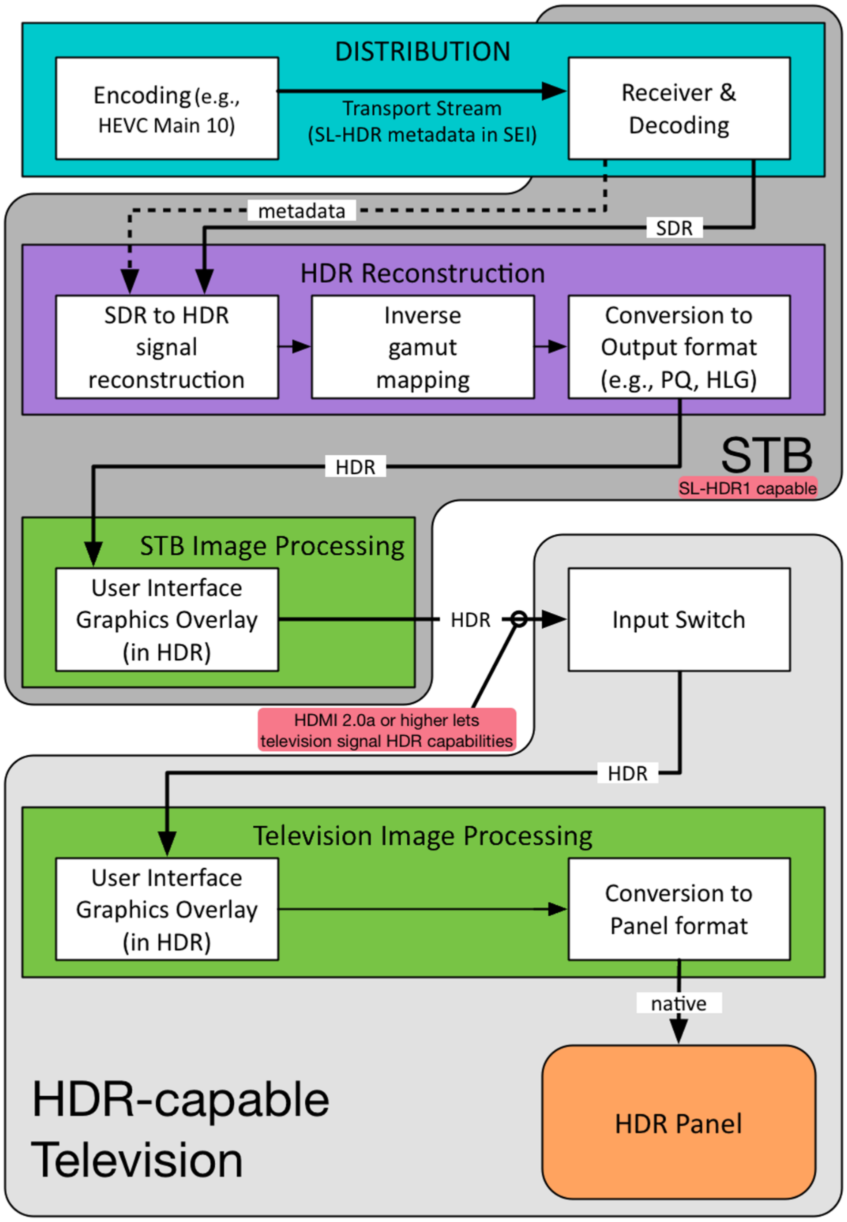

DISTRIBUTION
Encoding (e.g., HEVC Main 10)
Transport Stream (SL-HDR metadata in SEI)
Receiver & Decoding
metadata
SDR
HDR Reconstruction
SDR to HDR signal reconstruction
Inverse gamut mapping
Conversion to Output format (e.g., PQ, HLG)
HDR
STB
SL-HDR1 capable
STB Image Processing
User Interface Graphics Overlay (in HDR)
HDR
Input Switch
HDMI 2.0a or higher lets television signal HDR capabilities
HDR
Television Image Processing
User Interface Graphics Overlay (in HDR)
Conversion to Panel format
native
HDR Panel
HDR-capable Television

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

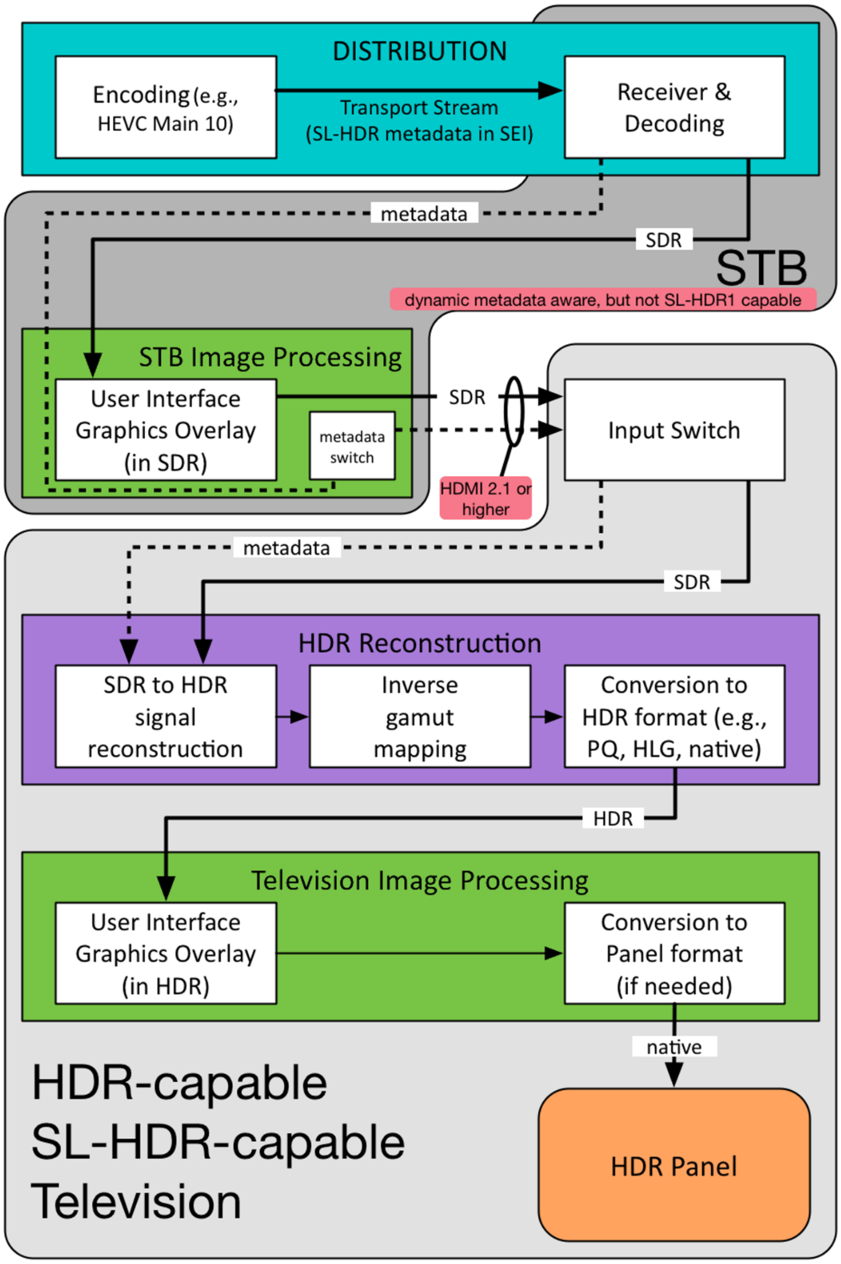

DISTRIBUTION
Encoding (e.g., HEVC Main 10)
Transport Stream (SL-HDR metadata in SEI)
Receiver & Decoding
metadata
SDR
STB
dynamic metadata aware, but not SL-HDR1 capable
STB Image Processing
User Interface Graphics Overlay (in SDR)
metadata switch
SDR
HDMI 2.1 or higher
Input Switch
metadata
SDR
HDR Reconstruction
SDR to HDR signal reconstruction
Inverse gamut mapping
Conversion to HDR format (e.g., PQ, HLG, native)
HDR
Television Image Processing
User Interface Graphics Overlay (in HDR)
Conversion to Panel format (if needed)
native
HDR Panel
HDR-capable SL-HDR-capable Television

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 11, 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 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 12.

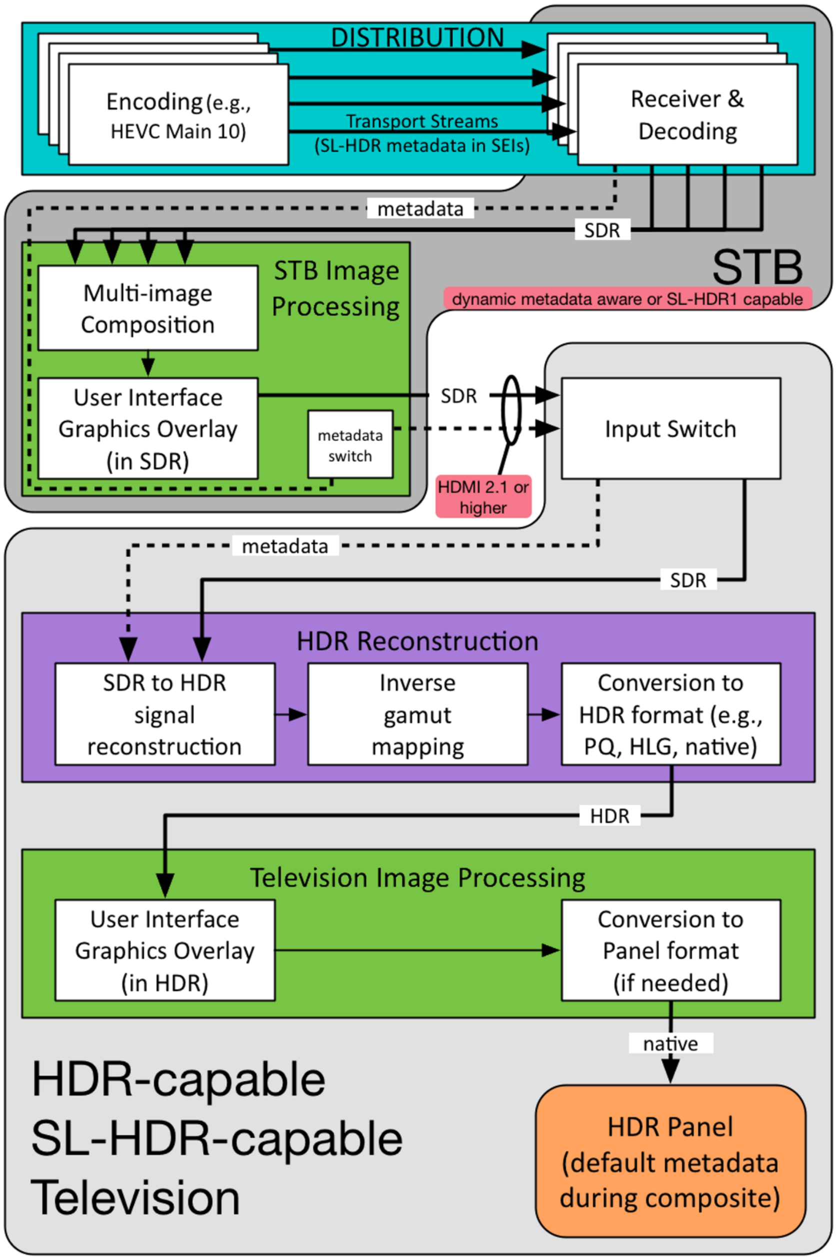

DISTRIBUTION
Encoding (e.g., HEVC Main 10)
Transport Streams (SL-HDR metadata in SEIs)
Receiver & Decoding
metadata
SDR
STB
dynamic metadata aware or SL-HDR1 capable
Multi-image Composition
STB Image Processing
User Interface Graphics Overlay (in SDR)
metadata switch
SDR
HDMI 2.1 or higher
Input Switch
metadata
SDR
HDR Reconstruction
SDR to HDR signal reconstruction
Inverse gamut mapping
Conversion to HDR format (e.g., PQ, HLG, native)
HDR
Television Image Processing
User Interface Graphics Overlay (in HDR)
Conversion to Panel format (if needed)
native
HDR-capable SL-HDR-capable Television
HDR Panel (default metadata during composite)

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 ETSI TS 103 433-1 [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 SMPTE ST 2094-10 [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.

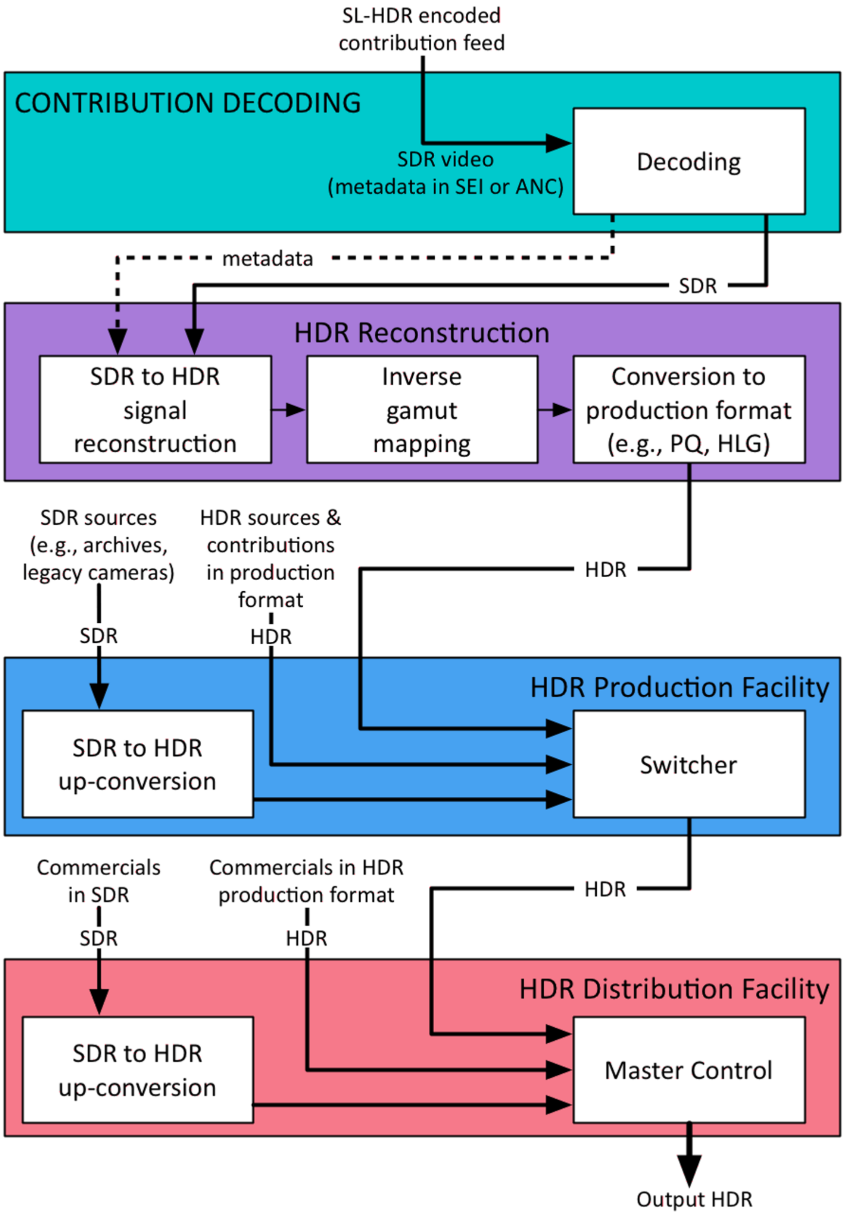


Figure 13. SL-HDR as a contribution feed to an HDR facility

The workflow for an HDR-ready affiliate receiving an SDR video with SL-HDR metadata as a contribution feed is shown in Figure 14. The decoding block and the HDR reconstruction block resemble the like-named blocks in Figure 8,with one potential exception: In Figure 14, the inverse gamut mapping block should use the invertible gamut mapping described in Annex D of ETSI 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 14, 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.

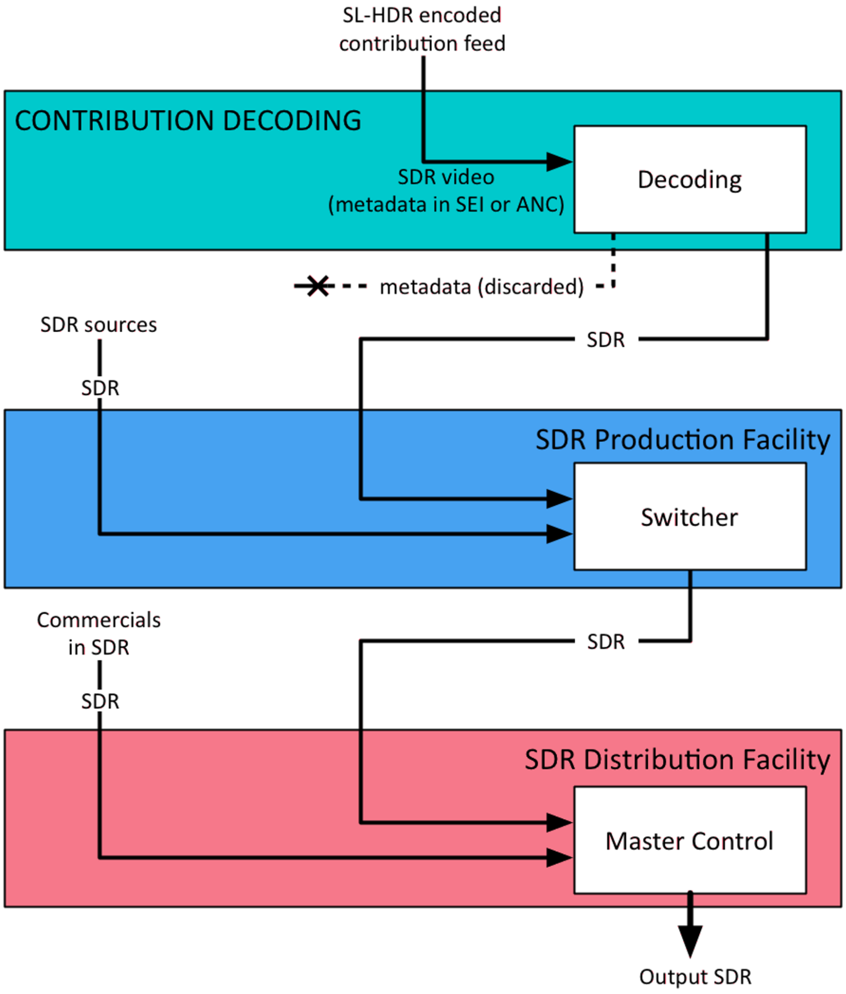


Figure 14. SL-HDR as a contribution feed to an SDR facility

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.

7.1.3.2. 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 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 [33] (in Annex A.2), 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.

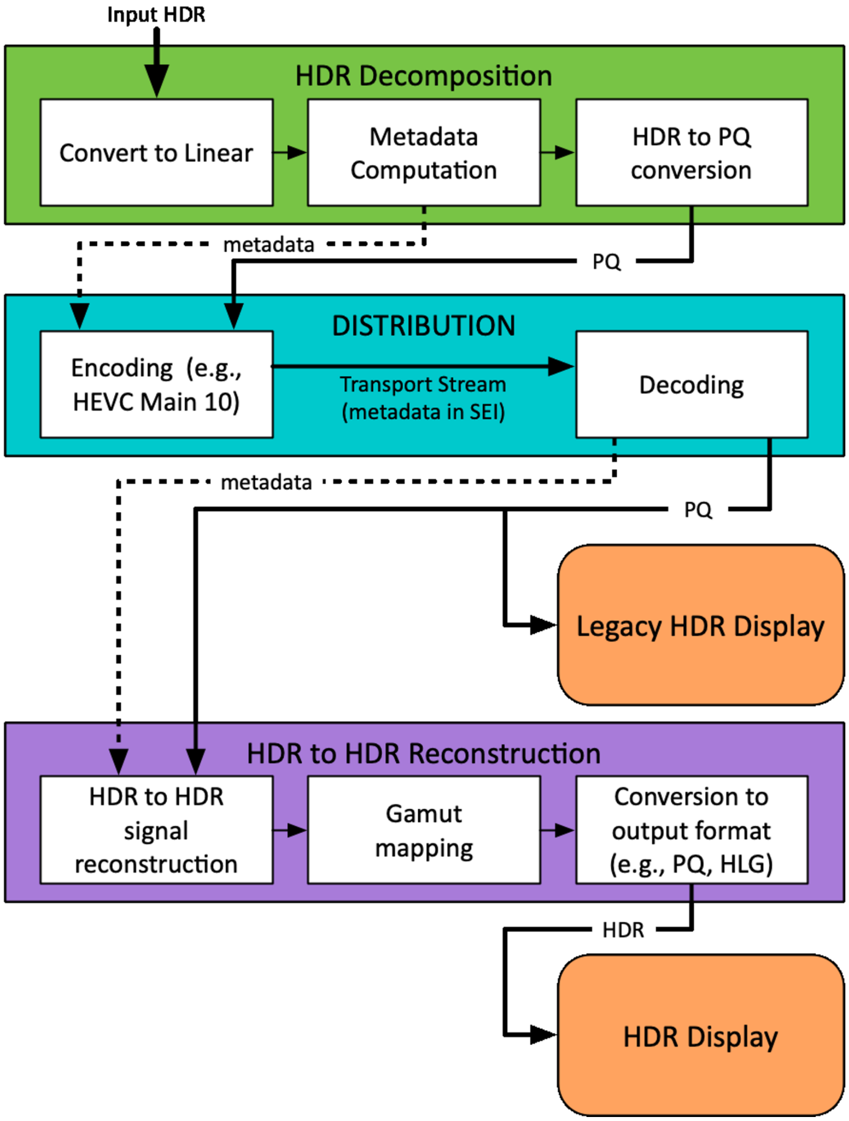


Figure 15. SL-HDR2 processing, distribution, reconstruction, for HDR presentation

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.

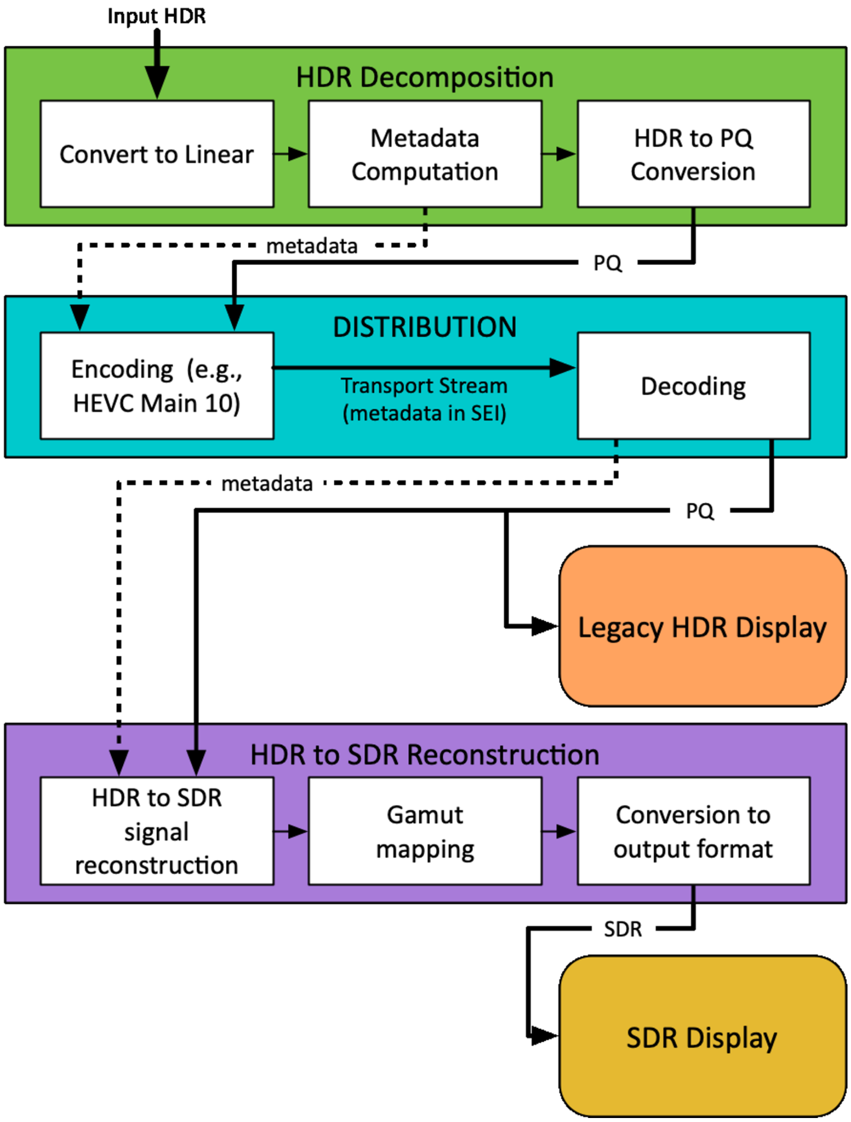


Figure 16. SL-HDR2 processing, distribution, reconstruction, and SDR presentation

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.

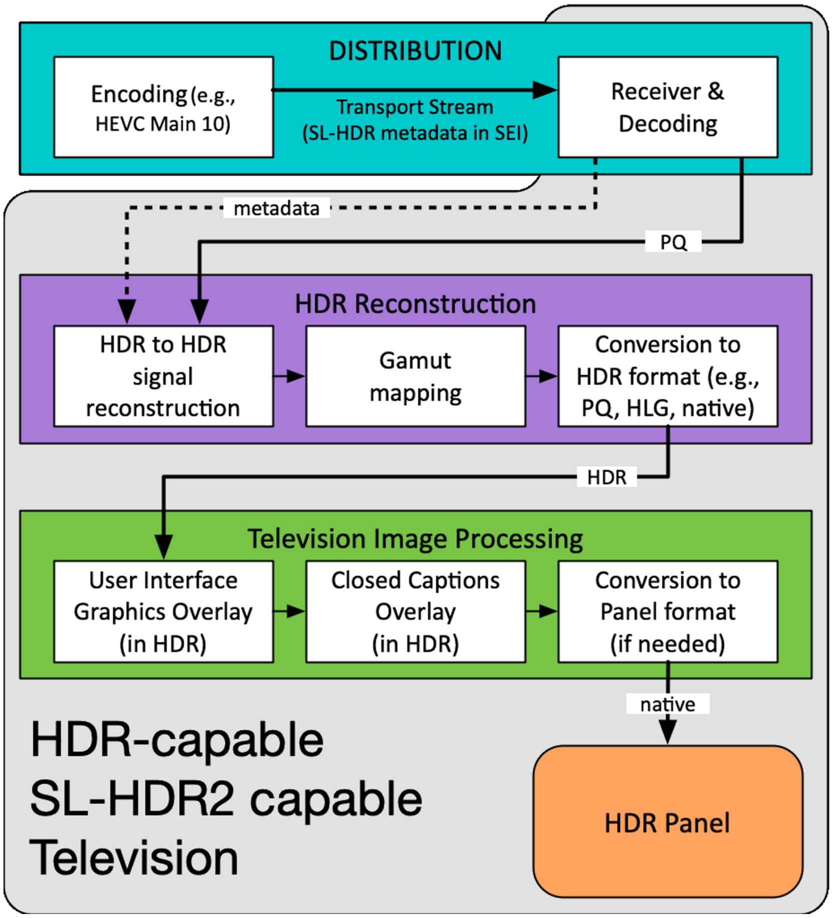

DISTRIBUTION
Encoding (e.g., HEVC Main 10)
Transport Stream (SL-HDR metadata in SEI)
Receiver & Decoding
metadata
PQ
HDR Reconstruction
HDR to HDR signal reconstruction
Gamut mapping
Conversion to HDR format (e.g., PQ, HLG, native)
HDR
Television Image Processing
User Interface Graphics Overlay (in HDR)
Closed Captions Overlay (in HDR)
Conversion to Panel format (if needed)
native
HDR-capable SL-HDR2 capable Television
HDR Panel

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.

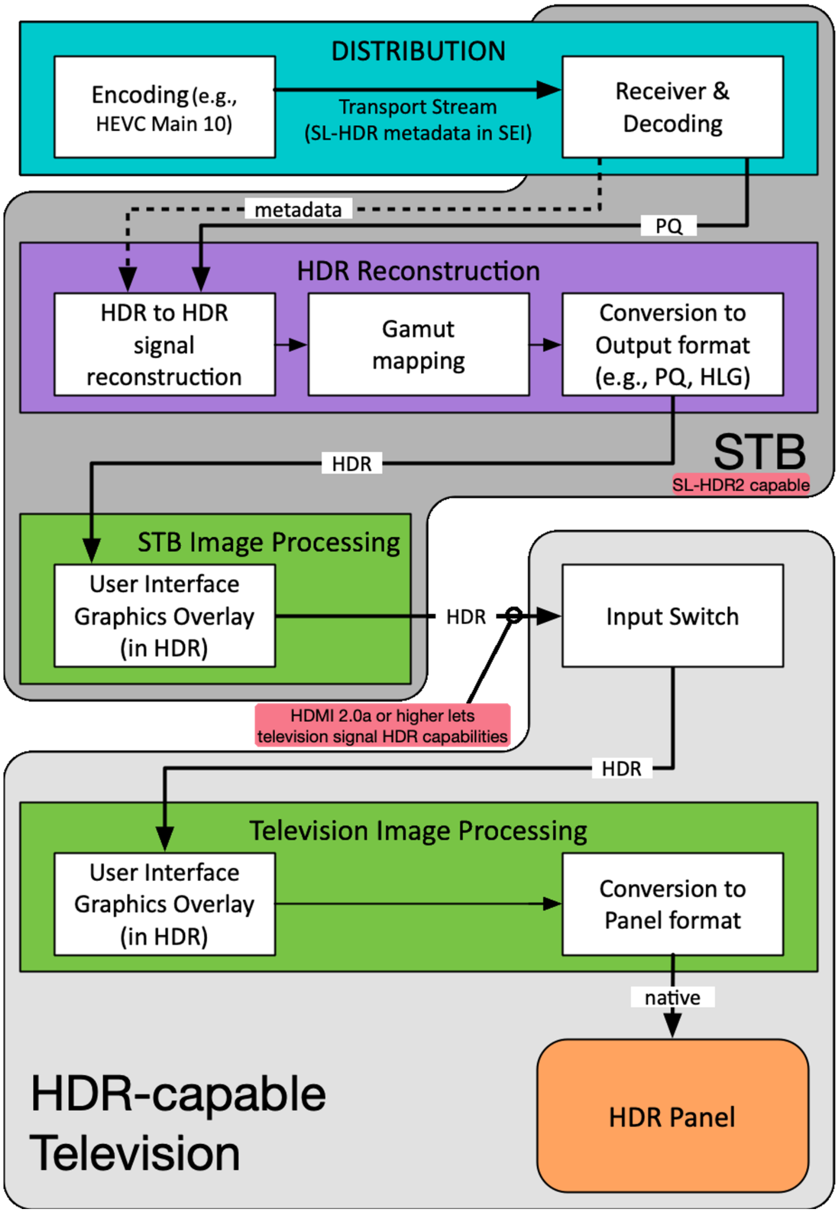

DISTRIBUTION
Encoding (e.g., HEVC Main 10)
Transport Stream (SL-HDR metadata in SEI)
Receiver & Decoding
metadata
PQ
HDR Reconstruction
HDR to HDR signal reconstruction
Gamut mapping
Conversion to Output format (e.g., PQ, HLG)
HDR
STB
SL-HDR2 capable
STB Image Processing
User Interface Graphics Overlay (in HDR)
HDR
Input Switch
HDMI 2.0a or higher lets television signal HDR capabilities
HDR
Television Image Processing
User Interface Graphics Overlay (in HDR)
Conversion to Panel format
native
HDR Panel
HDR-capable Television

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.

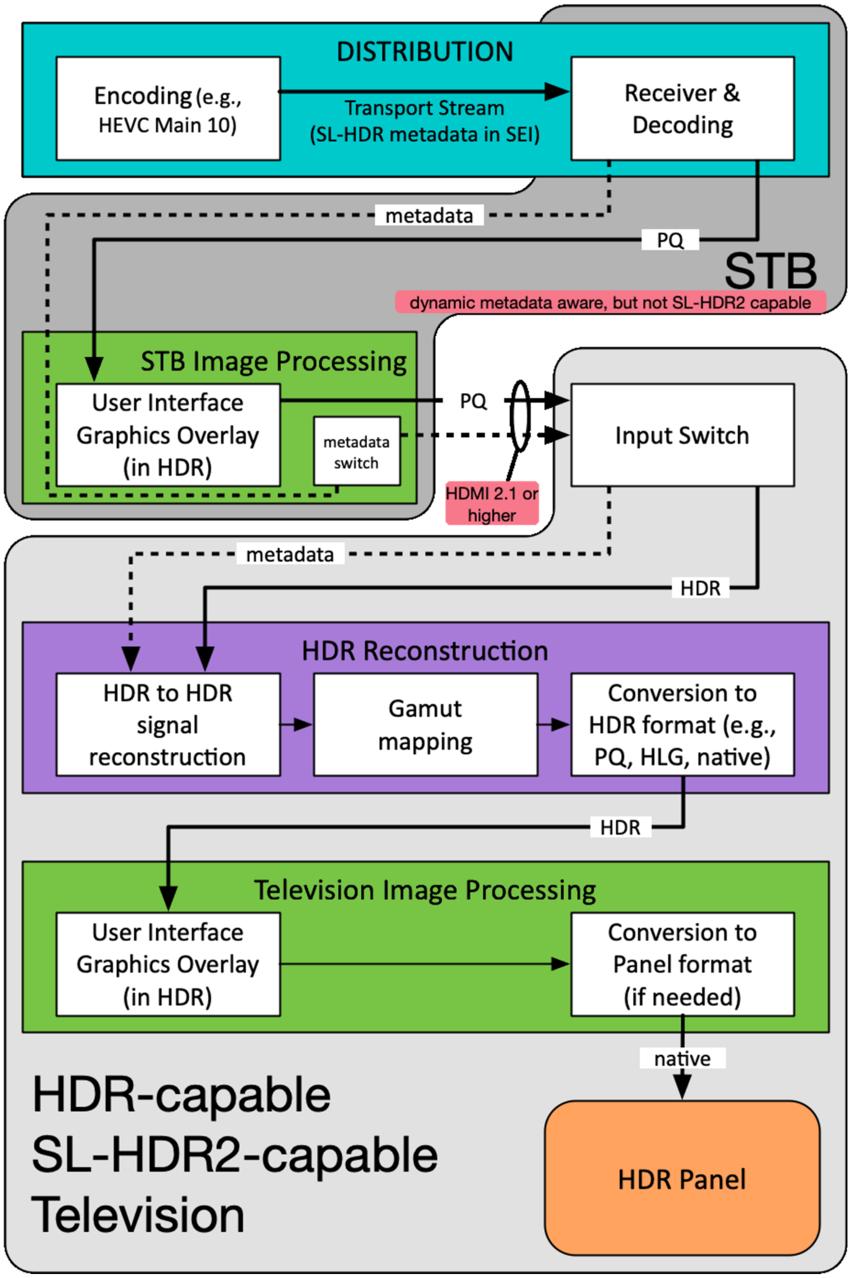

DISTRIBUTION
Encoding (e.g., HEVC Main 10)
Transport Stream (SL-HDR metadata in SEI)
Receiver & Decoding
metadata
PQ
STB
dynamic metadata aware, but not SL-HDR2 capable
STB Image Processing
User Interface Graphics Overlay (in HDR)
metadata switch
PQ
HDMI 2.1 or higher
Input Switch
metadata
HDR
HDR Reconstruction
HDR to HDR signal reconstruction
Gamut mapping
Conversion to HDR format (e.g., PQ, HLG, native)
HDR
Television Image Processing
User Interface Graphics Overlay (in HDR)
Conversion to Panel format (if needed)
native
HDR-capable SL-HDR2-capable Television
HDR Panel

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.

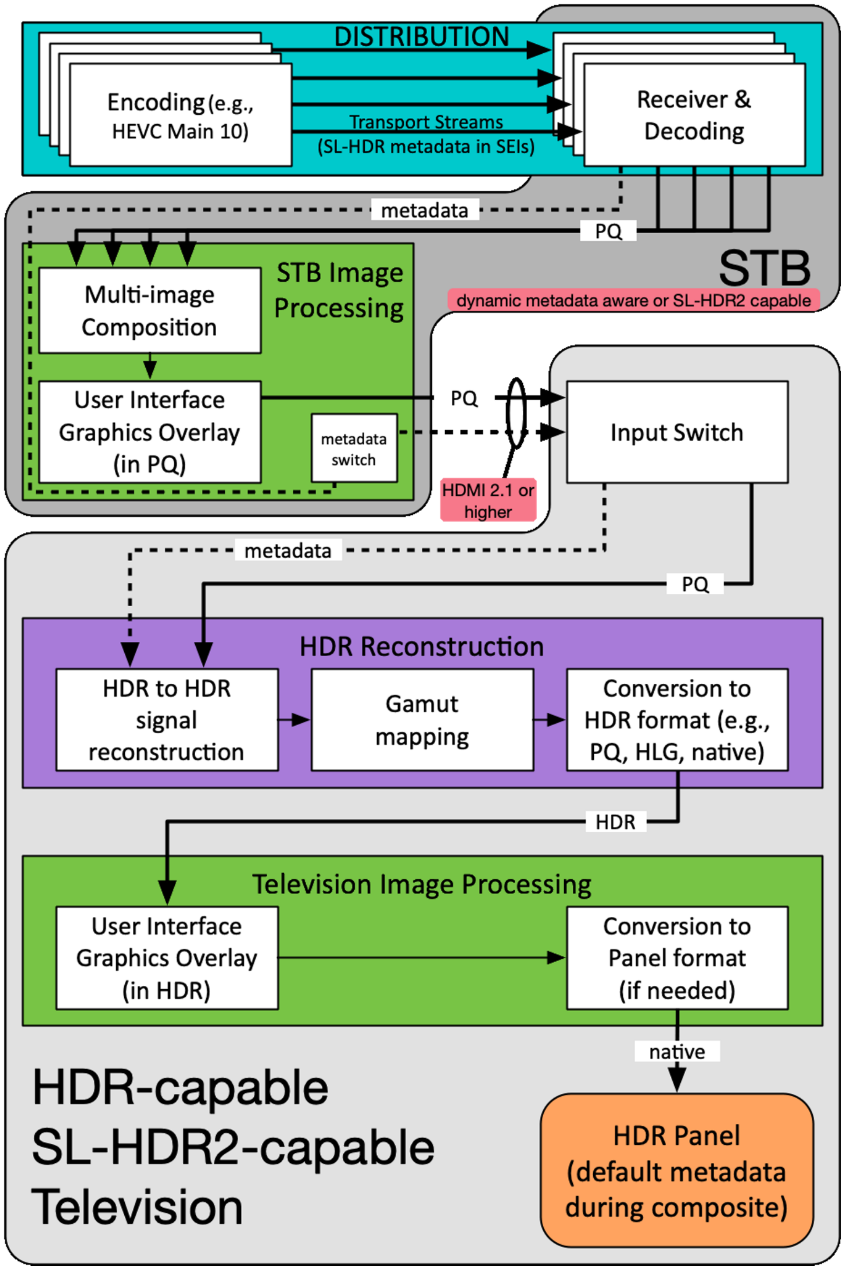

DISTRIBUTION
Encoding (e.g., HEVC Main 10)
Transport Streams (SL-HDR metadata in SEIs)
Receiver & Decoding
metadata
PQ
STB
dynamic metadata aware or SL-HDR2 capable
STB Image Processing
Multi-image Composition
User Interface Graphics Overlay (in PQ)
metadata switch
PQ
HDMI 2.1 or higher
Input Switch
metadata
PQ
HDR Reconstruction
HDR to HDR signal reconstruction
Gamut mapping
Conversion to HDR format (e.g., PQ, HLG, native)
HDR
Television Image Processing
User Interface Graphics Overlay (in HDR)
Conversion to Panel format (if needed)
native
HDR Panel (default metadata during composite)
HDR-capable SL-HDR2-capable Television

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 pre-processor 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 2108-1 [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.

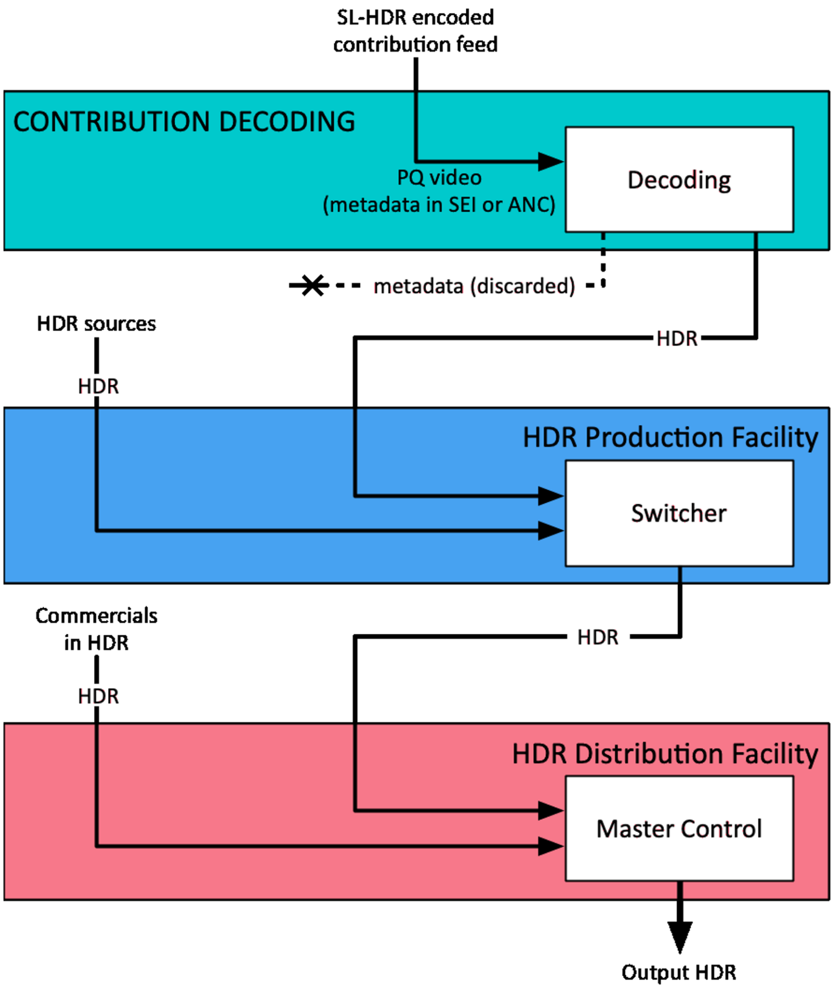


Figure 21. SL-HDR as a contribution feed to an HDR facility

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 ETSI 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 2094-10 [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.

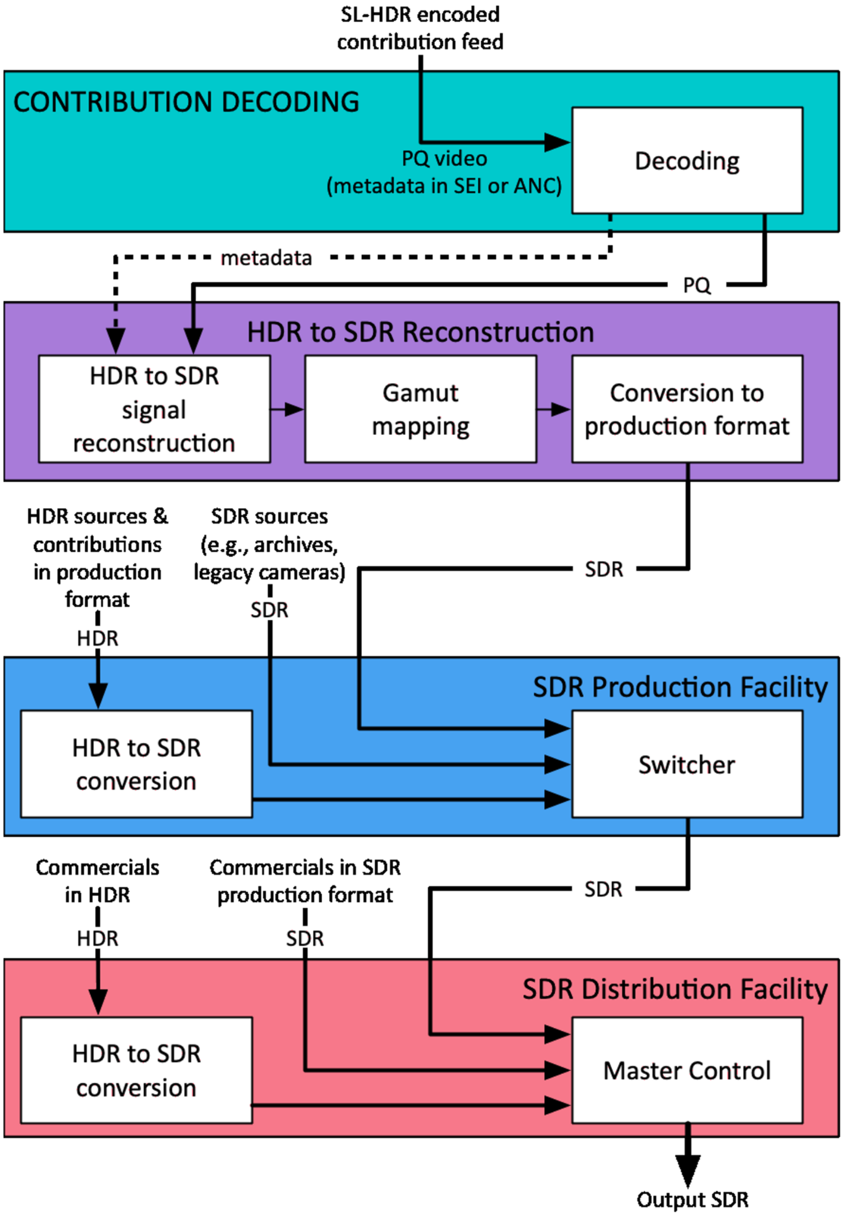


Figure 22. SL-HDR as a contribution feed to an SDR facility

7.2. Next Generation Audio (NGA)

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

- Immersive – An audio system that enables high spatial resolution in sound source localization in azimuth, elevation and distance, and provides an increased sense of sound envelopment
- Personalized – Enabling consumers to tailor and interact with their listening experience, e.g. selecting alternative audio experiences, switching between languages, enhancing dialogue intelligibility.
- Consistent – Playback experience automatically optimized for each consumer device, e.g. home and mobile
- Object-based Audio – Audio elements are programmed to provide sound from specific locations in space, irrespective of speaker location. By delivering audio as individual elements, or objects, content creators can simplify operations, reduce bandwidth, and provide a premium experience for every audience
- Scene-Based Audio – An arbitrarily large number of directional audio elements composing a 3D sound field are mixed in a fixed number of PCM signals according to the Higher-Order Ambisonics format. Once in the HOA format, the Audio Scene can be efficiently transmitted, manipulated, and rendered on loudspeaker layouts/headphones/soundbars.
- Flexible Delivery - NGA can be delivered to consumers over a number of different distribution platforms including terrestrial, cable, and satellite broadcast, IPTV, OTT, and mobile. It could also be delivered over a hybrid of broadcast and OTT
- Flexible Rendering - NGA can be experienced by consumers through headphones or speakers (e.g., TV speakers, home theater systems including ceiling speakers, sound bars) as shown in Figure 23.

There are three Next Generation Codecs described in the Guidelines, namely AC-4, DTS-UHD and MPEG-H. A quick overview of each system is described below.
See Indigo Book Section 8.1 (AC-4), Section 8.2 (DTS UHD) and Section 8.3 (MPEG-H)

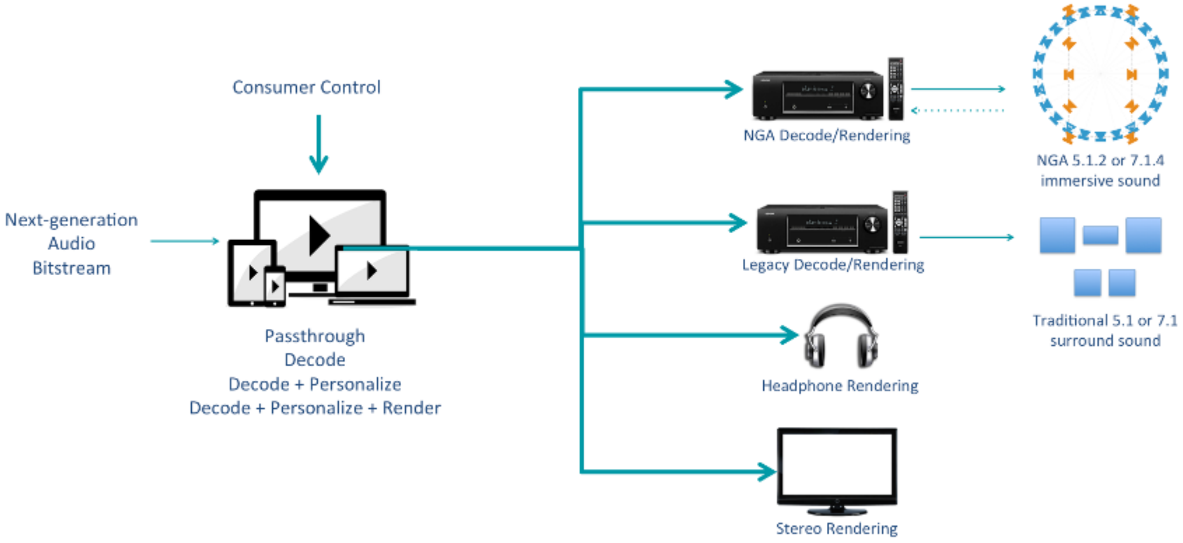

Consumer Control
Next-generation Audio Bitstream
Passthrough
Decode
Decode + Personalize
Decode + Personalize + Render
NGA Decode/Rendering
NGA 5.1.2 or 7.1.4 immersive sound
Legacy Decode/Rendering
Traditional 5.1 or 7.1 surround sound
Headphone Rendering
Stereo Rendering

Figure 23. NGA in the consumer domain

7.2.1. Audio Program Components and Preselections

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

NGA systems enable user control of certain aspects of the Audio Scene (e.g., adjusting the relative level of dialogue with respect to the ambient music and effects) by combining the Audio Program Components, present in one or more NGA streams, at the receiver side in user-selectable modes. In this way several Audio Program Components can be shared between different Audio Preselections, allowing more efficient delivery of additional services compared to legacy broadcast systems. For example, the same music & effects component can be used with a Spanish and an English dialog component, whereas a legacy broadcast would need to send two complete mixes, both including music and effects. This is a major advantage of NGA systems, where one stream contains more than one complete audio main, or multiple streams contain pieces of a complete audio main.

7.2.2. Carriage of NGA

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

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

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

7.2.3. Metadata

NGA codec systems have a rich set of audio metadata features and functions. Each codec has its own set of definitions; however, there is a common framework for audio metadata developed by the EBU called the Audio Definition Model (ADM) [61].

In general, there are three types of audio metadata:

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

7.2.4. Overview of Immersive Program Metadata and Rendering

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

An important consideration for a spatial (immersive) audio description model supporting audio objects is the choice of the spatial frame of reference. This will be utilized by the core Object-based Audio rendering algorithm (in playback devices) to map the source audio objects to the active speaker configuration/layout based on the positional metadata generated upstream in production.

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

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

Allocentric object position is therefore defined as an abstracted (unit) room where each object(s) 3-dimensional coordinates, (x,y,z) in [-1,1] × [-1,1] × [-1,1], correspond to the traditional balance controls found in mixing consoles (left/right, front/back and by extension to 3D bottom/top). Egocentric object positioning location of a point is specified by polar coordinates - azimuth (θ) elevation (φ) and radius (r) relative to the listeners position. Conversion between allocentric based object audio metadata and egocentric based object audio metadata is a lossless 3D geometric mathematical process carried out within rendering systems.

7.2.5. Audio Element Formats

The information contained in this section is provided courtesy of the Advanced Television Systems Committee from Standard A/342 Part 1, Audio Common Elements [55].

The NGA systems support three fundamental Audio Element Formats:

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

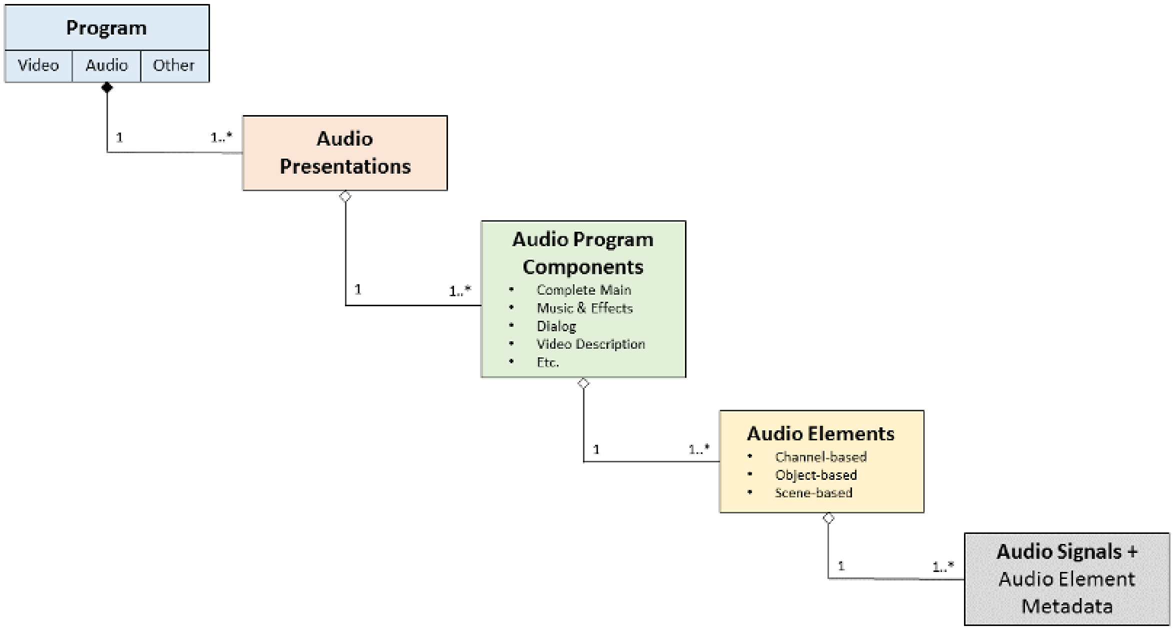


Figure 23. Relationship of key audio terms

Table 1. Mapping of Terminology Between NGA Technologies

Common Term	DASH-IF Term [59]	AC-4 Term [56]	MPEG-H Audio Term [57]	DTS-UHD Term[91]
Audio Element Metadata		Metadata, Object Audio Metadata	Metadata Audio Elements (MAE), Object Metadata (OAM)	Metadata Chunk
Audio Presentation	Preselection	Presentation	Preset	Presentation
Audio Program	Bundle	Audio Program	Audio Scene	Audio Program
Audio Program Component	Referred to as Audio Element	Audio Program Component	Group	Presentation/Object
Elementary Stream	Representation in an Adaptation Set	Elementary Stream	Elementary Stream	Elementary Stream

7.2.6. Audio Rendering

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

7.2.6.1. Video Description Service (VDS)[2]

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

7.2.6.2. Multi-Language

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

7.2.6.3. Personalized Audio

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

7.2.7. Dolby AC-4 Audio

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

[2] In the US, this service is now referred to as “Audio Description” (AD). FCC Report and Order 20-155 October 27, 2020

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 24, 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.

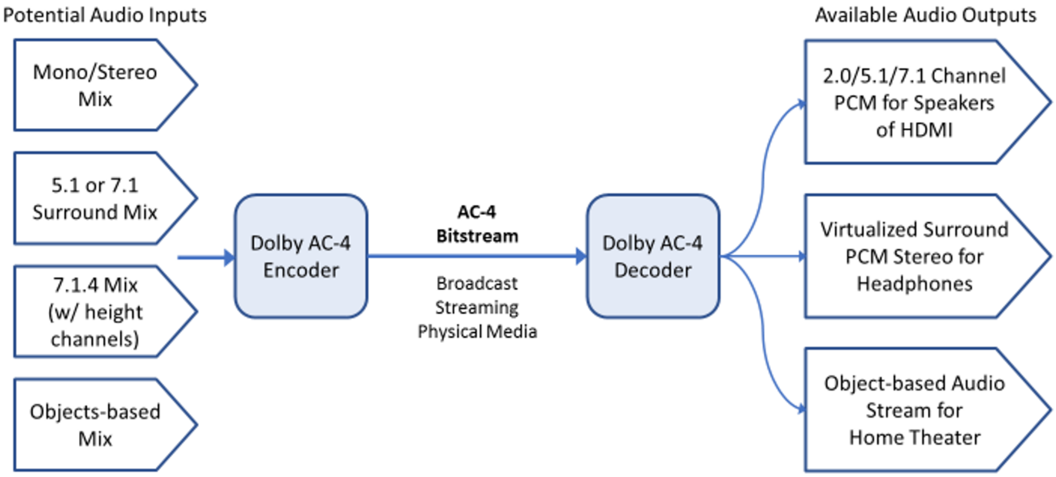


Figure 24. AC-4 Audio system chain

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 29 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 2. DE modes and metadata bitrates

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

7.2.8. DTS-UHD Audio

7.2.8.1. 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 ETSI TS 103 491 [91], and is included in the DVB Specification ETSI TS 101 154 [92] as well as also supported by the Society of Cable Television Engineers in SCTE 242-4 [93] and 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.

7.2.8.2. 7.2.8.2. System Overview

DTS-UHD audio format provides a number of significant improvements on legacy audio technologies, allowing enhanced audio controls for both personalization and enhanced accessibility. It is able to encode a number of different sources and deliver and render to multiple listening environments.

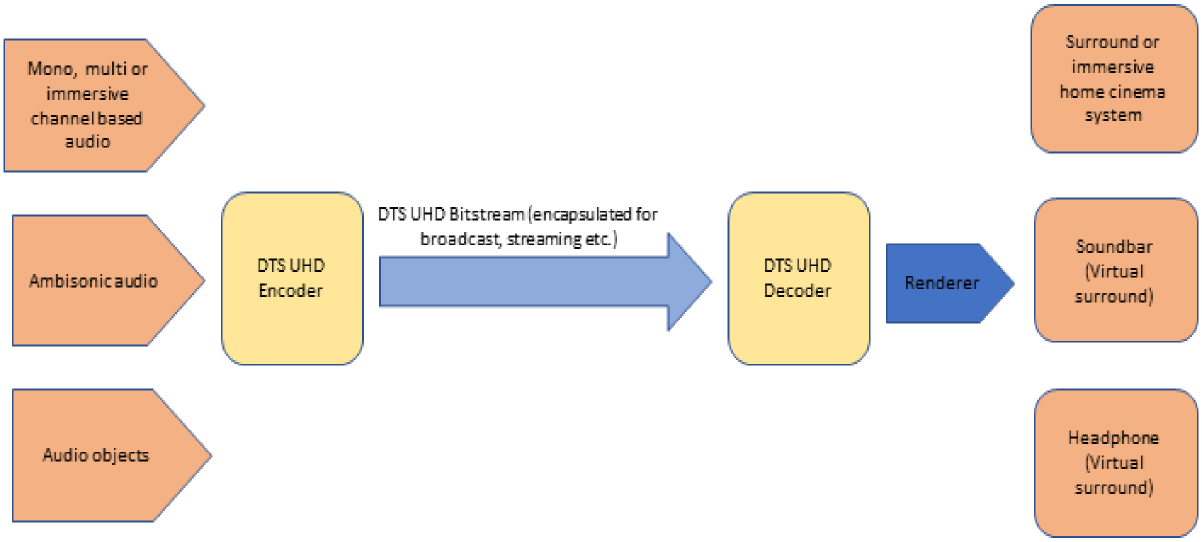

Mono, multi or immersive channel based audio
Ambisonic audio
Audio objects
DTS UHD Encoder
DTS UHD Bitstream (encapsulated for broadcast, streaming etc.)
DTS UHD Decoder
Renderer
Surround or immersive home cinema system
Soundbar (Virtual surround)
Headphone (Virtual surround)

Figure 25. 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 aspects 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.

7.2.9. MPEG-H Audio

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 26 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.

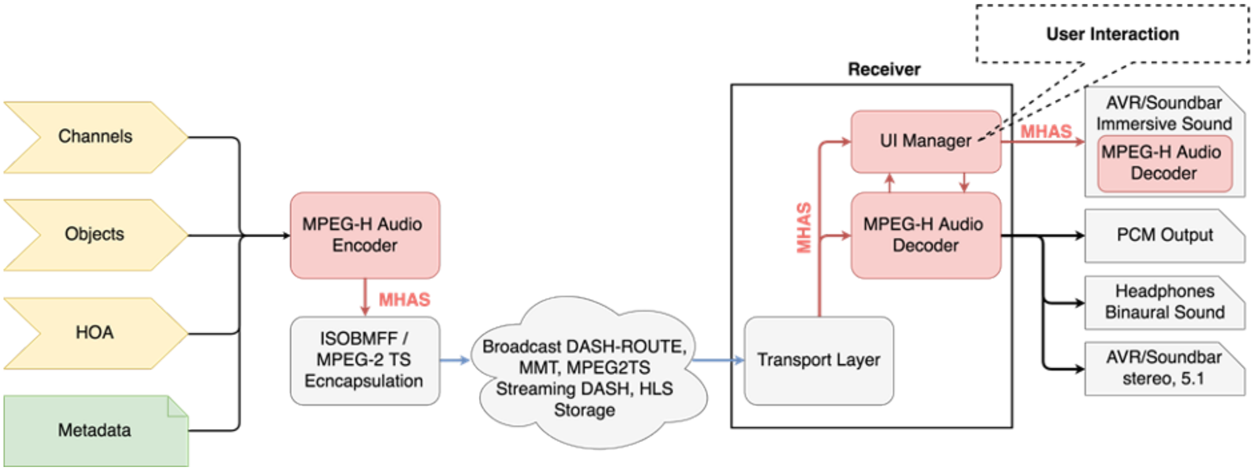


Figure 26. MPEG-H Audio system overview

The MPEG-H Audio Stream (MHAS), described in Sec 8.3.3 of the Indigo Book , contains the audio bitstream and various types of metadata packets and represents common layer for encapsulation into any transport layer format (e.g., MPEG-2 TS, ISOBMFF). The MPEG-H Audio enabled receiver can decode and render the audio to any loudspeaker configuration or a Binaural Audio representation for headphones reproduction. For enabling the advanced user interactivity features in cases where external playback devices are used, the UI Manager 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 Sec 8.3.1.4 of the Indigo Book.

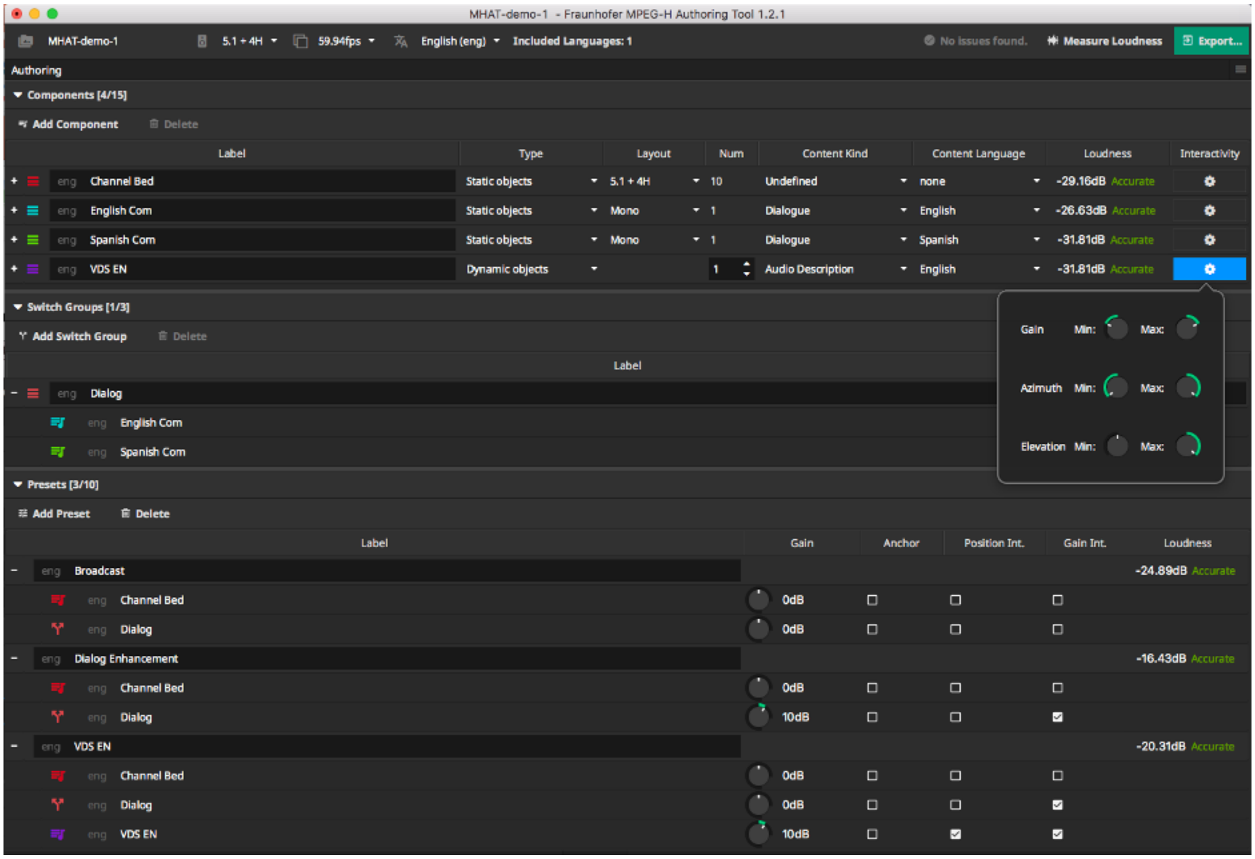
All MPEG-H Audio features that are described in the following sections are supported by the MPEG-H 3D Audio Low Complexity Profile Level 3 and are thus available in all broadcast systems based on the DVB and ATSC 3.0 specifications. See Table 3 for the characteristics of the Low Complexity Profile and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

Table 3. Levels for the Low Complexity Profile of MPEG-H Audio

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

7.2.9.1. 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 27. (See also INDIGO Book, section 8.3.2 MPEG-H Audio Metadata.)



MHAT-demo-1 - Fraunhofer MPEG-H Authoring Tool 1.2.1

MHAT-demo-1 | 5.1 + 4H | 59.94fps | English (eng) | Included Languages: 1 | No issues found. | Measure Loudness | Export...

Authoring

Components [4/15]

Add Component | Delete

Label	Type	Layout	Num	Content Kind	Content Language	Loudness	Interactivity
eng Channel Bed	Static objects	5.1 + 4H	10	Undefined	none	-29.16dB Accurate	⚙
eng English Com	Static objects	Mono	1	Dialogue	English	-26.63dB Accurate	⚙
eng Spanish Com	Static objects	Mono	1	Dialogue	Spanish	-31.81dB Accurate	⚙
eng VDS EN	Dynamic objects		1	Audio Description	English	-31.81dB Accurate	⚙

Switch Groups [1/3]

Add Switch Group | Delete

Label
eng Dialog
eng English Com
eng Spanish Com

Presets [3/10]

Add Preset | Delete

Label	Gain	Anchor	Position Int.	Gain Int.	Loudness
eng Broadcast					-24.89dB Accurate
eng Channel Bed	0dB	□	□	□	
eng Dialog	0dB	□	□	□	
eng Dialog Enhancement					-16.43dB Accurate
eng Channel Bed	0dB	□	□	□	
eng Dialog	10dB	□	□	☑	
eng VDS EN					-20.31dB Accurate
eng Channel Bed	0dB	□	□	□	
eng Dialog	0dB	□	□	☑	
eng VDS EN	10dB	□	☑	☑	

Figure 27. MPEG-H Authoring Tool example session

7.2.9.2. Universal Delivery

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

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

7.2.9.3. 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 3).

7.2.10. 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:

- Set-Top Box connecting to a TV over HDMI
- 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 28 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.

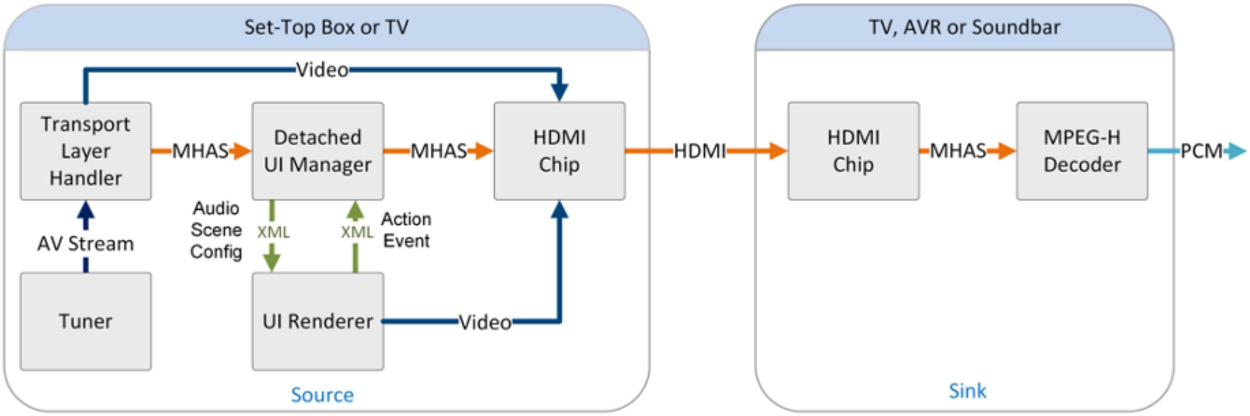


Figure 28. Distributed UI processing with transmission of user commands over HDMI

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

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

7.3. High Frame Rate Video (HFR)

For the purpose of this document, High Frame Rate (HFR) refers to frame rates of 100 fps or higher, including 100, 120/1.001[3] and 120, Standard Frame Rate (SFR) refers to frame rates of

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

60fps or lower, which are commonly used including 24/1.001, 24, 25, 30/1.001, 30, 50, 60/1.001 and 60.

According to a SMPTE/HPA paper authored by Mark Schubin[4] frame rates of 100, 120/1.001 and 120 fps add significant clarity to high motion video such as sports or action scenes. Schubin also notes that high dynamic range puts new demands on temporal resolution. He notes that, "Viewers of HDR imagery sometimes report increased perception of motion judder... Increased frame rate, therefore, might be necessary to accompany HDR."

Citing Schubin again, "... the [EBU[5]] found ... that in going from 60 frames per second (fps) to 120 fps or from 120 fps to 240 fps — a doubling of the frame rate — it is possible to achieve a full grade of improvement." Further, doubling the frame rate from 50/60 fps to 100/120 fps is a very efficient means of gaining that full grade of improvement when compared to going from 2K to 4K spatial resolution, as illustrated in Figure 29.

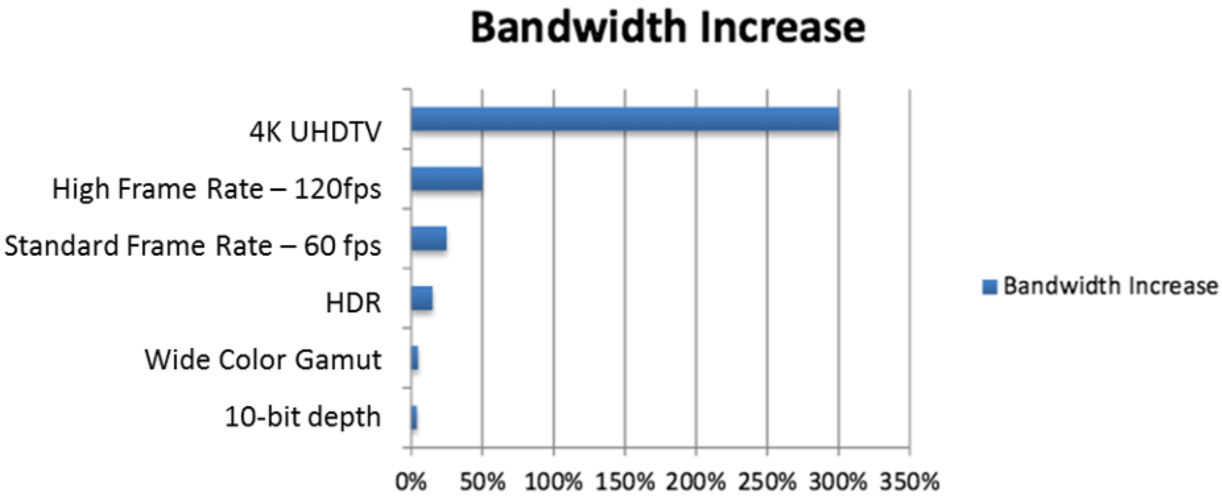


Figure 29. Bandwidth increases for various video format improvements compared to HD

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

[5] Rep. ITU-R BT.2246, The present state of ultra-high definition television; [130],p 26.

HFR has been included in newer DTT television standards including ATSC 3.0 [54] and DVB [63]. As such, the Ultra HD Forum considers HFR to be viable Ultra HD technology that can be layered onto Foundation Ultra HD for DTT. Both systems include a backward compatibility mechanism that enables 50/60 fps decoders to render a 50/60 fps version of the content while 100/120 fps decoders render the full HFR experience. See Section 7.3.1. for more information about backward compatibility.

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

Table 4. 2k High Frame Rate Content Parameters

Frame Rate	**Spatial Resolution**	**Scan Type**	**Dynamic Range**	**System Colorimetry**	**Bit Depth**	**Distribution Codec**	**HDMI Interface**[6]
100, 120[7]	HD	Progressive	SDR, HDR	709, 2020	10	HEVC Main 10 Level 5.1	100fps: at least 1.4 120 fps: at least 2.0

7.3.1. Backward Compatibility for HFR

Both DVB UHD-1 Phase 2 (ETSI TS 101 154 [63]) and ATSC 3.0 (A/341) [54] include frame rates up to 120 fps. Both documents further include optional temporal sub-layering for backward compatibility to a frame rate half of the HFR. According to A/341, achieving backward compatibility by rendering every other frame may cause unwanted strobing. ATSC 3.0 includes optional temporal filtering that reduces or removes strobing artifacts from the standard frame rate picture when temporal sub-layering is used. Further frame rate reduction to 25/30 fps will worsen any strobing, and the temporal filter included in ATSC 3.0 cannot prevent strobing artifacts at framerates below 60fps.

Both DVB and ATSC make use of the HEVC [69] Temporal Sub-layers technology to label every other frame for use by a 50/60 fps decoder.

In the case that an HFR video stream is available, an SFR stream may be extracted by dropping every other picture. HEVC temporal sub-layering identifies every other picture, which enables

[6] HDMI interfaces that support 10-bit, 1920x1080p, SDR high frame rate. HDMI 1.4 also supports 120fps with 4:2:2 chroma sub-sampling. HDR support requires HDMI 2.0a for HDR 10 and 2.0b for HLG.

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

division of the stream prior to decompression. Note that strobe effects may be present when dropping every other frame without applying any filtering. Filtering systems such as the one shown in Figure 30 can mitigate this effect.

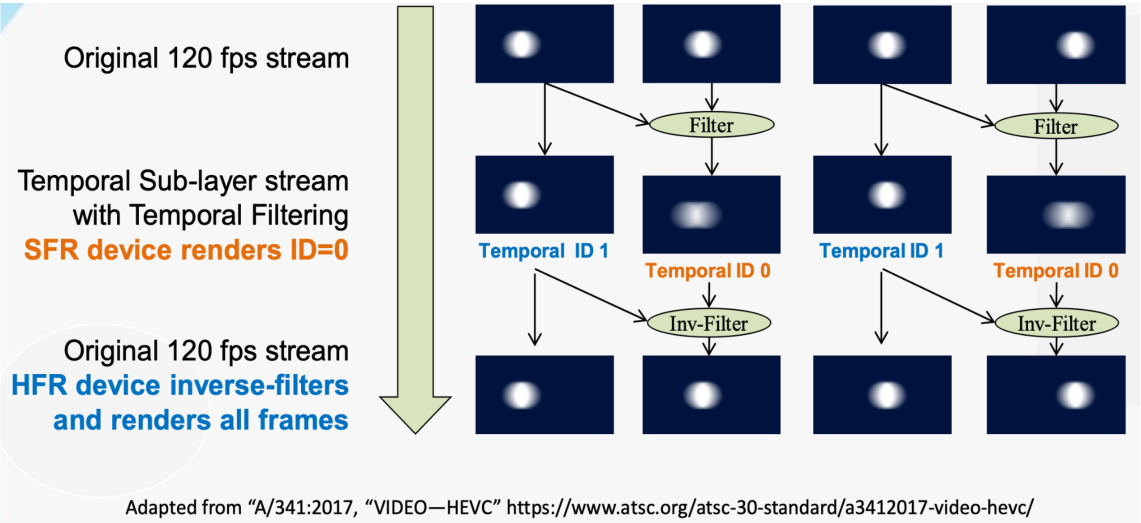
The SFR frames are Temporal ID = 0 and the additional frames needed for HFR are Temporal ID = 1. SFR devices render the frames with ID = 0 and HFR devices render all frames, i.e., ID = 0 and ID = 1.

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

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

The process of recording of HFR content in either compressed or uncompressed form should take into account the possibility that the content may undergo transformations in downstream processing to make the content backward compatible with SFR TVs, using one of the mechanisms described in this Section. Information that will be consumed by an SFR TV must be embedded in the images that will form the base layer of the temporally layered stream. For example, if CTA-608 [18]/CTA-708 [19] captions are stored in the uncompressed image data, this data should be stored in the image data that would become the base layer and not the enhancement layer (since the SFR TV will not have access to the latter).

ATSC 3.0 includes an additional feature for backward compatibility called Temporal Filtering. Temporal Filtering is a method by which consecutive HFR frames are averaged to create SFR frames, in order to maintain motion blur and prevent strobing. Averaging frames evenly may cause double images, depending on the shutter interval, so a weighted average may be used in order to optimize the SFR experience. The SFR device plays the filtered SFR frames. The HFR device recovers the original, pre-filtered HFR frames and renders all frames (i.e., the pre-filtered frames are optimized for HFR). See Figure 30 below.



Adapted from "A/341:2017, "VIDEO—HEVC" https://www.atsc.org/atsc-30-standard/a3412017-video-hevc/

Figure 30. ATSC 3.0 temporal filtering for HFR backward compatibility

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

A DVB broadcaster may ingest HFR content in compressed format that uses ATSC 3.0 temporal layering. The converse may also occur, where an ATSC 3.0 broadcaster ingests content in DVB dual layer format. Since these use cases will typically also involve frame rate conversion (between say 100p and 120p), which will require transcoders to be used, these transcoders can also convert the temporal layering from one format to the other, or convert between a temporally layered HFR stream and a single-layered HFR stream. When transcoding to an ATSC 3.0 temporally layered HFR stream, temporal filtering for judder reduction can also be implemented by the transcoder if so desired.

7.3.2. Production Considerations for HFR

As the mainstream end to end ecosystem still constrains HFR to 1080p 100 or 120 frame rates, interfaces between production systems such as cameras, switchers, storage and playout servers require no more than 3G SDI capability. Current state-of-the-art systems either incorporate these interfaces or are in the process of being revised to incorporate this capability.

The payload of HFR 1080p content can be carried via the SDI interface as described in SMPTE ST 425 [80], 2081 [83] and 2082-10 [84] document families using 3G or 12G interfaces. (6G is also possible, but not used in common practice.) Some examples include:

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

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

Carriage of HFR 1080p content via IP networks is described in SMPTE ST 2022 [81]. SMPTE ST 2022-6 [82] describes how to map SDI info to IP. The SMPTE ST 2110 [43]-[47] standards suite specifies the carriage, synchronization, and description of separate elementary essence streams over IP for real-time production, playout, and other professional media applications; i.e., it describes how each element of essence is mapped to IP.

Further work is required to determine the requirements of improvements to ecosystems to support HFR beyond 1080p.

7.4. Encoding

7.4.1. AVS2 & AVS3

The Digital Audio and Video Coding Standard Working Group (AVS Workgroup) of China delivered their Advanced Video Coding Standard (AVS2) to target UHD and HDR content for both broadcast and broadband communications and for storage. AVS2 was adopted by the State Administration of Radio, Film, and Television (SARFT) as the UHD video standard for industry [118] in May, 2016 and by the General Administration of Quality Supervision, Inspection and Quarantine (GAQSIQ) as the Chinese national standard [117]for UHD video, in December, 2016. Both were published in Chinese, and the English language version was standardized by the IEEE as 1857.4 [119] in July 2018.

This Annex was originally added to the Guidelines in April of 2018 and was directed exclusively to AVS2. Since then, a 3rd generation AVS codec (AVS3) and companion HDR format (Vivid HDR) have been endorsed by the China Ultra HD Video Industry Alliance (CUVA), formalized by the Digital Audio and Video Encoding and Decoding Technology Standard Working Group (AVS

Work Group) in October 2021. AVS3 is included, along with AVS2, as a part of China's "5G + 4K/8K Ultra HD Production and Broadcasting Demonstration Platform" project, led by the Central Radio and Television General Station. An English translation of the first phase main profile of AVS3 is available on request[8] and work on a second phase is underway. DVB, having evaluated the performance of the AVS3 codec, adopted AVS3 in its recent revision of the DVB-AVC specification[9], as a codec to be included in the DVB toolbox.

7.4.1.1. Why AVS2

AVS2 is the successor to the earlier video coding standard AVS+ [120], which was successor-in-turn to AVS1 [121][122]. AVS2 has double the coding efficiency of AVS1. Testing hosted by the State Administration of Radio, Film, and Television (SARFT) determined that AVS2 compared favorably to HEVC, producing slightly less image degradation relative to source images at the same bitrate. Using 4K video sequences (2160p 10-bit) specified by China's National Film and Television Administration, a test identifying specific builds of reference software demonstrated AVS2 to have a 3.0% average performance advantage relative to HEVC[10], while the decoder complexity is similar.

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

7.4.1.2. Deployment

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

The predecessors to AVS2 have seen widespread deployment: AVS+ is presently used widely in China, Sri Lanka, Laos, Thailand, Kyrgyzstan, and other countries; while AVS1 is further used in Burma, Cuba, and Uzbekistan. In December 2017, Guangdong Radio and Telev2ision (GRT)

[8] http://www.avs.org.cn/english/EnglishSpec.asp

[9] DVB BlueBook A001r21 (Interim draft TS 101 154 V2.8.1) - November 2022, which adds not only AVS3, but also MPEG's Versatile Video Coding (VVC) codec.

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

announced a pilot for China’s first 4K UHD channel and its use of AVS2. Growth of that channel is expected to reach 15M subscribers in the Guangdong province alone.

In June 2018, China Central Television (CCTV) conducted a live broadcast demonstration of the 2018 World Cup, which used AVS2 to distribute PQ images with SL-HDR2 metadata.

7.4.1.3. Technology

AVS2 uses a coding framework as shown in Figure 31. The residual between an image and a prediction of the image is compressed by the transform and quantization module. Thereafter, de-quantization and inverting the transform reconstitutes the residual and encoded image (green modules). Two classes of predictor are available: Intra Prediction (orange modules) to detect and exploit similarities within the encoded image itself, and inter prediction (gold modules) to detect and exploit similarities to other reconstituted images. Coefficients representing the compressed residual and the detected similarities from intra- or inter prediction are collected and further compressed by entropy coding to produce the AVS2 bitstream.

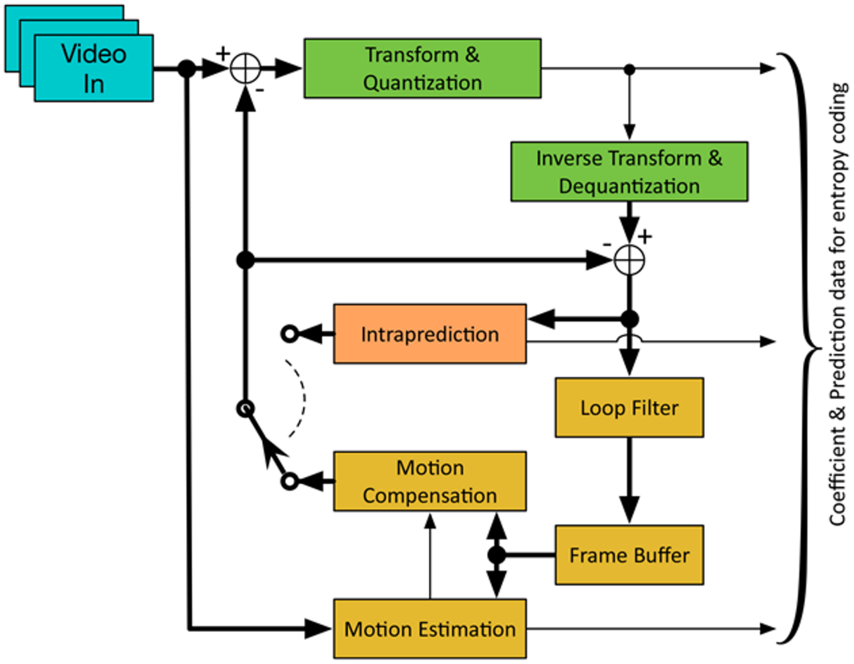


Figure 31. AVS2 coding framework

Improvements offered in AVS2 over prior codecs include new intra prediction modes (e.g., chrominance derived from luminance) and long-term reference frames in intra prediction modes. A significantly more detailed description of the AVS2 technologies and the advances in AVS2 can found in an IEEE paper published in 2015[11].

Where there are similarities between HEVC and AVS2, for example the overall processing flow, quad-tree partitioning, certain prediction modes, and motion vectors, transcoding from HEVC to AVS2 can be especially efficient[12].

[11] S. Ma, T. Huang and W. Gao, "The second generation IEEE 1857 video coding standard," 2015 IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP), Chengdu, China, 2015, pp. 171-175, doi: 10.1109/ChinaSIP.2015.7230385. https://ieeexplore.ieee.org/document/7230385

[12] Yucong CHEN, et al., “Efficient Software HEVC to AVS2 Transcoding”, Information, 2016, 7, 53

7.4.2. Content Aware Encoding (CAE)

7.4.2.1. Introduction

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

We will describe in this chapter how CAE works, how it can be applied to Ultra HD and the benefits of using CAE for the transmission of Ultra HD program material. CAE is not a standard, but a technique applied on the encoder side that is expected to be decoded by an HEVC Main 10 decoder. Regarding adaptive streaming, the only existing specification is iOS 11[13]. However MPEG DASH IF supports both CAE and ABR encoding.

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

7.4.2.2. Adaptive Bitrate Usage for Ultra HD

For OTT, ABR is already the most common way to deliver content. CAE is applied on top of ABR in the encoding process. Currently only iOS11[14] has done that, but we expect wider support such as from Android, DASH, and DVB in the future. For managed IP networks (Cable, Telcos), we also see ABR being used.

Cable operators can broadcast Live over either QAM or ABR or over IP (DOCSIS® 3.0 [77]). The IP delivery may be performed in Unicast as the traffic is not expected to be high, and may later be scaled using ABR Multicast CableLabs [58].

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

[13] https://developer.apple.com/documentation/http_live_streaming/hls_authoring_specification_for_apple_devices/

[14] ibid

7.4.2.3. Per-title Encoding

Content aware encoding was introduced in production by Netflix®[15] in 2015 using “per-title encoding”[16]. In summary Netflix discovered that the ABR ladder defined for the video encoding was very much dependent on the content and that for each title they would consider an optimized ladder where each step provides a just noticeable difference (JND) in quality (originally using PSNR, Netflix developed the VMFA metric) at the lowest bitrate. In addition, as the content complexity changes during a movie, the bitrate per resolution should also vary. Netflix later refined that model, by changing the ladder not per-title, but per-segment.

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

Some of the more recent implementations of CAE ladder generators reviewed[17] do not require full additional transcodes to be done ahead of time, making them more practical, and applicable in both VOD and Live use cases.

7.4.2.4. VBR Encoding

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

A drawback of traditional VBR streaming is that the bitrate of an encoded stream can be very high during complex scenes, putting OTT content providers at risk of exceeding the streaming bandwidth supported by the network. The maximum bitrate is chosen based on a combination of network bandwidth limitations and the video quality delivered during complex scenes. Setting a

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

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

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

ceiling for the maximum bitrate of the stream, known as Capped VBR (CVBR), resolves this issue by protecting the streaming bandwidth. But the technique is not infallible.

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

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

7.4.2.5. Content Aware Encoding Overview

Content Aware Encoding or Content-Adaptive Encoding (CAE) is a class of techniques for improving efficiency of encodings by exploiting properties of the content. By using such techniques, “simple” content, such as scenes with little motion, static images, etc. will be encoded using fewer bits than “complex” content, such as high-motion scenes, waterfalls, etc. By so doing, content-aware techniques aim to spend only a minimum number of bits necessary to ensure quality level needed for delivery. Since “simple” content is prevalent, the use of CAE techniques results in significant bandwidth savings and other benefits to operators (e.g., some systems may also reduce the number of encodings, deliver higher resolution in the same bits as the previous systems required for lower resolution, better overall quality, etc.). The CAE process is the “secret sauce” of an encoder company as described in several references[18].

[18] http://info.harmonicinc.com/Tech-Guide-Harmonic-EyeQ

http://media2.beamrvideo.com/pdf/Beamr_Content_Adaptive_Tech_Guide.pdf

https://www.brightcove.com/en/blog/2017/05/context-aware-encoding-improves-video-quality-while-cutting-costs

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

7.4.2.6. Principle Uses

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

CAE techniques can be applied at different levels, described in Table 5.

Table 5. CAE granularity

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

Table 5 Notes: *This might bring unacceptable additional delay (latency).

7.4.2.7. Content Aware Encoding applied to Ultra HD

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

Table 6 provides examples of three ABR encodings ladders. Note that these are examples provided to give the reader an indication of the bitrates that may be possible; however, the nature of the content and other factors will affect bitrate. All ladders use the same set of DVB-DASH-recommended resolutions [60], ranging from HD (720p) to UHD (2160p), but they differ in rates. The first ladder (shown in column 4) is a fixed CBR encoding design, assigning bitrates that are chosen independently, regardless of the type of content being encoded. The ladders shown in columns 5 and 6 are examples of CAE ladders generated for two different types of content. The CAE ladder in column 5 is produced for easier-to-encode content resulting in an average savings of more than 50%. The CAE ladder in column 6 is produced for more difficult content, resulting in an average savings of 40-50% vs. CBR encoding, depending on the content complexity.

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

Table 6. Examples of fixed and CAE encoding ladders for live sports

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

We draw in Figure 32 the bitrate vs. resolution of CAE vs. CBR at the same quality level. For simplicity for CAE, we use a more conservative example ladder, resulting in 40% savings, assuming the same visual quality at a given resolution.

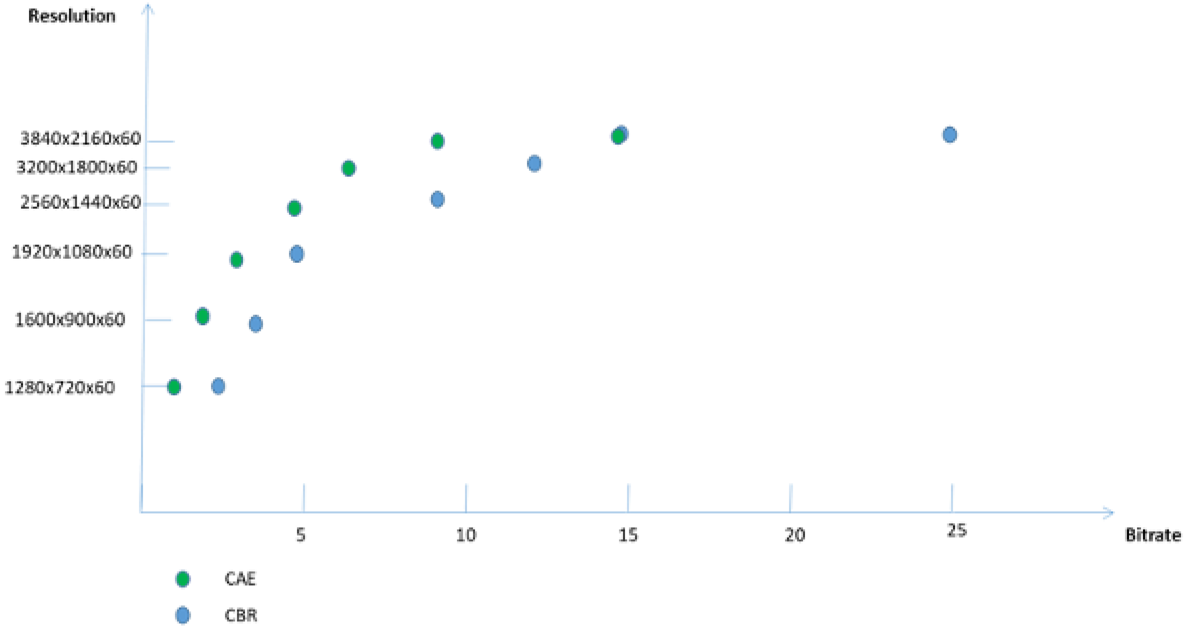


Figure 32. CAE encoding chart

7.4.2.8. Content Aware Encoding interoperability

The resulting bitstream from a CAE encoder is compliant with the guidelines for ABR delivery used in DVB-DASH [60] and Apple TV / HLS [67].

A key point for the player is that it should support variable bit rate bitstreams.

7.4.2.9. Application for Content Aware Encoding

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

From Belson[19], Figure 33 shows the Internet speed distribution over various regions of the world.

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

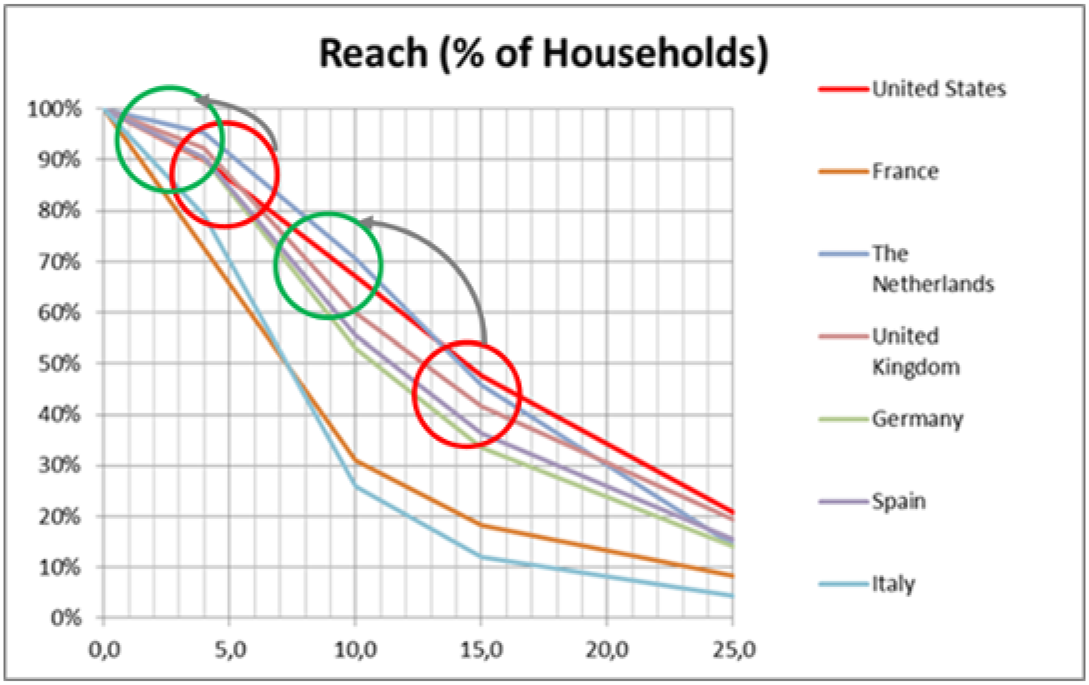


Figure 33. Internet speed distribution per countries (source Akamai)

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

2160p60 use case

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

1080p60 use case

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

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

7.4.2.10. CAE Sweet Spot for UHD

Based on the previous section finding, the CAE sweet spot is when 2160p60 can be delivered at a lower bitrate than the CBR case. We describe in Figure 34 the CAE sweet spot.

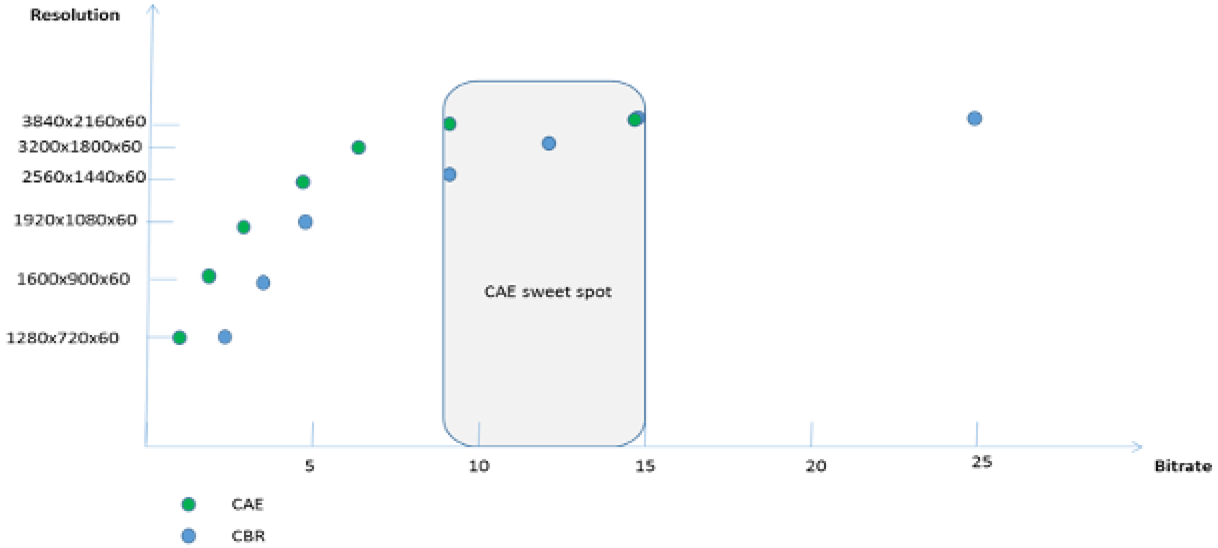


Figure 34. CAE sweet spot vs. CBR

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

7.4.2.11. Content Aware Encoding Benefits

CDN cost

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

Quality of experience

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

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

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

From a purely qualitative point of view, at junction bitrates (i.e., bitrates where the CAE encoding is at a higher resolution than the CBR encoding), the quality will be improved as a higher resolution will be displayed. The junction bitrates are depicted in Figure 35.

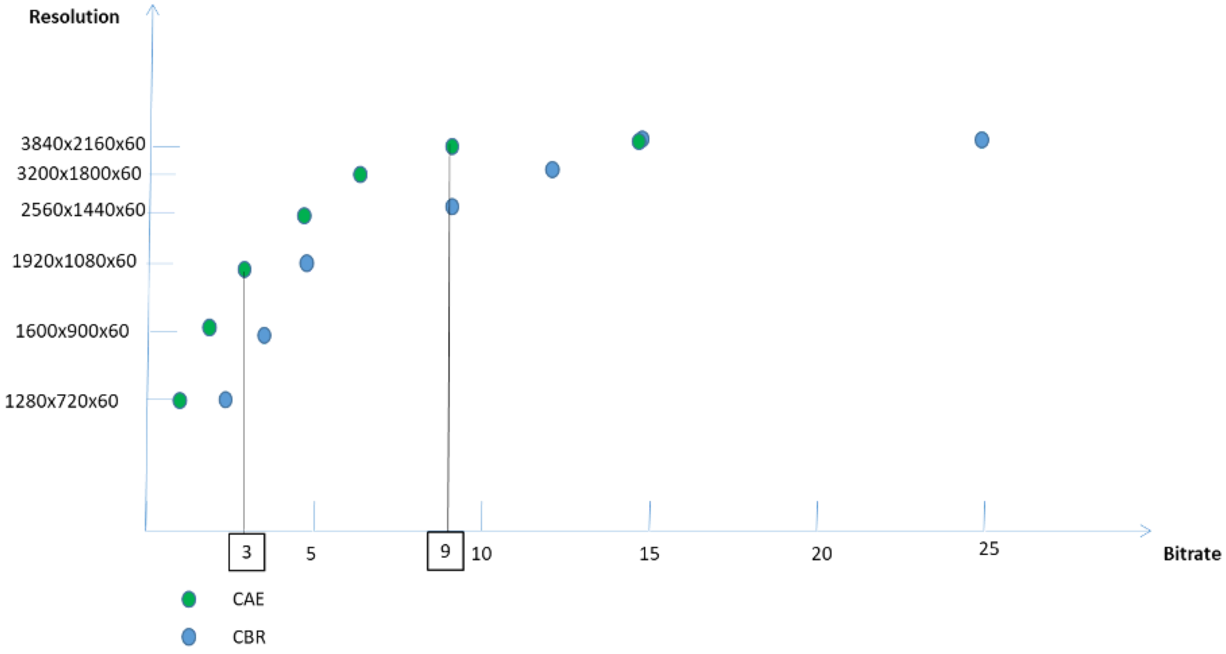


Figure 35. Junction bitrates chart

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

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

8. Considerations in Using Technology Beyond Foundation

See Section 10.4 in Violet Book.

8.1. Dynamic HDR Metadata

For PQ HDR content[21], as described in Orange Book, Section 7.2.1., HDR10 provides the static metadata elements in a PQ10-based HDR format as specified by SMPTE ST 2086, MaxFALL, and MaxCLL. That section identifies a number of limitations with these particular HDR metadata values, notably the difficulty with setting these values in a live environment, real-world experience suggesting that these values have been set to artificial numbers to force certain looks on consumer displays, and the inability to correctly set these values given limitations of mastering displays.

In addition to these limitations, the values of MaxFALL and MaxCLL are also very limited in that they are only currently specified to provide single values for the entirety of the program. The dynamic range of both narrative and live content can vary dramatically from scene to scene. As a result the static, program-wide metadata values, as strictly defined, are of limited use for a great deal of content that does not have a static, unchanging dynamic range. Interoperability tests show that receivers can recognize changes in the static metadata within the duration of a program; however, it is yet unknown how frequently or quickly such changes can be recognized. For example, it is not expected that static metadata would change on a frame-by-frame basis.

Finally, there is no standardized way of utilizing these values in the final consumer display, so displays differ significantly in reproduction of the image. In practice some displays may ignore the values altogether. This is not consistent with the goal of displaying the image as close to the creative intent as possible on the target display.

[21] HLG10 does not specify any display metadata as it is based on normalized scene-light, rather than the absolute luminance of the signal seen on the mastering display, as described in Orange Book Section 7.2.2. As such, the headroom (measured in f-stops) for HLG highlights above HDR Reference White, is approximately constant regardless of the display's nominal peak luminance. Moreover the HLG10 display EOTF, which is fully specified by the ITU-R BT.2100 [5], includes a variable display gamma to provide adjustment for a specific display's peak brightness capabilities, along with eye adaptation; thereby allowing HLG to function in brighter viewing environments. Thus static or dynamic metadata is not required for HDR productions using HLG10.

A number of Dynamic Metadata methodologies have been developed to address the limitations of PQ10 and HDR10. Dynamic Metadata refers to metadata that describes the image at a much finer temporal granularity, scene-by-scene or even frame-by-frame and produces significantly more information about the mastering and creative intent of the scenes. In addition, most of these methodologies provide detailed information about tone mapping in the consumer display with the goal of consistent images across different manufacturers' displays. The methodologies also are designed to preserve creative intent, with the final displayed image being as close to the mastered image as the consumer display has the ability to reproduce.

Some of these methodologies go further by capturing the metadata during the color grading session and passing that metadata to consumer displays to better reproduce the creative intent. Most metadata schemes also provide for automatic metadata creation, which is useful in workflows for live content.

In general, several of these dynamic metadata schemes are additive, in that they provide additional information about the carried PQ10 image, and the HDR10 static metadata remains intact alongside the dynamic metadata. In some cases, this can provide a simple backwards compatibility to an HDR10-only display - the dynamic metadata is simply ignored.

Finally, many of these methodologies have considered how the signal can be backwards compatible with SDR displays and have built-in methods for conversion. See Dolby Vision™ described in Section 7.1.1. and SL-HDR2 described in Section 7.1.3..

SL-HDR1 is another HDR dynamic metadata technology, which serves a different purpose than Dolby Vision or SL-HDR2. SL-HDR1 is intended to enable the service provider to emit an HDR/2020 service in an SDR/709 format that can be "reconstructed" to HDR/2020 by the receiver. HDR/2020 receivers that can interpret the SL-HDR1 metadata can present the HDR/2020 format to the viewer. The SDR/709 content can be displayed by receivers that cannot display HDR/2020. In this way SL-HDR1 provides a measure of backward compatibility for both HLG and PQ-based HDR content. It should be noted that SL-HDR1 requires 10-bit encoding, and so may not help address legacy SDR/709 receivers that are only capable of 8-bit decoding. See Section 7.1.3.1.

8.2. Next Generation Audio

8.2.1. Objects, static and dynamic

Complementing the visual enhancements that Ultra HD will bring to consumers, Next Generation Audio (NGA) provides compelling new audio experiences, as follows, with more details provided in Section 8 of the Indigo book:

- Immersive – An audio system that enables high spatial resolution in sound source localization in azimuth, elevation and distance, and provides an increased sense of sound envelopment
- Personalized – Enabling consumers to tailor and interact with their listening experience, e.g. selecting alternative audio experiences, switching between languages, enhancing dialogue intelligibility.
- Consistent – Playback experience automatically optimized for each consumer device, e.g. home and mobile
- Object-based Audio – Audio elements are programmed to provide sound from specific locations in space, irrespective of speaker location. By delivering audio as individual elements, or objects, content creators can simplify operations, reduce bandwidth, and provide a premium experience for every audience
- Scene-Based Audio – An arbitrarily large number of directional audio elements composing a 3D sound field are mixed in a fixed number of PCM signals according to the Higher-Order Ambisonics format. Once in the HOA format, the Audio Scene can be efficiently transmitted, manipulated, and rendered on loudspeaker layouts/headphones/soundbars.
- Flexible Delivery - NGA can be delivered to consumers over a number of different distribution platforms including terrestrial, cable, and satellite broadcast, IPTV, OTT, and mobile. It could also be delivered over a hybrid of broadcast and OTT
- Flexible Rendering - NGA can be experienced by consumers through headphones or speakers (e.g., TV speakers, home theater systems including ceiling speakers, sound bars) as shown in Figure 36.

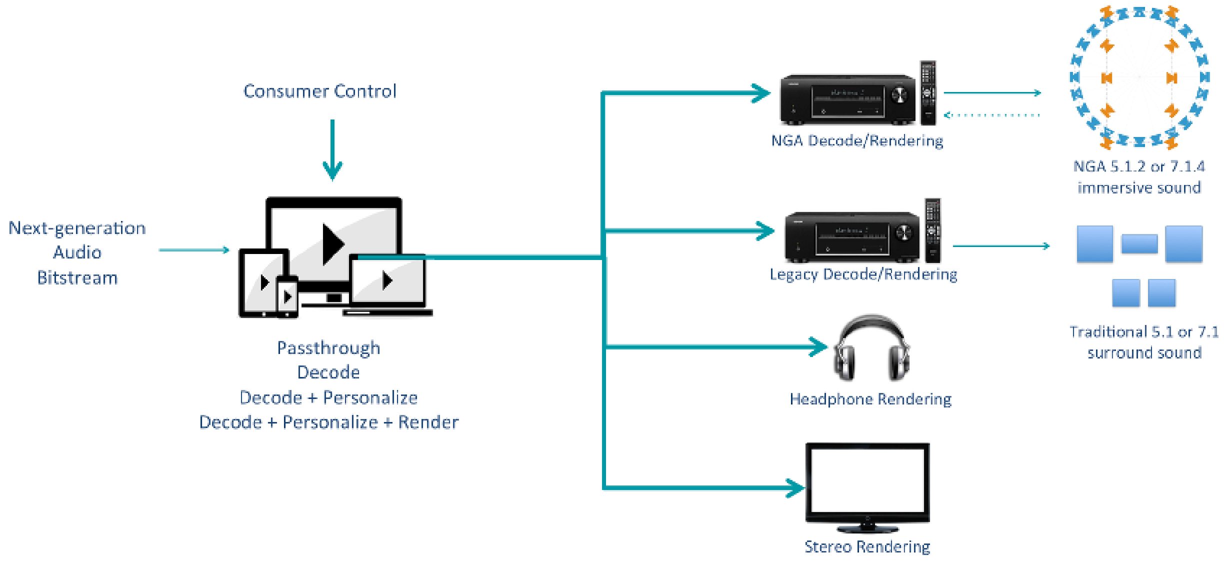

Consumer Control
Next-generation Audio Bitstream
Passthrough
Decode
Decode + Personalize
Decode + Personalize + Render
NGA Decode/Rendering
NGA 5.1.2 or 7.1.4 immersive sound
Legacy Decode/Rendering
Traditional 5.1 or 7.1 surround sound
Headphone Rendering
Stereo Rendering

Figure 36. NGA in the consumer domain

8.2.2. Personalization

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

There are two main concepts of personalized audio:

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

9. Maintaining Dynamic Range and System Colorimetry Parameters

Different dynamic range and system colorimetry parameters should not be mixed in a Real-time Program Service. For example, service operators should not shift between HLG10, PQ10 or SDR/BT.709. Decoders require time to adjust to different encoded data settings – as much as 2 seconds or more has been observed by Ultra HD Forum members – causing a poor consumer experience. OTT providers offering ABR streams must also ensure that the adaptive bitrate streams are all of the same transfer curve and system colorimetry[22]. Similar to the linear progression of a program through time, the progression of rendering successive levels of ABR streams requires quick adjustments on the part of decoders. If a Real-time Program Service must contain a switch point between dynamic range and system colorimetry, it is recommended that such switches be performed overnight or in a maintenance window and black frames be inserted at switch points to allow decoders and viewers to adjust to the new content characteristics.

It is possible to direct map SDR/BT.709 content into HLG10 or PQ10, to up-map SDR/BT.709 content to HLG10 or PQ10 and vice versa, and to convert content from PQ10 to HLG10 or vice versa (see ITU-R report BT.2408 [8]). The following subsections offer guidelines for normalizing content in the headend for this purpose. See also Section 10.4 in the Violet Book for conversion possibilities in consumer equipment for backward compatibility.

[22] See “Guidelines for Implementation: DASH-IF Interoperability Points”, Section 6.2.5, [16]

10. ANNEXES

10.1. Dynamic Resolution Encoding

10.1.1. Presentation

Video compression experts all know that when bandwidth is reduced, a good trade-off to preserve quality and limit visible compression artifacts is to reduce the resolution. Of course, the best resolution for a given bitrate highly depends on the video content. In December 2015, Netflix popularized the concept of variable resolution encoding, in its blog “Per-Title Encoding Optimization”[23]. At the time, the best resolution was selected for each VOD content. In the following years, Netflix improved the concept with Dynamic Resolution Selection, applied for each scene. As Netflix’s market is VOD, this selection can be made offline but a viewing-based selection would be much too time consuming. To mitigate this issue, Netflix developed an objective Video Quality (VQ) metric, called VMAF, to help in the resolution selection process automation. This technique cannot be applied to live content, as it would require too much processing power and can’t be used in real time.

Harmonic has implemented a similar concept on live content with a very dynamic selection, applied for each video delivery segment that is a few seconds in duration.

The method proposed below is based on a machine learning (ML) mechanism that learns how to pick the best resolution to be encoded in a supervised learning environment. At run time, using the already existing pre-processing stage, the live encoder can decide on the best resolution to encode, without adding any processing complexity or delay. This results in higher quality of experience (QoE) or lower bitrate, as well as lower CPU footprint vs. a classical fixed ladder approach.

The first section below explains how the Dynamic Resolution Selection can work in a live workflow. The second section presents the identified use cases for DRE deployments. The next two sections go through the results of quality and interoperability evaluation. The last paragraph concludes with a discussion of results and outlooks.

[23] Per-Title Encode Optimization, December 2015, https://netflixtechblog.com/per-title-encode-optimization-7e99442b62a2.

10.1.2. Dynamic Resolution Selection for Live video delivery

The per-scene encoding optimization from Netflix makes use of multiple encodings as well as resolution selection based on real VQ measurement, as shown in Figure 36. This solution would be much too computationally intensive and would result in too much delay for live streaming.

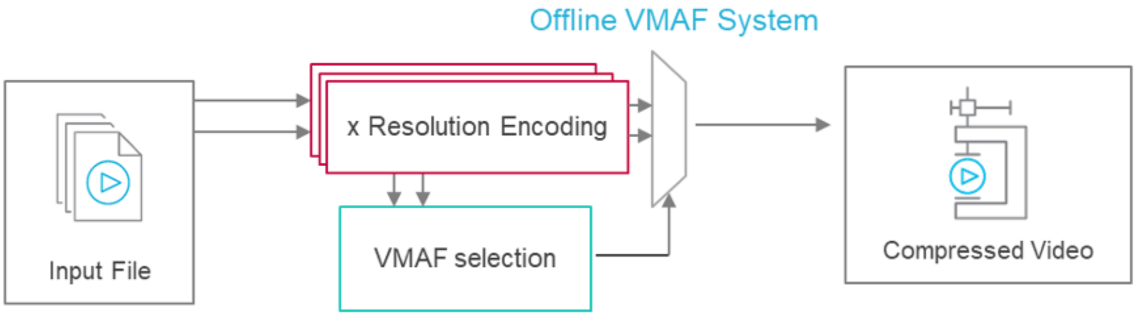


Figure 36. VOD Dynamic Resolution Encoding

A live encoding system is characterized by a limited look-ahead of picture analysis before encoding and delivery stages. Therefore, it is not realistic to take a decision for a global scene that can be of a variable duration. Instead, a very dynamic decision scheme is built with a selection of resolution applied for each video segment of limited duration (typically two to three seconds). Features from the video pre-analysis stage have been used to train a ML-based prediction model, offline, in a supervised learning environment, as shown in Figure 37 below. Each video segment is encoded with various resolutions at a given target bitrate (constant or capped bitrate). At run time, using the already existing pre-analysis stage and the same target bitrate, the live encoder can decide on the best resolution to encode the pictures of the current segment by using the prediction model created offline. The best resolution is the one that provides the best visual quality.

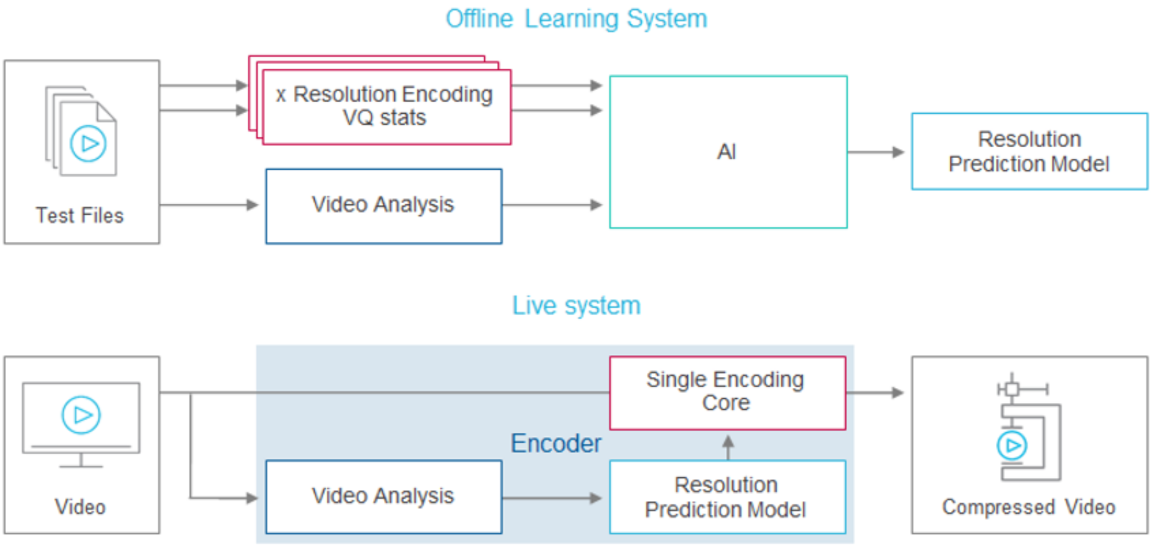


Figure 37. AI-based Live Dynamic Resolution Encoding

The increase of processing complexity is very limited since the prediction model is a decision tree algorithm driven by features that were already computed. The additional delay is, as always, a trade-off with quality. A pre-analysis of the full segment duration can better tackle a change happening in the last part of the segment but would add a significant delay. The use of the classical look-ahead adds no delay but takes the decision on the first part of the segment and may react to a change with one segment delay. One way to limit the impact is to work with segments aligned with scene changes.

10.1.3. Dynamic Resolution Encoding benefits

10.1.3.1. OTT Streaming

The first use case of Dynamic Resolution Encoding (DRE) that comes to mind is OTT streaming with the most widely used DASH and HLS delivery formats. In this type of delivery, a ladder of profiles with various bitrates is built for video encoded representations so that the client can adapt to the bandwidth fluctuations by requesting the appropriate representation. The ladders are built using average statistics of best resolution per bitrate and do not take into account the individual video content characteristics. With a DRE scheme, the resolution will vary within all or selected bitrate profiles.

Use of DRE approach can result:

- in bandwidth savings for the same QoE by using a lower constant (CBR) or capped (cVBR with Content Aware Encoding) bitrate for the highest profiles thanks to a lower resolution when the content is too challenging for the highest resolution,
- in a better QoE by using a higher resolution on static scenes at the lowest bitrate profiles (higher sharpness) or a lower resolution on temporally complex scenes at the highest bitrate profiles (less compression artifacts),
- in storage savings by reducing the number of profiles for the ladder,
- in CPU savings by lowering the resolutions for the most complex scenes on the highest profiles and by reducing the number of profiles.

The implementation of DRE depends on the OTT delivery format. With DASH-based OTT streaming, the manifest can indicate the maximum resolution for each bitrate profile and not the real resolution that every segment will use. To make sure the decoder in the OTT player can properly handle such a stream, the video representations will have inband signaling of resolution using avc3 MP4 brand for AVC, hev1 MP4 brand for HEVC, and vvi1 MP4 brand for VVC in the MP4 container. This means that the player/decoder will get the resolution from the high-level syntax of the encoded stream. When being served with a lower resolution that the one indicated in the manifest for the requested profile, the player/decoder will decode it and upscale to the nominal resolution set in the manifest.

For HLS-based OTT streaming, one Initialization file per resolution is created. Each time the resolution changes, a #EXT-X-MAP tag with reference to the proper Initialization file is added in the playlist.. This scenario is possible because HLS implies, per construction for live content, a dynamic playlist updated at every new segment made available on the origin. The video representations can have classical out-of-band signaling using avc1 MP4 brand for AVC, hvc1 MP4 brand for HEVC, vvc1 MP4 brand for VVC in each segment or an inband signaling of resolution: avc3 MP4 brand for AVC, hev1 MP4 brand for HEVC, vvi1 MP4 brand for VVC in the MP4 container as described for DASH delivery. The latter option allows to have common segments for both DASH and HLS delivery.

10.1.3.2. Broadcast Delivery

DRE can also be applied to broadcast delivery, where the use of a lower resolution on temporally complex scenes can result in bandwidth and CPU savings. The traditional way of adapting the encoding scheme to the video content characteristics for the broadcast delivery was to share the bandwidth of a transponder among multiple video channels and to allocate a bitrate to each channel in function of the content characteristics in a very dynamic way, using a statistical multiplexing engine.

With a segment-based IP broadcast as specified by ATSC 3.0, with a DASH segment being serialized using ROUTE protocol, the dynamic allocation of bitrate does not exist anymore and the bandwidth for the channel may be constrained much more than what could be possible in a shared transponder. Therefore, DRE can be of high interest to preserve QoE in this constrained environment as well as to reduce the transponder cost.

For this use case, as for DASH OTT delivery, the segments being pushed in the route encapsulator should have inband signaling to ensure that the decoding engine on the ATSC 3.0 receivers will properly adapt to the moving resolutions on the transmitted streams. Even for a traditional TS-based broadcast making use of statistical multiplexing, DRE can have a value by smoothing the peak bitrate requests since the usage of a lower resolution for the most complex scenes will result in lower bitrate needs for the same QoE. It will provide a better QoE for congestion cases where all channels are complex at the same time. This can happen even more when the number of channels is small in the transponder, which is more frequent when broadcasting UHD channels. Therefore, DRE can favor the development of UHD channels.

10.1.4. Video Quality Evaluation Results

At the 2019 NAB and IBC Shows, Harmonic demonstrated a better QoE by using a higher resolution on static scenes at low bitrates.

For this paper, we focused our video quality evaluation on the use of a lower resolution to save bandwidth. The evaluation has two goals: the first goal is to confirm that there is no perception of resolution change within a scene, and the second goal is to measure the bandwidth savings. The subjective evaluations were performed by three Harmonic experts.

We considered two use cases with different codecs and maximum resolution:

1. a 1080p59.94 AVC delivery @ 4Mbps, with 1080p, 720p and 540p resolutions,
2. a 4K 59.94 HEVC delivery @ 6Mbps, with 2160p, 1440p, 1080p and 720p resolutions.

10.1.4.1. HD AVC Use Case

In this use case, the content is delivered at a constant and challenging bitrate of 4Mbps and can make use of one of the three following resolutions: 1920x1080p, 1280x720p or 960x540p at 59.94 fps, for each segment set to two seconds. We took five 4K HDR PQ BT.2020 @ 59.94fps content and one 8K HDR PQ BT.2020 @ 59.94 fps content produced by Harmonic and converted them to lower resolutions using a Lanczos filter and to SDR BT.709 using a proprietary HDR-to-SDR converter. The duration of content varies from 20 seconds to one minute. We used Harmonic's AI resolution selection and traditional encoding engine to produce the streams.

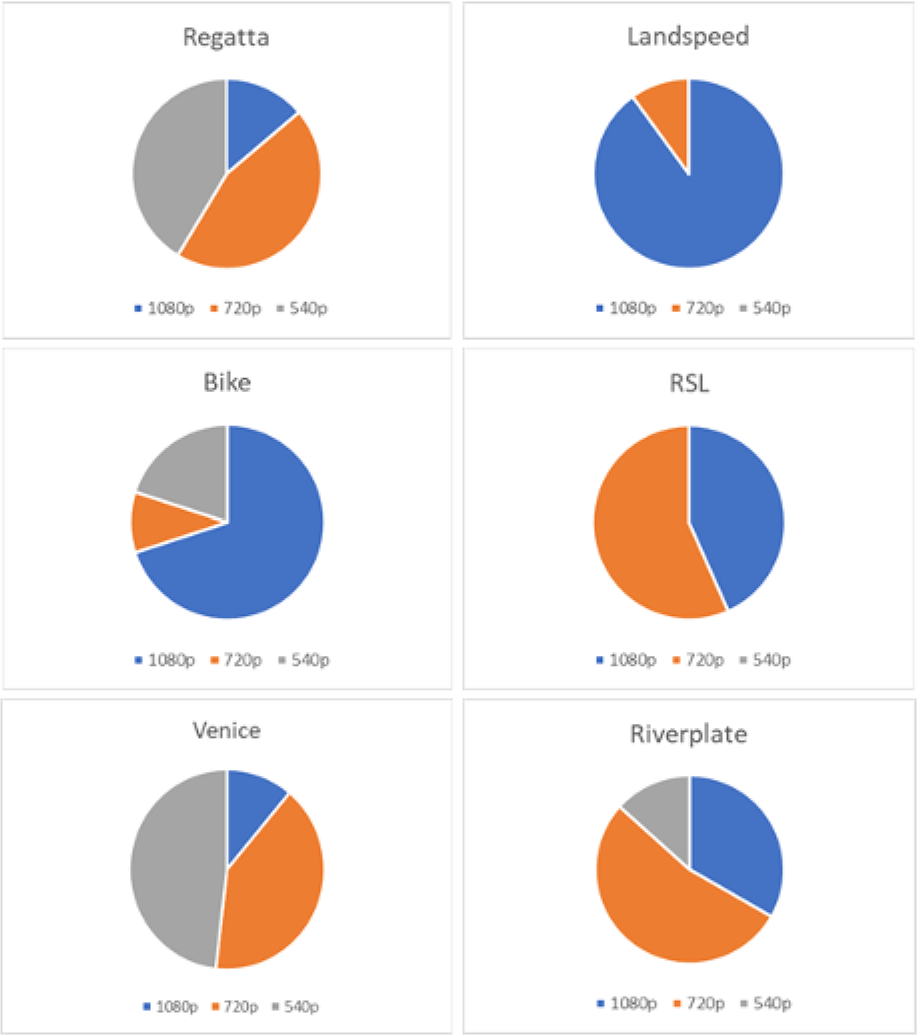


Figure 38. Shows the shares of resolution choices:

We can observe that there is a good variety of resolution choices at this bitrate. The resolution may change within a scene when the video characteristics change (typical case of a camera pan). The subjective evaluation of each individual DRE stream shows that there is no perception of resolution changes within a scene. We built off-line split screens comparing the stream encoded at a constant 1920x1080p resolution with the DRE stream up to 1920x1080p, using FFMPEG cropping and re-encoding at a high bitrate to preserve initial qualities, as shown in Figure 39. A DRE gain is perceived on complex scenes and on grass textures of River Plate and RSL sequences. No bad choices of resolution (no significant loss of details or sharpness) are observed.

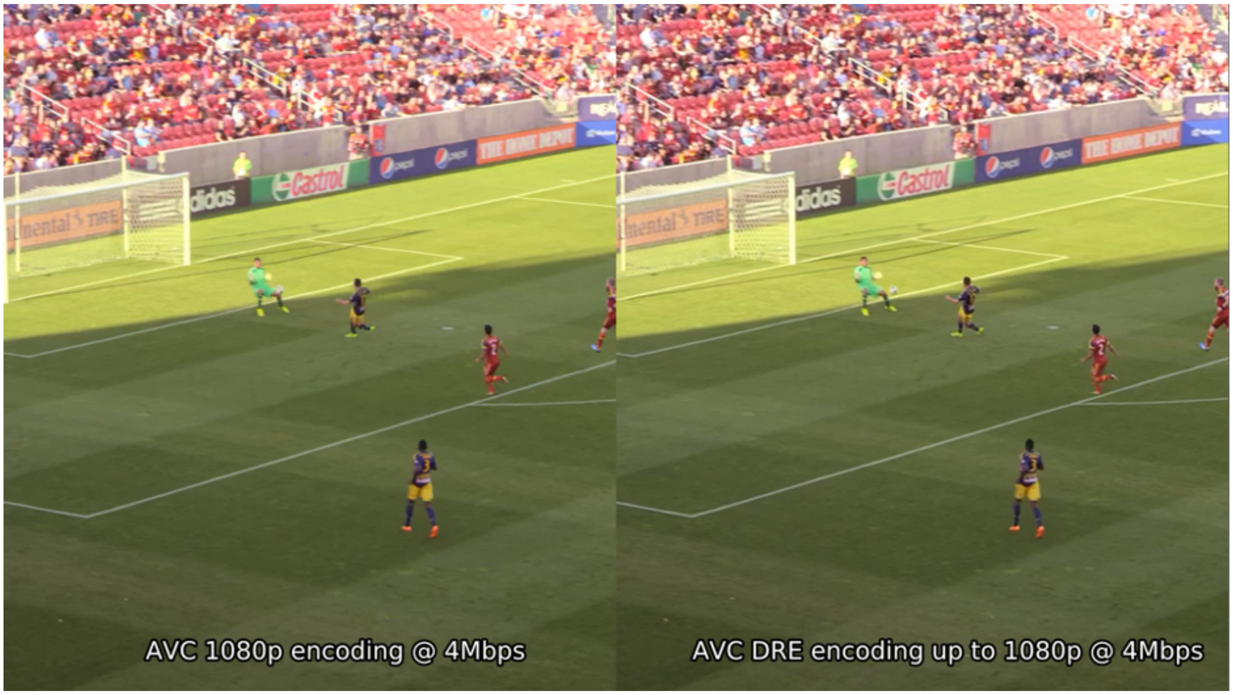

AVC 1080p encoding @ 4Mbps
AVC DRE encoding up to 1080p @ 4Mbps

Figure 39. 1080p/DRE split screen

To determine what the DRE bandwidth savings can be, we made the same split screens comparing the @ 4 Mb/s DRE stream with the constant resolution 1080p stream @ 5Mbps (20% savings) and 6 Mbps (33% savings). The subjective evaluation showed that 20% bitrate savings can be achieved on River Plate, close to 20% on RSL and close to 33% on Regatta.

For each DRE stream, Table 7 shows the average VMAF score, the best VMAF gain of DRE on a segment compared with 1080p encoding and the associated VMAF score for the segment.

Table 7. VMAF measurements of DRE streams

Content	Average VMAF	Best DRE VMAF gain	Associated VMAF
Regatta	75.73	+13.06	62.84
Landspeed	92.62	+0.88	93.27
Bike	91.21	+6.37	85.18
RSL	88.8	+1.45	91.01
Venice	87.36	+6.07	83.3
Riverplate	89.48	+10.16	83.31

The VMAF scores evaluation shows that up to +13 points can be achieved using a lower resolution (on Regatta) and the best gains are obtained on the segments where the VMAF scores are much lower than the average VMAF score on the sequence, which proves that DRE is important to preserve the quality on the most complex scenes.

10.1.4.2. Ultra HD HEVC Use Case

In this use case, the content is delivered at a constant and challenging bitrate of 6 Mbps and can make use of one of the four following resolutions: 3840x2160p, 2560x1440p, 1920x1080p or 1280x720p at 59.94 fps, for each segment set to two seconds. We took the same five 4K HDR PQ BT.2020 @ 59.94fps content and one 8K HDR PQ BT.2020 @ 59.94 fps content as in the previous use case. They were converted to lower resolutions using a Lanczos filter and to SDR BT.709 using a proprietary HDR-to-SDR converter.

We used Harmonic’s AI resolution selection and traditional encoding engine to produce the streams.

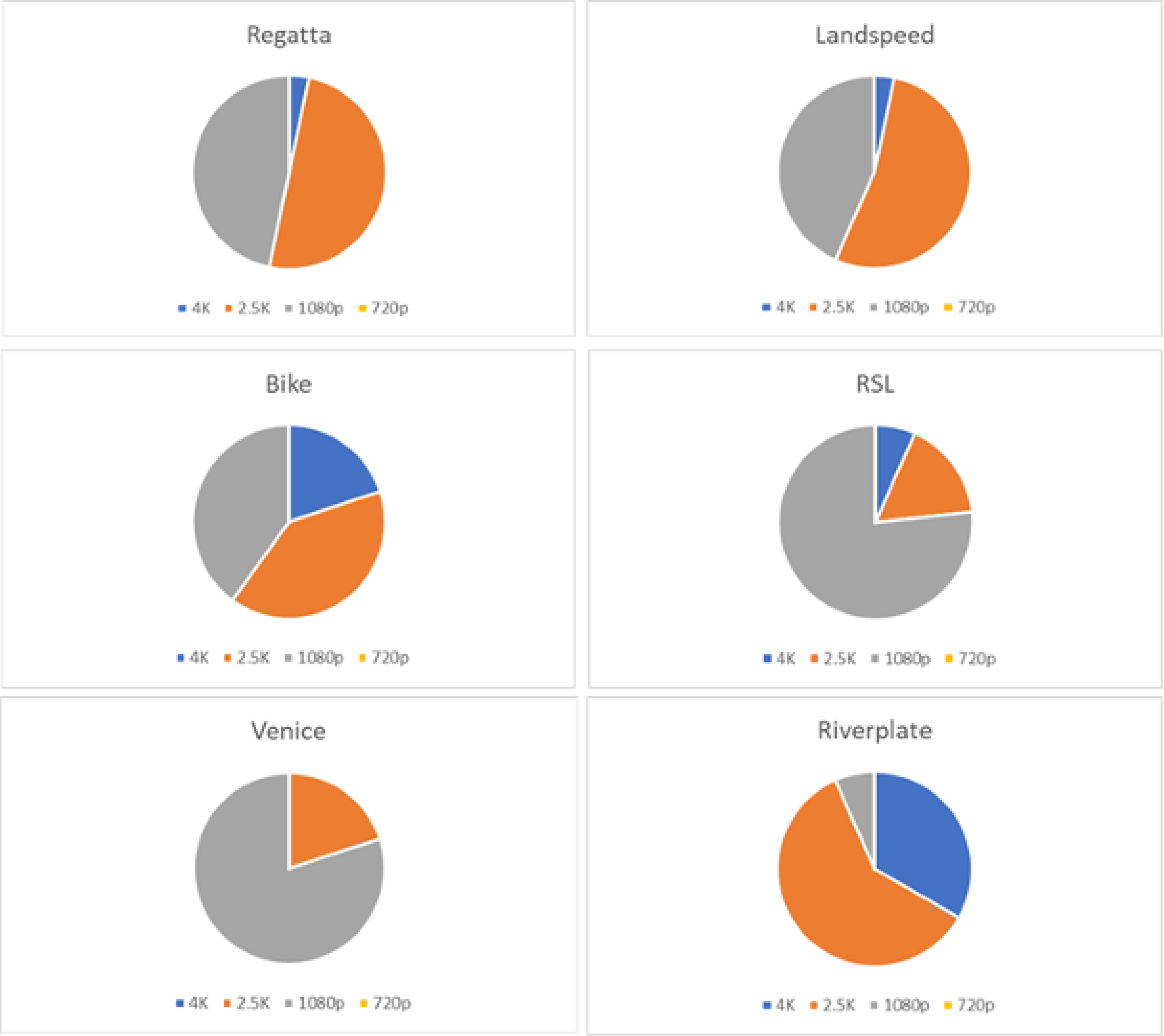


Figure 40. HEVC resolution selection shares per content

We can observe that there is, once again, a good variety of resolution choices at this bitrate. The resolution may change within a scene when the video characteristics change (typical case of a camera pan). The subjective evaluation of each individual DRE stream shows that there is no perception of resolution changes within a scene. We built offline split screens comparing the stream encoded at a constant 4K resolution with the DRE stream up to 4K, using FFMPEG cropping and re-encoding at a high bitrate to preserve initial qualities, as shown in Figure 41. A

DRE gain is perceived on complex scenes and on grass textures of Riverplate and RSL sequences. No bad choices of resolution (no significant loss of details or sharpness) are observed.

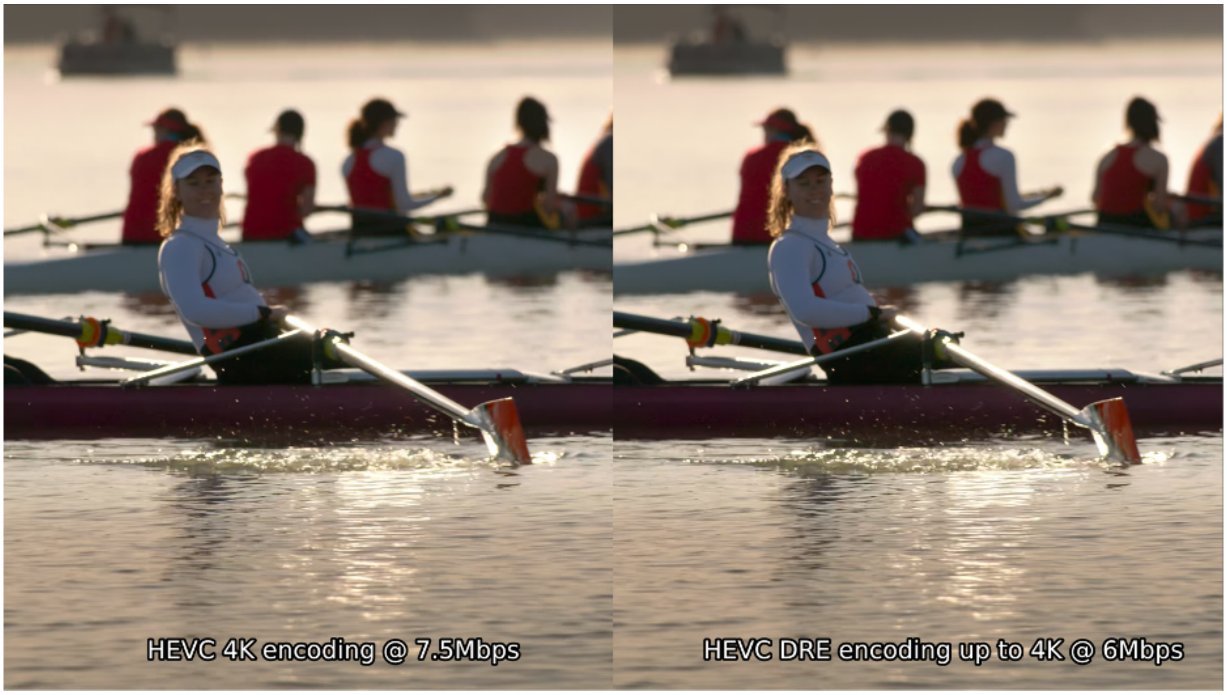

HEVC 4K encoding @ 7.5Mbps
HEVC DRE encoding up to 4K @ 6Mbps

Figure 41. 4K/DRE split screen

To determine what the DRE bandwidth savings can be, we made the same split screens comparing the @ 6 Mbps DRE stream with the constant 4K stream @ 7.5Mbps (20% savings) and 9 Mbps (33% savings). The subjective evaluation showed that 20% bitrate savings can be achieved on Venice and Regatta, 33% on Bike, Landspeed, RSL, RiverPlate.

For each DRE stream, Table 8 shows the average VMAF score, the best VMAF gain of DRE on a segment compared with 4K encoding and the associated VMAF score for the segment.

Table 8. VMAF measurements of DRE streams

Content	Average VMAF	Best DRE VMAF gain	Associated VMAF
Regatta	84.83	+6.62	82.37
Landspeed	90.22	+4.87	83.33
Bike	91.73	+5.32	89.95
RSL	90.65	+5.81	71.38
Venice	87.47	+6.46	85.22
Riverplate	95.62	+4.99	93.02

The VMAF scores evaluation shows that up to +6.5 points can be achieved using a lower resolution (on Regatta and Venice), and the best gains are obtained on the segments where the VMAF scores are much lower than the average VMAF score on the sequence, which proves that DRE is important to preserve the quality on the most complex scenes.

In addition to the bandwidth savings or QoE improvements, DRE also brings non-negligible CPU savings as the CPU cycles decrease with the encoded resolutions. On the HEVC test set, 20% CPU savings can be achieved compared with encoding at a constant 4K resolution.

10.1.5. Syntax requirements

10.1.5.1. DASH OTT Delivery

In each representation, the two parameters height and width usually indicate the resolution of the content for this profile. This is used by the player to check if it can handle this content with regards to its processing capabilities. If DRE had to become used at a larger scale, some addition in the MPEG DASH standard[24] or in the DASH-IF IOP[25] [4] may be included to mention more specifically that these height and width information correspond to the max possible resolution.

10.1.5.2. HLS Delivery

The HLS specification[26] describes the resolution element in the manifest by saying “The value is a decimal-resolution describing the optimal pixel resolution at which to display all the video in the Variant.” This is therefore not a description of the actual segment resolution and there is no need to change anything in the HLS specification.

10.1.5.3. Segment-based IP Broadcast Delivery (ATSC 3.0)

Nothing (more than in the OTT DASH case) may be needed as the payload is DASH segments.

10.1.5.4. DVB TS Broadcast Delivery

Resolution changes are authorized but no constraint is given on the seamless switch support for the Integrated Receiver Decoder (IRD).

10.1.6. Interop Test Results

Streams using a dynamic resolution have been tested with the following devices:

[24] ISO/IEC 23009-1:2019, Information Technology — Dynamic Adaptive Streaming Over HTTP (DASH) — Part 1: Media Presentation Description and Segment Formats (4th Edition), https://www.iso.org/standard/79329.html

[25] Guidelines for Implementation: DASH-IF Interoperability Points V4.3: On-Demand and Mixed Services, HDR Dynamic Metadata and other Improvements, https://dash-industry-forum.github.io/docs/DASH-IF-IOP-v4.3.pdf

[26] HTTP Live Streaming 2nd Edition, draft-pantos-hls-rfc8216bis-12. https://datatracker.ietf.org/doc/draft-pantos-hls-rfc8216bis/

- DASH and HLS players,
- ATSC 3.0-compliant 4K TVs (DASH ROUTE),
- DVB-T2 4K TVs.

10.1.6.1. OTT Streaming

DASH AVC DRE streams with inband signaling of resolution are well supported by the reference dash.js and Shaka players. DASH HEVC streams with inband signaling are well supported by the reference Shaka player in the edge browser.

HLS HEVC DRE streams with inband signaling when accessed through a playlist that refers to the right Init file each time the resolution changes are well supported by the iPad using the native player in Safari and the Shaka player in edge browser. Since the playlist refers to a new Init file each time the resolution changes, the fragmented MP4 file may not require inband signaling. At the time of writing this article, this test has not been done yet. HLS AVC streams have not been tested either.

10.1.6.2. DVB-T2 Broadcast Use Case

The HEVC DRE stream has been encapsulated into TS for DVB-T2 tests. Tests made with recent 4K TV sets of major brands (2020/2021 models), in collaboration with TDF in France, showed that the stream is well supported but with a few black frames (often below 100 ms) on many TVs when the resolution changes in the stream. Only one TV set was close to a seamless behavior with a one-frame freeze.

10.1.6.3. DASH ROUTE Broadcast Use Case

The DASH HEVC DRE stream has been played in loop to create a virtual live DASH feed pushed into an ATSC 3.0 DASH ROUTE server. The channel was combined with other live linear channels into a physical layer pipe (PLP) and then transmitted over the air. The DRE channel was received on RF inputs of three major brands of ATSC 3.0-compatible 4K TV sets. One of them was able to display the DRE channel perfectly well. The other two showed visible imperfections at the resolution changes. These artifacts do not come from the video decoder but from the rendering engine. Harmonic is working with these manufacturers on improving the behavior of the rendering engine to present a smooth, seamless up/downscale.

10.1.6.4. Interoperability Summary

This section summarizes the different results we have obtained by testing DRE. TV 1 to 4 represent various 4K TV brands.

Table 9. Interop Summary of Results

Use case	Dash.js	Shaka	iOS	TV1	TV2	TV3	TV4
ABR (DASH or HLS)	OK	OK	OK	NT	NT	NT	NT
TS broadcast	NA	NA	NA	Fail	Fail	Fail	Fail
Segment-based broadcast delivery	NA	NA	NA	OK	UI	UI	NA

Notes Table 9: DRE testing results.
NT: Not tested. UI: Under investigation

10.1.7. Discussion of Results and Outlook

The tests we made on two different use cases (HD AVC and 4K HEVC) show that DRE can push the boundaries of video compression by preserving the quality of complex scenes at lower bitrates thanks to the use of a lower resolution while keeping the sharpness of pictures with the highest possible resolution on static scenes with details.

At the same bandwidth, when going into challenging bitrates, the DRE offers better quality than encoding at the highest resolution as demonstrated by the split screens between constant resolution and DRE streams. The subjective assessments are confirmed by objective VMAF measures, which show a significant gain brought by the DRE for the most complex scenes.

In addition to bandwidth savings, the DRE can bring non-negligible CPU savings.

DRE can reinvent live OTT streaming by building content-aware dynamic profiles instead of constant profiles based on average statistics, as done currently. This will lead to bandwidth savings, better QoE, storage and CPU savings. This can be deployed with existing codecs like AVC and HEVC, since DASH or HLS players support per construction the resolution changes. Segment-based IP broadcast, such as ATSC 3.0 DASH ROUTE, should also be able to support it using the HEVC standard. This requires more interoperability work that will best fit under the CTA umbrella. In Brazil, the SBTVD Forum is currently considering DRE as a possible enhancement to the VVC-based TV 3.0 standard for the new broadcast system, which will be deployed in 2023 and onward. Considering the very limited bandwidth of the targeted terrestrial transponder and the impossibility of sharing bandwidth among multiple channels, as in the past, DRE seems to be a promising solution to trigger up to 4K delivery.

For more traditional broadcast delivery, such as DVB-T2 systems, the interoperability tests showed that resolution changes at a high frequency are not realistic for the HEVC deployments and DRE will target future VVC-based broadcast deployments.

Other broadcast networks such as IPTV, DTH and QAM could also be considered, but the variety of clients and the lack of DVB standardization for resolution change does not make us confident this path could be pursued.

(End of Yellow Book)