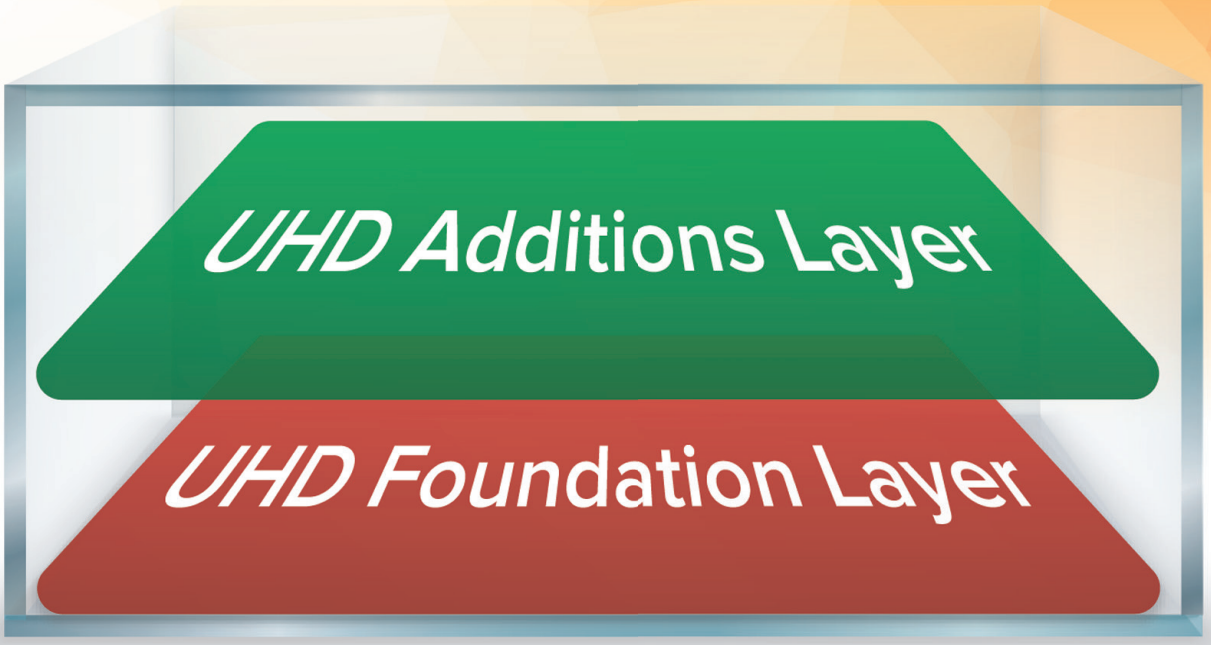
Ultra HD FORUM

GUIDELINES

Revision 2.1 - September 12, 2019



4K NGA WCG HDR HFR CAE

Ultra HD Forum Guidelines

September 12, 2019
Version: 2.1

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

Foreword

You are holding in your hands first update of the Ultra HD Forum Guidelines, unified into a single document. This book introduces the concepts of a foundation layer and an enhancement layer.

This work represents over four years of collaborative effort by all those who have contributed to the Guidelines Work Group. Our new guidelines would not have been possible without the leadership of Jim DeFilippis from Fraunhofer, chair of the Guidelines Work Group, who has managed to keep the group focused on the target, the “Guidelines” that I hope you will enjoy reading.

This new version 2.1 includes many updates including:

- DTS-UHD audio specification
- Another single HDR/SDR real world production description with the BBC 2019 FA Cup
- Improved HDR/WCG description in section 6.1 and an updated HFR Sec 13
- Clarification and vocabulary standardisation for conversion, colorimetry and mapping
- More external references and better alignment with standards bodies such as CTA

Key contributors to this update were Ian Nock, Ben Bodner, Bill Redmann, Chris Seeger, Pete Sellar, Andrew Cotton, Richard Doherty, Yurij Reznik and of course Jim DeFilippis.

If you want to know more about Ultra HD, and how it can be deployed, I invite you to join the Ultra HD Forum.

You can start by visiting our website: www.ultrahdforum.org.

Thierry Fautier, Ultra HD Forum President
Amsterdam, September 2019

Acknowledgements

This document is the result of many iterations of work by the members of the Ultra HD Forum Guidelines Working Group. We would like to thank these members who have worked hard at producing these guidelines and industry best practices.

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

Notice

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

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

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

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

Table of Contents

FOREWORD III
ACKNOWLEDGEMENTS V
NOTICE VII
1. PURPOSE AND SCOPE 1
2. REFERENCES 3
2.1 Reference List 3
2.2 Summary of ITU-R BT.709, BT.2020, & BT.2100 for linear broadcast 10
3. TERMS AND ACRONYMS 11
3.1 Terms 11
3.2 Acronyms and Abbreviations 14
4. PHASES AND TIMEFRAMES 17
4.1 Foundation UHD Technologies 17
4.2 Additional UHD Technologies 20
5. USE CASES 22
5.1 Digital Terrestrial Transmission 22
5.2 MVPD Platform Delivery 22
5.3 IP Network Delivery 23
6. PRODUCTION AND POST PRODUCTION 25
6.1 HDR/WCG Technologies 26
6.1.1 Perceptual Quantization (PQ) and PQ10 26
6.1.2 Hybrid Log-Gamma (HLG) and HLG10 27
6.1.3 Recommendation ITU-R BT.2100 28
6.1.4 Static Metadata – SMPTE ST 2086, MaxFALL, MaxCLL 29
6.1.5 HDR10 29
6.1.6 Foundation UHD HDR Technologies 30
6.1.7 HDR10 Metadata Generation 30
6.1.8 HDR10 Metadata Carriage 30
6.1.9 Signaling Transfer Function, System Colorimetry and Matrix Coefficients 31
6.1.10 Peak Brightness: Production, Ref. Monitors, Consumer Displays and Archives 35
6.1.11 Studio Video over IP 36
6.1.12 Adding Dynamic HDR Metadata to Foundation UHD 37
6.2 Production for Pre-recorded Content 38
6.2.1 Camera Requirements 39
6.2.2 Reference Monitor 40
6.2.3 On-Set / Near-Set Monitoring 41
6.2.4 Color Grading 41
6.2.5 Channel-based Immersive Audio Post Production 42
6.2.6 Additional UHD Technologies beyond Foundation UHD – NGA 43
6.3 Production for Live Content 43
6.3.1 Live Production in Trucks or Studio Galleries 44
6.3.2 Production with encodings other than PQ and HLG 44
6.3.3 Channel-based Immersive Audio Production 45
6.3.4 Additional UHD Technologies beyond Foundation UHD – NGA 47
7. SECURITY 48
7.1 Content Encryption 48
7.2 Forensic Watermarking 49
7.2.1 Introduction 49

7.2.2 Use Cases 49
7.2.3 Distribution 50
7.2.4 One-Step Watermarking 51
7.2.5 Two-Step Watermarking Integration 51
7.2.6 Use Case: ABR VOD 56
8. REAL-TIME PROGRAM SERVICE ASSEMBLY 58
8.1 Maintaining Dynamic Range and System Colorimetry Parameters 58
8.2 Conversion from SDR/BT.709 to PQ10/HLG10 58
8.3 Conversion between Transfer Functions 61
8.4 Conversion from PQ10/HLG10 to SDR/BT.709 61
8.5 Avoiding Image Retention on Professional and Consumer Displays 62
8.5.1 Background 62
8.5.2 Definition of Static Images 62
8.5.3 Recommendations 62
8.6 Additional UHD Technologies beyond Foundation UHD 63
9. DISTRIBUTION 64
9.1 Production Processing and Contribution 65
9.1.1 Video 66
9.1.2 Audio 68
9.1.3 Closed Captions and Subtitles 68
9.2 Broadcast Center Processing and Primary Distribution 68
9.3 Final Distribution from MVPD/OTT/DTT Provider Processing 71
9.3.1 Bit Depths 71
9.3.2 Video 71
9.3.3 Adaptive Bitrate (ABR) Streaming 73
9.3.4 Audio 74
9.3.5 Closed Captions and Subtitles 74
9.3.6 Considerations for UHD Technologies beyond Foundation UHD 74
9.4 Transport 75
10. DECODING AND RENDERING 76
10.1 Decoding 76
10.2 Rendering 76
10.3 Overlays Inserted at the Consumer Device 77
10.4 Considerations for UHD Technologies beyond Foundation UHD 77
11. FORMAT INTEROPERABILITY 80
11.1 Legacy Display Devices 81
11.2 Down-conversion at the Service Provider 81
11.3 Down-conversion at the STB 82
11.4 Spatial Resolution Up-conversion of Legacy Services 83
11.5 Interoperability of Atmos Immersive Audio 85
11.6 Considerations for UHD Technologies beyond Foundation UHD 85
12. HIGH DYNAMIC RANGE 86
12.1 Dolby Vision 86
12.1.1 Dolby Vision Encoding/Decoding Overview 86
12.1.2 Dolby Vision Cross Compatibility 87
12.1.3 Dolby Vision Color Volume Mapping (Display Management) 88
12.1.4 Dolby Vision in Broadcast 88
12.2 Dual Layer HDR 91
12.3 SL-HDR1 94

12.4 SL-HDR2 103
13. HIGH FRAME RATE 113
13.1 Introduction 113
13.2 HFR Video Format Parameters 114
13.3 Backward Compatibility for HFR 114
13.4 Production Considerations for HFR 116
14. NEXT GENERATION AUDIO 117
14.1 Common Features of NGA 117
14.1.1 NGA Use Cases 118
14.1.2 Audio Program Components and Preselections 118
14.1.3 Carriage of NGA 119
14.1.4 Metadata 119
14.1.5 Overview of Immersive Program Metadata and Rendering 119
14.1.6 Audio Element Formats 120
14.1.7 Audio Rendering 121
14.2 MPEG-H Audio 122
14.2.1 Introduction 122
14.2.2 MPEG-H Audio Metadata 125
14.2.3 MPEG-H Audio Stream 129
14.3 Dolby AC-4 Audio 131
14.3.1 Dynamic Range Control (DRC) and Loudness 134
14.3.2 Hybrid Delivery 135
14.3.3 Backward Compatibility 135
14.3.4 Next Generation Audio Metadata and Rendering 135
14.3.5 Overview of Immersive Program Metadata and rendering 135
14.3.6 Overview of Personalized Program Metadata 139
14.3.7 Essential Metadata Required for Next-Generation Broadcast 140
14.3.8 Metadata Carriage 143
14.4 DTS-UHD Audio 146
14.4.1 Introduction 146
14.4.2 System Overview 146
14.4.3 DTS-UHD Bitstream 147
14.4.4 Metadata 148
14.4.5 Audio Chunks 151
14.4.6 Organization of Streams 151
14.4.7 Multi-Stream Playback 154
14.4.8 Rendering 156
14.4.9 Personalization 158
15. CONTENT AWARE ENCODING 159
15.1 Introduction 159
15.1.1 Adaptive Bitrate Usage for UHD 159
15.1.2 Per-title Encoding 159
15.1.3 VBR Encoding 160
15.2 Content Aware Encoding Overview 160
15.2.1 Principles 161
15.3 Content Aware Encoding applied to UHD 161
15.4 Content Aware Encoding interoperability 163
15.5 Application for Content Aware Encoding 163
15.5.1 Internet bandwidth 163
15.5.2 CAE Sweet Spot for UHD 164

15.6 Content Aware Encoding Benefits 165
15.6.1 CDN cost 165
15.6.2 Quality of experience 165
16. ANNEX A: REAL WORLD FOUNDATION UHD DEPLOYMENTS 167
16.1 CBS and DirecTV Major Golf Tournament 167
16.2 Amazon Major Parade 168
16.3 NBC Universal Olympics and 2018 World Cup 168
16.4 Sky/BBC Royal Wedding and Major Tennis Tournament 169
16.5 BBC 2019 Football Association Challenge Cup 171
17. ANNEX B: IC_TC_P COLOR REPRESENTATION 174
18. ANNEX C: ACES WORKFLOW FOR COLOR AND DYNAMIC RANGE 175
19. ANNEX D: ISO 23001-12, SAMPLE VARIANTS 177
20. ANNEX E: AVS2 179
20.1 Why AVS2 179
20.2 Deployment 179
20.3 Technology 180

Index of Tables and Figures

Table 1 Summary Comparison of ITU-R BT.709, BT.2020, and BT.2100 10
Table 2 Foundation UHD Workflow Parameters 18
Table 3 Foundation UHD Content Parameters 18
Table 4 Foundation Decoder Capabilities 19
Table 5 Foundation Service Formats 20
Table 6 UHD over IP Networks 24
Table 7 File-Based Signaling for SDR/BT.709 32
Table 8 File-Based Signaling for SDR/BT.2020 33
Table 9 File-based Signaling for HDR/BT.2020 34
Table 10 Pre-recorded Content Interface Descriptions 39
Table 11 Compression and Distribution Nodes and Interfaces 64
Table 12 SDI Input Standards for 2160p Content 65
Table 13 Contribution Bitrates and Key Parameters 67
Table 14 Primary Distribution Bitrates and Key Parameters 70
Table 15 Existing Practices for Real-Time Program Service Distribution Formats 71
Table 16 Final Distribution Bitrates and Key Parameters 72
Table 17 Example “Real World” Bitrates as early as 2016 72
Table 18 Example Bitrates for Video Streams 74
Table 19 2K high frame rate content parameters 114
Table 20 Mapping of terminology between NGA technologies 121
Table 21 Levels for the Low Complexity Profile of MPEG-H Audio 123
Table 22 DE modes and metadata bitrates 133
Table 23 Common target reference loudness for different devices 134
Table 24 Common DRC curves 150
Table 25 CAE granularity 161
Table 26 Examples of fixed and CAE encoding ladders for live sports 162
Table 27 ACES Workflow Model 175

Figure 1 Content Production and Distribution Workflow 26
Figure 2 Pre-recorded Content Production Workflow and Interfaces 38
Figure 3 Channel-based Immersive Audio Post-Production 42
Figure 4 Channel-based Immersive Audio Live Production 46
Figure 5 Illustration of Watermark Identifier 49
Figure 6 One-Step Watermark Performed on the Client Side 51
Figure 7 Two Examples of Two-step Watermarking Systems 52
Figure 8 Watermark Embedding Using a Unique Variant Sequence 53
Figure 9 Transport at the Media Layer Using MPEG SEI NALUs 54
Figure 10 Transport at the Container Layer Using a Track in an ISOBMFF File 55
Figure 11 Pre-processing in the Baseband Domain for Two-step Watermarking 56
Figure 12 ABR Playlist Serialization 57
Figure 13 Sample Live Workflow with Mixed Format Source Content 60
Figure 14 Distribution Nodes and Interfaces 64
Figure 15 Contribution Workflow 66
Figure 16 Primary Distribution to an MVPD or Station 69
Figure 17 Primary Distribution to an OTT Provider 70
Figure 18 Down-conversion at the Headend 82

Figure 19 Down-conversion at the STB 83
Figure 20 Spatial Resolution Up-conversion of Legacy Services 84
Figure 21 Encoder functional block diagram....86
Figure 22 Decoder function block diagram 87
Figure 23 Example display device color volumes 88
Figure 24 Example broadcast production facility components....89
Figure 25 HDR broadcast production facility with BT.2100 PQ workflow- transition phase 90
Figure 26 HDR broadcast production facility with BT.2100 PQ workflow- SDI metadata....91
Figure 27 Example dual-layer encoding and distribution....93
Figure 28 SL-HDR processing, distribution, reconstruction, and presentation 95
Figure 29 Direct reception of SL-HDR signal by an SL-HDR1 capable television....96
Figure 30 STB processing of SL-HDR signals for an HDR-capable television....97
Figure 31 STB passing SL-HDR to an SL-HDR1 capable television 98
Figure 32 Multiple SL-HDR channels received and composited in SDR by an STB 99
Figure 33 SL-HDR as a contribution feed to an HDR facility 101
Figure 34 SL-HDR as a contribution feed to an SDR facility 102
Figure 35 SL-HDR2 processing, distribution, reconstruction, for HDR presentation....104
Figure 36 SL-HDR2 processing, distribution, reconstruction, and SDR presentation 105
Figure 37 Direct reception of SL-HDR signal by an SL-HDR2 capable television....106
Figure 38 STB processing of SL-HDR signals for an HDR-capable television....107
Figure 39 STB passing SL-HDR to an SL-HDR2 capable television 108
Figure 40 Multiple SL-HDR channels received and composited in HDR by an STB 109
Figure 41 SL-HDR as a contribution feed to an HDR facility 111
Figure 42 SL-HDR as a contribution feed to an SDR facility 112
Figure 43 Bandwidth increases for various video format improvements 113
Figure 44 ATSC 3.0 temporal filtering for HFR backward compatibility 116
Figure 45 NGA in the consumer domain....117
Figure 46 Relationship of key audio terms 121
Figure 47 MPEG-H Audio system overview....123
Figure 48 MPEG-H Authoring Tool example session....124
Figure 49 Distributed UI processing with transmission of user commands over HDMI 125
Figure 50 Example of an MPEG-H Audio Scene information 127
Figure 51 Audio description re-positioning example 128
Figure 52 Loudness compensation after user interaction 129
Figure 53 MHAS packet structure 129
Figure 54 Example of a configuration change from 7.1+4H to 2.0 in the MHAS stream....131
Figure 55 Example of a configuration change from 7.1+4H to 2.0 at the system output....131
Figure 56 AC-4 Audio system chain 132
Figure 57 AC-4 DRC generation and application....134
Figure 58 Object-based audio renderer....136
Figure 59 Common panning algorithms 137
Figure 60 Serialized EMDF Frame formatted as per SMPTE ST 337 [36]....144
Figure 61 DTS-UHD System Overview 147
Figure 62 DTS-UHD Audio Frame Structure Example 148
Figure 63 Default Playback 152
Figure 64 Specific Object and Group Selection 153
Figure 65 Playback using Default settings....153
Figure 66 Example of Selecting Playback of Audio Presentation 2 154
Figure 67 Example of Selecting Desired Objects to Play Within a Single Stream 154
Figure 68 Example of multi-stream decoding 156

Figure 69: Point Source Object Renderer Coordinate System....157
Figure 70: 7.x Output Configuration with Predefined Virtual Speakers....158
Figure 71: Object Interactivity Manager....158
Figure 72 CAE encoding chart....163
Figure 73 Internet speed distribution per countries (source Akamai)....164
Figure 74 CAE sweet spot vs. CBR....165
Figure 75 Junction bitrates chart....166
Figure 76 CBS and DirecTV Major Golf Tournament Workflow....167
Figure 77 Amazon Major Parade Workflow....168
Figure 78 NBCU Olympics and 2018 World Cup UHD Workflow....169
Figure 79 Sky/BBC Royal Wedding and Major Tennis Tournament Workflow....170
Figure 80 BBC 2019 Football Association Challenge Cup Workflow....171
Figure 81 ACES Workflow Model....175
Figure 82 Transport at the Container Layer Using a Track in an ISOBMFF File....177
Figure 83 AVS2 coding framework....180

1. Purpose and Scope

The purpose of this document is to describe consistent methods for the creation and delivery of Ultra HD content for consumer distribution along with a uniform set of characteristics that may be combined to produce content that can be considered “Ultra HD” The scope includes delivery via the Internet, satellite, terrestrial broadcast and cable as transmission methods. It does not include encode and delivery of content via storage media, such as Blu-ray® disc, HDD, SCSA devices, or similar, nor does it include encode and delivery of Digital Cinema content.

The goal is to create consistency across the industry for ensuring interoperability and a high-quality experience for consumers. While this document provides context with respect to content creation, the primary purpose of this document is to define guidelines for proper delivery of UHD content from the studio producer or the live event to the consumer via a linear (real-time) service.

This document recommends profiles and practices to be used across each of the elements in a distribution chain to maximize end-to-end interoperability. References supporting the recommendations are provided to the extent possible. However, in many cases, industry practices are advancing more quickly than existing documentation. In the cases where technologies are in the process of being developed and/or standardized, these guidelines also provide associated time line expectations where possible. All the recommendations represent the consensus view of the Ultra HD Forum based on these references, its members’ expertise and experience, and/or results from Ultra HD Forum Interop events.

The Ultra HD Forum intends the UHD Guidelines to be a "living document" and plans to release new revisions as more data becomes available, more learning is accumulated across deployments, and more interops and trials are conducted, while keeping the same scope for the document.

For the purpose of this document, the Ultra HD Forum is considering the following UHD content and service types, which have different workflow characteristics:

- Content Types:
  - Live content – content that is distributed to consumers in real-time as it is produced, such as sports, news, awards shows, reality programs, talent shows, debates, etc. Live production workflows do not include a post-production step and creative intent is set in the control room or truck. Note that content produced in this manner may also be captured for subsequent re-broadcast.
  - Pre-recorded content – content that is fully produced prior to distribution to consumers, such as sitcoms, dramas, advertisements, documentaries, etc. Pre-recorded production workflows include post-production steps, and creative intent is set during post-production.
- Service Types:
  - Real-time Program Services – services consisting of a linear, pre-scheduled stream of content that is assembled in real-time for distribution to consumers such as a broadcast television channel, cable network, etc. Real-time Program Services are comprised of Live and/or Pre-recorded content and may also include graphic overlays, such as station logos, emergency text crawls, etc.

- On-Demand Services – While services such as Hulu, Netflix and MVPD VOD are largely out of scope, if the content offered on these platforms was originally Live and recorded for later distribution then these Guidelines may be relevant. See below for further explanation.

The primary focus of the Ultra HD Forum is on Real-time Program Services because they may be the most challenging for an Ultra HD end-to-end workflow. On-Demand Services are largely out of scope. However, guidelines contained herein related to producing Live content may be valuable to On-Demand service providers who are offering content that was originally produced for Live distribution using a linear workflow, and has been repurposed as a VOD asset, e.g., via caching at the point of final distribution, for start-over, catch-up, or trick play. With this in mind, the scope of this document is defined as follows:

In scope:

- Pre-recorded and Live content production
  - Cameras
  - Monitoring
  - Color grading for Pre-recorded content
  - HDR/WCG technologies
  - Channel-based Immersive Audio
- Metadata
- Security
- Distribution and Compression
  - Content master format
  - Content mezzanine format
  - Encoding codecs, methods and recommendations
  - Transcode codecs, methods and recommendations
  - Approximate ranges of bitrates through the entire processing chain
  - Distribution and transport methods
- Real-time program stream assembly
- Conversion between SDR and HDR formats and among different HDR formats
- Interface guidelines for connecting systems and functions throughout the production and delivery of the content
- Backward compatibility for legacy systems

Out of scope:

- Filming techniques (e.g., lighting, camera settings, etc.)
- TV settings
- Encoder settings
- Subjective analysis of overall content quality
- TV technology guidelines (e.g., OLED vs. Quantum Dots)
- Color grading guidelines (e.g., luma, saturation and contrast preferences)
- Fixed media delivery and digital cinema

2. References

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

2.1 Reference List

[1] Recommendation ITU-T H.222.0 | ISO/IEC 13818-1:2000, "Information Technology—Generic coding of moving pictures and associated audio information - Part 1: Systems"

[2] Recommendation ITU-R BT.709, "Parameter values for the HDTV standards for production and international programme exchange"

[3] Recommendation ITU-R BT.2020, "Parameter values for ultra-high definition television systems for production and international programme exchange"

[4] Recommendation ITU-R BT.1886, "Reference electro-optical transfer function for flat panel displays used in HDTV studio production"

[5] Recommendation ITU-R BT.2100, "Image parameter values for high dynamic range television for use in production and international programme exchange", June 2017, http://www.itu.int/rec/R-REC-BT.2100

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

[7] Recommendation ITU-R BT.2087-0, "Colour conversion from ITU-R BT.709 [2] to ITU-RBT.2020 [3]"

[8] Recommendation ITU-R BT.2408, "Operational Practices in HDR Television Production"

[9] SMPTE ST 2084, "High Dynamic Range Electro-Optical Transfer Function of Mastering Reference Displays"

[10] SMPTE ST 2086, "Mastering Display Color Volume Metadata Supporting High Luminance and Wide Color Gamut Images"

[11] SMPTE HDR Study Group report, https://www.smpte.org/sites/default/files/Study%20Group%20On%20High-Dynamic-Range-HDR-Ecosystem.pdf

[12] DVB UHD-1 Phase 1 specification, http://www.etsi.org/deliver/etsi_ts/101100_101199/101154/02.01.01_60/ts_101154v020101p.pdf

[13] DVB DASH specification, https://www.dvb.org/resources/public/standards/a168_dvb-dash.pdf

[14] Recommendation ITU-R BT.814, "Specifications and alignment procedures for setting of brightness and contrast on displays"

[15] HDMI 2.0a specification, http://www.hdmi.org/manufacturer/hdmi_2_0/hdmi_2_0a_faq.aspx

[16] DASH-IF Interoperability Points: Guidelines for Implementation, version 4.3

[17] DCI Specification, Version 1.2 with Errata as of 30 August 2012 Incorporated

[18] CTA-608-E R-2014, "Line 21 Data Services", http://www.cta.tech/Standards/Standard-Listings/R4-3-Television-Data-Systems-Subcommittee.aspx

[19] CTA-708-E (ANSI), "Digital Television (DTV) Closed Captioning" , http://www.cta.tech/Standards/Standard-Listings/R4-3-Television-Data-Systems-Subcommittee.aspx

[20] ETSI 300 743, "Digital Video Broadcasting (DVB); Subtitling systems", http://www.etsi.org/deliver/etsi_en/300700_300799/300743/01.03.01_60/en_300743v010301p.pdf

[21] ETSI 300 472, "Digital Video Broadcasting (DVB); Specification for conveying ITU-R System B Teletext in DVB bitstreams", http://www.etsi.org/deliver/etsi_en/300400_300499/300472/01.03.01_60/en_300472v010301p.pdf

[22] SCTE-27, "Subtitling Methods for Broadcast Cable" http://www.scte.org/documents/pdf/standards/SCTE_27_2011.pdf

[23] W3C: "TTML Text and Image Profiles for Internet Media Subtitles and Captions (IMSC1)", [Candidate] Recommendation, W3C, www.w3.org.

[24] ATSC: "Techniques for Establishing and Maintaining Audio Loudness for Digital Television," Doc. A/85, Advanced Television Systems Committee, Washington, D.C., 12 March 2013, http://atsc.org/recommended-practice/a85-techniques-for-establishing-and-maintaining-audio-loudness-for-digital-television/

[25] ISO/IEC: Doc. ISO/IEC 14496-10 "MPEG-4 -- Part 10: Advanced Video Coding"[1]

[26] ISO/IEC: Doc. ISO/IEC 23008-2:2015 "Information technology -- High efficiency coding and media delivery in heterogeneous environments -- Part 2: High efficiency video coding"[2]

[27] ISO/IEC 14496-3:2009, "Informational Technology, Coding of Audio-Visual objects, Part 3: Audio"

[28] ISO/IEC 14496-12:2015, "Information technology—Coding of audio-visual objects—Part 12: ISO base media file format"

[29] ETSI 102 366 v1.3.1 (2014-08), "Digital Audio Compression (AC-3, Enhanced AC-3) Standard"

[30] SMPTE RP 431-2, "D-Cinema Quality — Reference Projector and Environment"

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

[32] W3C: "HTML5, A vocabulary and associated APIs for HTML and XHTML" https://www.w3.org/TR/html5/

[33] ETSI TS 103 433-1 v1.2.1 (2016-08) "High-Performance Single Layer Directly Standard Dynamic Range (SDR) Compatible High Dynamic Range (HDR) System for use in Consumer Electronics devices (SL-HDR1)", http://www.etsi.org/deliver/etsi_ts/103400_103499/10343301/01.02.01_60/ts_10343301v010201p.pdf

[34] ETSI TS 103-433-2 v1.1.1 (2018-01) "High-Performance Single Layer High Dynamic Range (HDR) System for use in Consumer Electronics devices; Part 2: Enhancements for Perceptual Quantization (PQ) transfer function based High Dynamic Range (HDR) Systems (SL-HDR2)",

[1] Also published by ITU as ITU-T Recommendation H.264.

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

https://www.etsi.org/deliver/etsi_ts/103400_103499/10343302/01.01.01_60/ts_10343302v010101p.pdf
[35] ETSI TS 103 420, "Object-based audio coding for Enhanced AC-3 (E-AC-3)".
[36] SMPTE ST 337, "Format for Non-PCM Audio and Data in AES 3 Serial Digital Audio Interface"
[37] Recommendation ITU-R BS.1770-4, "Algorithms to measure audio programme loudness and true-peak audio level"
[38] SMPTE ST 2067-21, "Interoperable Master Format — Application #2E," 2016
[39] MovieLabs, "MovieLabs Specification for Enhanced Content Protection, v1.1"
[40] ISO/IEC 23001-12:2015, "Information technology -- MPEG systems technologies - Part 12: Sample Variants in the ISO base media file format"
[41] ETSI TS 102 034 v1.4.1 (2009-08), "Transport of MPEG-2 TS Based DVB Services over IP Based Networks"
[42] Internet Engineering Task Force (IETF) RFC 3550, "RTP: A Transport Protocol for Real-Time Applications", https://www.ietf.org/rfc/rfc3550.txt
[43] SMPTE ST 2110-10:2017, "Professional Media over IP Networks: System Timing and Definitions"
[44] SMPTE ST 2110-20:2017, "Professional Media over IP Networks: Uncompressed Active Video"
[45] SMPTE ST 2110-21:2017, "Professional Media over IP Networks: Traffic Shaping and Delivery Timing for Video"
[46] SMPTE ST 2110-30:2017, "Professional Media over IP Networks: PCM Digital Audio"
[47] SMPTE ST 2110-40:2018, "Professional Media over IP Networks: SMPTE ST 291-1 Ancillary Data"
[48] SMPTE ST 2108-1:2018, "HDR/WCG Metadata Packing and Signaling in the Vertical Ancillary Data Space"
[49] OpenCable Specifications, OC-SP-EP-I01-130118: "Encoder Boundary Point Specification" (01/2013). https://apps.cablelabs.com/specification/encoder-boundary-point-specification/
[50] SMPTE ST 2065-1:2012, "Academy Color Encoding Specification (ACES)"
[51] ATSC: A/300:2017, "ATSC 3.0 System", October 19, 2017, https://www.atsc.org/atsc-30-standard/a3002017-atsc-3-0-system/
[52] ATSC: A/322:2017, "Physical Layer Protocol", June 6, 2017, https://www.atsc.org/wp-content/uploads/2016/10/A322-2017a-Physical-Layer-Protocol.pdf
[53] ATSC: A/331:2017, ", "Signaling, Delivery, Synchronization, and Error Protection", December 6, 2017, https://www.atsc.org/wp-content/uploads/2017/12/A331-2017-Signaling-Deivery-Sync-FEC-3.pdf
[54] ATSC: A/341:2018, "Video-HEVC with Amendments No. 1 and No. 2", March 9, 2018, https://www.atsc.org/wp-content/uploads/2017/05/A341-2018-Video-HEVC-1.pdf
[55] ATSC: A/342-1:2017, "Audio Common Elements", January 24, 2017, https://www.atsc.org/wp-content/uploads/2017/01/A342-1-2017-Audio-Part-1-5.pdf
[56] ATSC: A/342-2:2017, "AC-4 System", February 23, 2017, https://www.atsc.org/wp-content/uploads/2017/02/A342-2-2017-AC-4-System-5.pdf
[57] ATSC: A/342-3:2017, "MPEG-H System", March 3, 2017, https://www.atsc.org/wp-content/uploads/2017/03/A342-3-2017-MPEG-H-System-2.pdf

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

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

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

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

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

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

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

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

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

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

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

[69] ISO/IEC: 23008-2, "Information technology -- High efficiency coding and media delivery in heterogeneous environments -- Part 2: High efficiency video coding", May 2015, https://www.iso.org/standard/67660.html [3]

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

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

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

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

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

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

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

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

[77] SCTE: 135:2013, "Data-Over-Cable Service Interface Specification version 3.0", March 14, 2013, http://www.scte.org/documents/pdf/Standards/ANSI_SCTE_135-1 2013.pdf

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

[79] SMPTE ST 424:2012, "3 Gb/s Signal/Data Serial Interface", October 8, 2012, https://ieeexplore.ieee.org/document/7290519

[80] SMPTE ST 425-1:2017, "Source Image Format and Ancillary Data Mapping for the 3 Gb/s Serial Interface", November 1, 2017, https://ieeexplore.ieee.org/document/8113731;
ST 425-3:2015, "Image Format and Ancillary Data Mapping for the Dual Link 3Gb/s Serial Interface", June 21, 2015, http://ieeexplore.ieee.org/servlet/opac?punumber=7290046 ; and
ST 425-5, "Image Format and Ancillary Data Mapping for the Quad Link 3Gb/s Serial Interface", June 21, 2015, http://ieeexplore.ieee.org/servlet/opac?punumber=7291843

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

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

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

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

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

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

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

[88] EBU Tech Report 038, March 2017, "Subjective evaluation of HLG for HDR and SDR distribution" https://tech.ebu.ch/publications/tr038

[89] EBU Recommendation 129, November 2018, "Advice to Broadcasters on Avoiding Image Retention on TV Production Displays", https://tech.ebu.ch/docs/r/r129.pdf

[90] Dolby Vision Profiles and Levels, v1.3.1.1, April 22, 2019, https://www.dolby.com/us/en/technologies/dolby-vision/dolby-vision-profiles-levels_v1.3.1.1.pdf

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

[92] ETSI TS 101 154 v2.5.1 (2019-12), "Digital Video Broadcasting (DVB); Specification for the use of Video and Audio Coding in Broadcasting Application based on the MPEG-2 Transport Stream", https://www.etsi.org/deliver/etsi_ts/101100_101199/10154/02.05.01_60/ts_101154v020501p.pdf

[93] ANSI/SCTE 242-4 2018, Next Generation Audio Coding Constraints for Cable Systems: Part 4 – DTS-UHD Audio Coding Constraints, https://www.scte.org/SCTEDocs/Standards/ANSI_SCTE%20242-4%202018.pdf

[94] ANSI/SCTE 243-4 2018, Next Generation Audio Carriage Constraints for Cable Systems: Part 4 – DTS-UHD Audio Carriage Constraints, https://www.scte.org/SCTEDocs/Standards/ANSI_SCTE%20243-4%202018.pdf

[95] ITU report BT.2408-2 "Guidance for operational practices in HDR television production" (04/2019), https://www.itu.int/pub/R-REP-BT.2408

[96] ANSI/SCTE 135-1, -2, -3, -4, -5 2019, DOCSIS 3.0, https://www.scte.org/SCTEDocs/Standards/ANSI_SCTE%20135-01%202018.pdf

[97] ANSI/SCTE 220-1, -2, -3, -4, -5 2016 DOCSIS 3.1, https://www.scte.org/SCTEDocs/Standards/ANSI_SCTE%20220-1%202016.pdf

[98] ETSI EN 302 769 v1.1.1 (2010-04), "Digital Video Broadcasting (DVB); Frame structure channel coding and modulation for a second generation digital transmission system for cable systems (DVB-C2)", https://www.etsi.org/deliver/etsi_en/302700_302799/302769/01.01.01_60/en_302769v010101p.pdf

[99] SMPTE ST 292-1:2018 1.5 Gb/s Signal/Data Serial Interface

[100] R. W. G. Hunt, "Light and Dark Adaptation and the Perception of Color*," J. Opt. Soc. Am. **42**, 190-199 (1952), https://www.osapublishing.org/josa/abstract.cfm?uri=josa-42-3-190

[101] J.C. Stevens and S.S. Stevens, "Brightness Function: Effects of Adaptation," J. Opt. Soc. Am. 53, 375-385 (1963), https://doi.org/10.1364/JOSA.53.000375

[102] J. Kautz, H. Kim, T. Weyrich, "Modeling Perception under Extended Luminance Levels", ACM TOG, Vol 28 Issue 3, August 2009, https://dl.acm.org/citation.cfm?id=1531333

[103] T. Borer, "Display of High Dynamic Range Images Under Varying Viewing Conditions," Proc. SPIE 10396, Applications of Digital Image Processing XI, 103960H,

September 2017, https://www.spiedigitallibrary.org/conference-proceedings-of-spie/10396/2274253/Display-of-high-dynamic-range-images-under-varying-viewing-conditions/10.1117/12.2274253.short

[104] CTA 861.4, "Updates to Dynamic HDR Metadata Signaling", March 2019, https://members.cta.tech/ctaPublicationDetails/?id=83193112-294f-e911-867a-0003ff528858

2.2 Summary of ITU-R BT.709, BT.2020, & BT.2100 for linear broadcast, production and distribution

The Ultra HD Forum Guidelines document refers to ITU-R BT.709 [2], BT.2020 [3], and BT.2100 [5] that address transfer function, system colorimetry, matrix coefficients, and more. The following table is a summary comparison of those three documents. Please note that this is merely a summary, and the reference documents contain considerably more information.

Table 1 Summary Comparison of ITU-R BT.709, BT.2020, and BT.2100

	ITU-R BT.709	ITU-R BT.2020	ITU-R BT.2100
Spatial Resolution	2K	4K, 8K	2K, 4K, 8K
Framerates*	24, 25, 30, 50, 60	24, 25, 30, 50, 60, 100, 120	24, 25, 30, 50, 60, 100, 120
Interlace/Progressive	Interlace, Progressive	Progressive	Progressive
System Colorimetry**	BT.709	BT.2020	BT.2100
Dynamic Range	SDR (BT.1886 [4])	SDR (BT.1886 [4])	HDR (PQ and HLG, BT.2100 [5])
Bit Depth	8, 10	10, 12	10, 12
Signal Format	RGB, YC_BC_R	RGB, YC_BC_R	RGB, YC_BC_R, IC_TC_P

Table 1 notes:

*Framerates include both integer and fractional values (including 120/1.001 for BT.2020 and BT.2100).

**The color primaries and gamut of BT.2020 and BT.2100 are identical. Refer to the ITU-R documents for actual color primary values. In this Guidelines document, when referring to color BT.2020 and BT.2100 mean the same primaries and gamut.

3. Terms and Acronyms

3.1 Terms

This guideline contains the following terms and definitions:

Access Unit (AU)	Self-contained audio stream packet.
Adaptive Bit Rate	A technique used in streaming multimedia over computer networks, in which multiple versions of a single content source are provided, each encoded at different bitrates; the client device monitors the available bandwidth and CPU capacity in real time, and switches between streaming the different encodings, choosing the highest bitrate (i.e., highest quality) according to available resources.
Audio Objects	An audio element that consists of an audio signal and audio metadata, which includes rendering information (e.g., gain and position) that may dynamically change. Audio Objects with positional information that does not dynamically change are referred to as “static” objects.
Binaural Audio	Process that reproduces audio for headphones, including immersive audio.
Bit Depth	The number of bits used per component. It describes the number of increments for both brightness and color.
Color Gamut	The subset of colors that can be accurately represented within a given system colorimetry, or by a certain source or output device.
Color Volume	Combined color gamut and luminance characteristics.
Color Volume Transform	A technique used to map a coordinate in one color volume to a coordinate in another color volume.
Commentary	Audio program element assigned to voice/announcer information
Convergence/ Divergence	For audio object, the amount of the ‘spread’ of the audio in acoustic space
Core Decode	Minimal decode specification, usually limited to stereo or 5.1 audio programs.
DCI-P3	Color gamut defined in SMPTE RP 431-2 [30].
Dialog Enhancement	Feature for the hearing challenged viewer or where there is high ambient noise to enhance the intelligibility of the dialog or commentary audio or for the preference of the viewer.
Downmixing	For Channel-based audio formats, the ability for the decoder to reproduce the higher order speaker channel arrangement to a lesser speaker channel arrangement (i.e. 5.1 to 2.0).
Electro-Optical Transfer Function	The transfer function that maps digital pixel values to values of display light.

Forensic Watermarking	Forensic Watermarking is a technology that modifies multimedia content (e.g., a video, a song, a piece of text) to encode a Watermark Identifier without introducing artifacts that would be perceptible by a human being. The Watermark Identifier encoded by a Forensic Watermark can be recovered even if the content is altered after the watermarking operation.
Foundation UHD	Term used in this document to for content that conforms to the parameters shown in Table 3.
Full Decode	Decode specification that provides for full immersive or higher spatial resolution sound program reproduction.
HLG10	The Hybrid Log-Gamma OETF described in BT.2100 [5] together with BT.2020 [3] system colorimtry and 10-bit depth.[4] [see also Section 6.1.2]
HDR10	A PQ10-based format further capable of providing SMPTE ST 2086, MaxFALL, and MaxCLL metadata (see also Section 6.1.5).
High Dynamic Range	Greater than or equal to the contrast ratio that could be derived from 13 f-stops.
High Frame Rate	Content with a relative rate greater than 24 frames per second for motion pictures and greater than 60 fps for television content.
Hybrid Log-Gamma	Hybrid Log-Gamma OETF, EOTF, and OOTF as defined in BT.2100 [5].
Immersive Audio	An audio system that enables high spatial resolution in sound source localization in azimuth, elevation and distance, and provides an increased sense of sound envelopment.
Inverse Tone Mapping	Process to convert SDR/709 video to HDR/2020 video. Also refereed to 'up-mapping'
ISO Base Media File Format	File format for media as defined by ISO/IEC 14496-12 [68]
Loudness Normalization	Process within the audio codec that ensures consistent audio loudness across all renders, downmixes and preselections.
MaxCLL	Maximum Content Light Level – Represents the brightest pixel in the entire video stream (CTA 861.G [31])
MaxFALL	Maximum Frame-Average Light Level – Represents the maximum frame average pixel light value per frame of the entire video stream (CTA 861.G [31]).
Modulation Transfer Function	The contrast performance of an optical system such as a lens as a function of spatial frequency.
Multichannel Video Programming Distributor	A service provider that delivers video programming services, usually for a subscription fee (pay television).
Next Generation Audio	Immersive sound with dynamic and static objects, interactive and personalized audio delivery system with improved audio compression quality. NGA supports three fundamental audio element formats: Channel Sets, Audio Objects (static and/or dynamic), and Scene-based audio.

[4] Note: HLG10, as used in DVB [12] is further limited to the Non-Constant Luminance $Y'C'_BC'_R$ signal format and narrow range quantization as defined in [5].

Nit	Unit of luminance measurement, weighted by the human visual system, formally specified in "candela per meter squared" (cd/m^2); the term "nits" is used in this document for convenience.
Opto-Electronic Transfer Function	The transfer function that maps scene light captured by the camera into digital pixel values.
Opto-optical Transfer Function	The overall transfer function that maps scene light captured by the camera to light values produced by the display.
Parametric	Audio encoding method that uses side-information to reconstruct the original audio information.
Perceptual Quantization	A high dynamic range EOTF used for HDR. PQ is specified in BT.2100 [5].
PQ10	The Perceptual Quantization EOTF described in BT.2100 [5] which requires BT.2020 [3] system colorimetry and 10-bit depth (see also Section 6.1.1) but has no associated dynamic or static metadata.[5]
Preselection	Set of Audio Program components representing a version of the Audio Program that may be selected for simultaneous decoding. An Audio Preselection is a subset of available Audio Program Components of one Audio Program.
Random Access Point	A collection of audio or video data packets that allow entry into a content stream without restarting the decoding process.
Renderer	A part of an NGA receive device, post decoding, that combines various sound program components (channels and objects) into the available reproduction channels while maintaining the original program intent and consistent audio loudness.
Resolution	The number of vertical and horizontal pixels available on a display device.
Set of Variants	A Set of Variants is a collection of Variants for a given segment of a multimedia asset. Variants contain the same perceptual content but different marks and can be used interchangeably. Sets of Variants for a given asset are typically generated during the first step in a two-step watermarking system.
Signal Format	Describes a triplet-based system that use different perceptual elements when combined properly make a complete image representation. Examples include $Y{\cdot}C_bC_r$, RGB, IC_tC_p
Standard Dynamic Range	Content graded as per BT.1886 [4] and BT. 709 [2] for HD television.
System Colorimetry	Specifies chromaticity of the color primaries and white point, allowing for a consistent reproducible representation of images. BT.2020 [3] and BT.709 [2] are examples of system colorimetries.
Tone Mapping	Process to convert HDR/2020 video to SDR/709 video. Also known as 'down-mapping'

[5] Note: PQ10, as used in DVB [12] is further limited to the Non-Constant Luminance $Y'C'_BC'_R$ signal format and narrow range quantization as defined in [5].

Variant	A Variant is an alternative representation of a given segment of a multimedia asset. Typically, a Variant is a pre-watermarked version of the segment using a Forensic Watermarking technology. The size of the segment varies for different Forensic Watermarking technologies: a few bytes, a frame, a group of pictures, a video fragment.
Variant Sequence Generator	A Variant Sequence Generator (VSG) selects a single Variant in each Set of Variants to produce a Variant Sequence. The VSG is part of the second step in a two-step watermarking system.
Variant Sequence	A Variant Sequence is a sequence of Variants that encodes a desired Watermark Identifier.
UHD-1	UHD at resolution of 3840 H by 2160 V (this is a 4K resolution).
UHD-2	UHD at resolution of 7680 H by 4320 V (this is an 8K resolution).
Wide Color Gamut	Color gamut wider than the gamut of BT.709 [2].
Watermark Identifier	A serialization number that is embedded in a multimedia asset using a Forensic Watermarking technology to make the asset unique. Examples of data used as a Watermark Identifier are session IDs, client IDs, device IDs, firmware versions, timestamps, etc. The Watermark Identifier is also routinely referred to as the *payload* or the *message* in the watermarking literature. See also Figure 5.

3.2 Acronyms and Abbreviations

ABR	Adaptive Bit Rate
ACES	Academy Color Encoding System
AVC	Advanced Video Coding
AVR	Audio/Video Receiver
BL	Base Layer
CA	Conditional Access
CAE	Content Aware Encoding or Content Adaptive Encoding
CBA	Channel Based Audio
CBR	Constant Bit Rate
CVBR	Capped Variable Bit Rate
CDN	Content Delivery Network
CG	Character Generator
CGI	Computer Generated Imagery
DASH	Dynamic Adaptive Streaming over HTTP
DOCSIS	Data Over Cable Service Interface Specification
DRC	Dynamic Range Control
DRM	Digital Rights Management
DTT	Digital Terrestrial Transmission

DVE	Digital Video Effects
EL	Enhancement Layer
EMB	Watermark EMBedder
ENC	Video ENCoder
EOTF	Electro-Optical Transfer Function
EPB	Encoder Boundary Point
HD	High Definition
HDR	High Dynamic Range
HEVC	High Efficiency Video Coding
HFR	High Frame Rate
HLG	Hybrid Log-Gamma
HLS	HTTP Live Streaming
HOA	High Order Ambisonics
HTTP	Hyper Text Transfer Protocol
IP	Internet Protocol
IPTV	Internet Protocol Television
ISO	International Standards Organization
ISOBMFF	ISO Base Media File Format [28]
ITM	Inverse Tone Mapping
JOC	Joint Object Coding
LUT	Look Up Table
MPD	Media Presentation Description
MTF	Modulation Transfer Function
MVPD	Multichannel Video Programming Distributor
NALU	Network Abstraction Layer Unit
NGA	Next Generation Audio
OBA	Object Based Audio
OETF	Opto-Electronic Transfer Function
OOTF	Opto-Optical Transfer Function
OTT	Over-the-Top (i.e., Internet-based transmission of content)
PCM	Pulse-Code Modulation
PES	Packetized Elementary Stream
PQ	Perceptual Quantization
PVR	Personal Video Recorder
RTP	Real-Time Transport Protocol
SD	Standard Definition
SEI	Supplemental Enhancement Information
SDR	Standard Dynamic Range
SFR	Standard Frame Rate

SHVC	Scalable High-Efficiency Video Coding (see Annex H of [69]
STB	Set Top Box
TM	Tone Mapping
TSD	Transport Stream Decoder
UDP	User Datagram Protocol
UHD	Ultra High Definition (see “Foundation UHD” in Section 3.1 above for use of this term within the scope of this document)
URI	Uniform Resource Identifier
VBR	Variable Bit Rate
VDS	Video Description Service
VSG	Variant Sequence Generator
VOD	Video-on-Demand
WCG	Wide Color Gamut
WM	WaterMark
WM ID	Watermark Identifier
xDSL	Digital Subscriber Line (x indicates any variety, e.g., ADSL, HDSL, SDSL, etc.)

4. Phases and Timeframes

These guidelines describe a number of UHD-related technologies. The Ultra HD Forum notes that some of these technologies can be considered "Foundation" UHD technologies, which are core aspects of UHD content and services, such as Wide Color Gamut (WCG). The Ultra HD Forum also describes additional technologies, which service providers may find to be valuable enhancements or enablers to a core UHD service, such as dynamic HDR metadata.

The high-level media attributes for Foundation UHD technologies are listed in Section 4.1. The high-level media attributes for additional technologies are found in Section 4.2 and in several Annexes. Detailed recommendations are grouped into Production / Post-Production, Distribution (includes compression and distribution for contribution, primary and final delivery stages), and Decoding / Rendering.

4.1 Foundation UHD Technologies

Foundation UHD content and services are often those that were commercially available as early as 2016, and thus have reached a level of market adoption and maturity. For the purposes of this document, Foundation UHD comprises the following characteristics:

- Resolution – greater than or equal to 1080p and lower than or equal to 2160p, (progressive format; BT.2100 [5] does not include interlaced formats)
- Wide Color Gamut – color gamut wider than BT.709 [2]
- High Dynamic Range – greater than or equal to the contrast ratio that could be derived from 13 f-stops of dynamic range
- Bit depth – 10-bit
- Frame rates – up to 60fps (integer frame rates are preferred; note that cinematic content may opt to use lower frame rates, e.g., see DCI specification [17])
- Audio – 5.1 channel surround sound or channel-based Immersive Audio (2.0 stereo is possible; however, 5.1 or channel-based Immersive Audio are preferred)
- Closed Captions/Subtitles – CTA 708/608, ETSI 300 743, ETSI 300 472, SCTE-27, IMSC1

The following terms are used in this document for HDR and HDR plus WCG:

- HLG: The Hybrid Log-Gamma OETF defined in BT.2100 [5]
- HLG10: The Hybrid Log-Gamma OETF described in BT.2100 [5] together with BT.2020 [3] system colorimetry and 10-bit depth.
- PQ: The Perceptual Quantization EOTF defined in BT.2100 [5]
- PQ10: The Perceptual Quantization EOTF described in BT.2100 [5] together with BT.2020 [3] color system colorimetry and 10-bit depth
- HDR10: The Perceptual Quantization EOTF with BT.2020 [3] system colorimetry, 10-bit depth, and ST 2086 [10] static metadata, and the MaxCLL and MaxFALL static metadata [26]; an HDR10 system or format is capable of handling these static metadata when present

Table 2 Foundation UHD Workflow Parameters

Content Creation & Mastering	Defined and documented standard workflows for Live and Pre-recorded content
Service Type	Real-time Program Services; On-Demand content that was originally offered as Live content
Network Type	Unicast (including Adaptive Bit Rate) Broadcast, Multicast
Transport/Container/Format	MPEG TS, Multicast IP, DASH ISO BMFF
Interface to TVs (source format)	IP connected (for OTT content delivered via managed or unmanaged network) HDMI (for services delivered via a STB, e.g., OTT, MVPD)
Backward Compatibility	Native (HLG), simulcast (HDR10/PQ10), decoder based (optional)

Table 3 Foundation UHD Content Parameters

Spatial Resolution	1080p* or 2160p
System Colorimetry	BT.709 [2], BT.2020 [3]
Bit Depth	10-bit
Dynamic Range	SDR, PQ, HLG
Frame Rate**	24d(23.976), 25, 30(29.97), 50, 60(59.94)
Video Codec	HEVC, Main 10 Profile, Level 5 or 5.1 (single layer)***
Audio Channels	Stereo or 5.1 or channel-based Immersive Audio
Audio Codec	AC-3, E-AC-3, E-AC-3 + JOC, HE-ACC, AAC-LC
Captions/Subtitles Coding	CTA 608/708, ETSI 300 743, ETSI 300 472, SCTE-27, IMSC1

Table 3 Notes:

*1080p together with WCG and HDR fulfills certain use cases for Foundation UHD services and is therefore considered to be an Ultra HD format for the purpose of this document. 1080p without WCG or HDR is considered to be an HD format. The possibility of 1080i or 720p plus HDR and WCG is not considered here. HDR and WCG for multiscreen resolutions may be considered in the future.

**Fractional frame rates for 24, 30 and 60 fps are included for compatibility with current plant video clock reference, but ultimately not preferred. Fractional frame rates will be necessary during migration from legacy video systems. The lower frame rates may be common for cinematic content.

*** For use in China, the AVS2 codec, Main10 profile, is used in addition to HEVC. See Annex E: AVS2.

For the purpose of this document, including the above constraint on 1080p content, various combinations of these Foundation UHD parameters can be combined to produce UHD content. Additional, non-Foundation technologies may also be employed.

The Foundation UHD codec is HEVC[6] for distribution to consumers, due to its support for HDR/WCG (10-bit) as well as coding efficiency, which makes 4K content more feasible. AVC specifications have been updated to include support for 10-bit depth with BT.2020 system colorimetry and support of PQ/HLG High Dynamic Range formats. However, the Ultra HD

[6] For use in China, the AVS2 codec, Main10 profile, may be used instead of HEVC. See Annex E: AVS2.

Forum finds that deployed consumer decoders are mostly hardware-based with no feasible mechanism to upgrade to match these latest updates, and thus these new AVC capabilities are not generally usable for distribution purposes. For AVC, consumer decoders generally support 8-bit 4:2:0 formatted content and normally with limitations with regard to the maximum bit rates that would normally preclude decoding 4K content at the compression rates that AVC can achieve.

Using AVC for production, contribution and mezzanine workflows is more viable because changing to encoders with support of the HDR/10-bit capabilities is something that is more easily accomplished.

Table 4 and Table 5 categorizes decoders and services in terms of Foundation UHD capability. Additionally, Table 5 offers some indications of which types of decoders are compatible with which service formats. Foundation decoder and service formats provide the base encoding formats for a number of enhancement features which may be implemented without causing service incompatibility with Foundation supporting devices. This will normally be described in the sections covering the enhancement features.

Table 4 Foundation Decoder Capabilities

Type No.	Color Container	Resolution	Frame rate	Bit Depth	HDR	SDR BT2020	HDMI	HDCP	UHDF Foundation
1	**BT.709**	**1080**	**P25/30**	**8**	**No**	**No**	**1.4**	**1.x**	**No**
2	BT.709	1080	P50/60	8	No	No	1.4	1.x	No
3	BT.709	2160	P25/30	8	No	No	1.4	1.x	No
4	BT.709	2160	P50/60	8	No	No	2.0	2.2	No
5	BT.2020	1080	P50/60	10	No	Yes	2.0	2.2	No
6	**BT.2020**	**2160**	**P50/60**	**10**	**No**	**Yes**	**2.0**	**2.2**	**Yes**
7	BT.2020	1080	P50/60	10	PQ10	Yes	2.0a	2.2	Yes
8	BT.2020*	1080	P50/60	10, 8	PQ10	No	2.0a	2.2	Yes
9	BT.2020	2160	P50/60	10	PQ10	Yes	2.0a	2.2	Yes
10	BT.2020*	2160	P50/60	10, 8	PQ10	No	2.0a	2.2	Yes
11	BT.2020	1080	P50/60	10	HLG10/PQ10	Yes	2.0b	2.2	Yes
12	**BT.2020**	**2160**	**P50/60**	**10**	**HLG10/PQ10**	**Yes**	**2.0b**	**2.2**	**Yes**
13	BT.2020*	1080	P50/60	10, 8	HLG10/PQ10	No	2.0b	2.2	Yes
14	BT.2020*	2160	P50/60	10, 8	HLG10/PQ10	No	2.0b	2.2	Yes

Table 4 notes:

- The Ultra HD Forum finds that decoder types 1, 6, and 12 (in **bold** text) comprise the vast majority of decoders, with type 6 being more common in Europe than in other regions due to support for backward compatibility with HLG formatted HDR.
- Decoders capable of HDR are assumed to also be capable of SDR.
- Decoders capable of BT.2020 are assumed to also be capable of BT.709 at 10- and 8-bit depths.
    - *The Ultra HD Forum InterOp Work Group has found that some decoders capable of HDR/BT.2020 are capable of SDR/BT.709, but not also capable of SDR/BT.2020.

- Decoders capable of 2160p resolution include support for 1080p resolution through up-conversion.
- Frame rates indicated are maximum supported for the decoder and include the support for lower frame rates, i.e., 24, 25, and 30, including fractional frame rates for 24, 30 and 60.

Table 5 Foundation Service Formats

Service Format Description	Attributes					Table 4 Decoders	UHDF Foundation
	Color Container	Resolution	Frame rate	Bit Depth	HDR		
HD SDR	BT.709	1080	P30	8	No	All	No
HDp60 SDR	BT.709	1080	P50/60	8	No	2 and above	No
UHD SDR	BT.709	2160	P30	8	No	3, 4, 6, 9, 10, 12 and 14	No
UHD SDR	BT.709	2160	P50/60	8	No	4, 6, 9, 10, 12 and 14	No
HDp60 SDR2020	BT.2020	1080	P50/60	10	No	5, 6, 7, 9, 11, and 12	No
UHD SDR2020	BT.2020	2160	P50/60	10	No	6, 9, and 12	Yes
HDp60 PQ10	BT.2020	1080	P50/60	10	Yes	7 and above	Yes
HDp60 HLG10	BT.2020	1080	P50/60	10	Yes	11 and above	Yes
HDp60 HLG10*	BT.2020	1080	P50/60	10	Yes	11 and above (HDR), 5 through 10 (SDR)	Yes
UHD PQ10	BT.2020	2160	P50/60	10	Yes	9, 10, 12, and 14	Yes
UHD HLG10	BT.2020	2160	P50/60	10	Yes	12 and 14 (HDR)	Yes
UHD HLG10*	BT.2020	2160	P50/60	10	Yes	12 and 14 (HDR), 6, 9, and 10 (SDR)	Yes

Table 5 notes:

*Indicates the Service Format signals HLG10 using the SDR/BT.2020 backward compatible method. See Section 6.1.9.

4.2 Additional UHD Technologies

In addition to the Foundation UHD technologies, the Ultra HD Forum provides guidance on a number of additional technologies that can enable or enhance Foundation UHD content and services.

The Ultra HD Forum notes that these additional UHD technologies can be “layered” onto Foundation UHD technologies in order to upgrade the consumer experience. For example, 7.1+4 immersive audio can be transmitted instead of 5.1 surround sound and/or dynamic HDR metadata can be included with HDR10 content and/or dynamic HDR metadata can be included

with HDR10-based. Content-aware encoding can be employed to increase encoding efficiency. Content producers and service providers may elect to implement one or more additional UHD technologies according to market drivers, the capabilities of the end-to-end ecosystem and consumer preferences. See Sections 12 through Annex E: AVS2.

5. Use Cases

The following use cases are intended to provide context for the guidelines defined within this document. They are not intended to be exhaustive, yet they cover those associated with the content creation and distribution ecosystem.

5.1 Digital Terrestrial Transmission

In February 2017, Korea launched terrestrial UHD TV commercial services using ATSC 3.0 with 4K spatial resolution and Next Gen Audio encoding. Japan, the U.S. and Europe may also commercially deploy UHD Digital Terrestrial Transmission (DTT) services.

DTT of 2160p content is an expensive proposition in terms of pixels/Hz. Some broadcasters, such as those in Korea, may have access to sufficient spectrum to deliver Ultra HD content in 2160p resolution. However, in other parts of the world, such as the U.S. or Europe, broadcasters' network capacity may be limited, especially if legacy HD/SD simulcasting and/or channel sharing is necessary. In this case, broadcasters may choose to offer advanced services in 1080p, 50/60fps, HDR/WCG format, which may be deployed in under 10Mbps. Where simulcasting and channel/sharing are not necessary and HEVC compression is sufficient, 2160p content can be broadcast DTT. It is possible that in some countries, broadcasters will use satellite to deliver an Ultra HD experience until spectrum is allocated on terrestrial networks.

In countries where bandwidth is constrained, the expectation is that a single HDR/WCG service with direct backwards compatibility may be most desirable (i.e., not simulcast with a separate stream in SDR/BT.709). Because broadcasters target TVs directly, they must also consider backward compatibility with deployed consumer electronics or other infrastructure that is not capable of processing some aspects of Ultra HD content so simulcasting may be a necessity nonetheless.

5.2 MVPD Platform Delivery

Programming distributors may wish to provide a consumer with a high-quality experience via cable, satellite or IPTV or, as a secondary delivery system, OTT. The programming distributor may deliver content to the following devices:

- Set-Top Box (STB)
- Other media device (e.g., TV, tablet, smart phone)

The content will need to be transcoded into multiple formats and bit-rates to provide the best experience to the consumer. The following will need to be considered:

- 1080p with HDR/WCG for low bit-rate delivery

Consumers may be at home or mobile and will access content across multiple devices. Some devices may provide the consumer a superior experience based on decoding and display capability, and content should be created with these variables in mind.

5.3 IP Network Delivery

Adaptive Bit Rate (ABR) is an HTTP delivery solution suitable for multiple streaming formats, such as HLS (HTTP Live Streaming, implemented by Apple®) or DASH. ABR can be used over a managed network (e.g., DOCSIS 3.x [95][96] or DVB Digital Cable [97]) or an unmanaged network (i.e., the public Internet). Content-aware Encoding can be an enabling technology capable of added efficiency.

In a managed network, ABR content can be delivered by an MVPD that is also an Internet service provider (ISP), e.g. a cable operator or telco operator. The content is sent over a managed network in IP Services. This is also referred to as IPTV. There are a number of technologies that can be used, e.g., DOCSIS 3.x. MVPDs who are ISPs may also send their IP Services over networks managed by other ISPs in what is known as "TV Everywhere."

In an unmanaged network, a type of service provider referred to as an edge provider, e.g., Netflix®, Amazon® and others, sends IP services over multiple ISP networks on the public Internet. This is known as over the top (OTT).

Today Live event content producers such as sports league owners and others may choose to provide consumers with a high-quality Ultra HD audio/video experience of a real-time event over an IP network. When content is captured and mastered using parameters described in this document and delivered to compatible displays, IP networks can deliver a high-quality consumer experience.

So that a high-quality experience can be delivered over varying network conditions, the content is transcoded into multiple versions suitable for various bitrates to form sets of encoded content at different bitrate levels. The sets allow seamless switching between the higher and lower bitrate versions of the real-time content, i.e., "adapting" as network conditions vary. In order to make the most efficient use of available bandwidth to the consumer, a content producer will get the best results using advanced picture encoding technologies (e.g., HEVC), which have been engineered specifically for such applications (see also Section 9.3.2).

The following represents the minimum, under normal network conditions, that should be supported from camera to the consumer's television:

- 1080p resolution with HDR, WCG, and 10-bit depth

UHD ABR content may be delivered via an IP network to the following devices.

MVPD/ISPs:

- STBs managed by MVPDs that connect to a television.

OTT or TV Everywhere:

- OTT Streaming Boxes (STB) that connect to a television (e.g., game consoles, Roku® boxes and similar devices).
- OTT-capable media devices (e.g., smart TVs, tablets and smart phones) that include a display panel.

Content producers will continue to deliver programming to customers that are using 'legacy' devices (i.e., devices that only support SDR, BT.709 [2] system colorimetry and stereo audio). Content distribution network (CDN) partners may need to host both UHD and legacy files.

It should be noted that if the content is delivered to the consumer device via a home Wi-Fi connection, the quality of service may be impacted by the available bandwidth on the home LAN. A wired connection to the device may be preferred.

Table 6 describes various options for delivering UHD content via IP networks. A list of commercial services currently employing these (and other) methods can be found at the Ultra HD Forum website, https://ultrahdforum.org/resources-categories/information-on-uhd-deployments/.

Table 6 UHD over IP Networks

Operator	Protocol	Network	Format	Video
IPTV	IP multicast	Fiber / xDSL	UDP	Single bitrate
IPTV	ABR unicast (managed network)	Fiber / xDSL	HTTP (HLS/DASH)	ABR
IP cable	ABR unicast (managed network)	DOCSIS 3.x	HTTP (HLS/DASH)	ABR
OTT TV (live)	ABR unicast (un-managed network)	Fiber / xDSL	HTTP (HLS/DASH)	ABR

6. Production and Post Production

The UHD Forum is concerned with establishing viable workflows both for Real-time Program Services and On Demand content that was originally offered live. Real-time Program Services (aka Linear TV) make frequent use of Pre-recorded material, such as edited inserts, interstitials, etc., which involve production and post-production.

Live content has specific requirements and operating practices that are unlike Digital Cinema, Blu-ray™ disc mastering, or other Pre-recorded content practices. Ultra HD workflows and technologies that are designed for these other delivery methods may not apply to Live content production.

Production practices for Foundation UHD audio are similar to those used in current HD content creation. Audio follows multi-channel workflows established for multi-channel 5.1 surround delivery using (as appropriate) AC-3 [29], E-AC-3+JOC (an instance of Dolby Atmos®[7]) [35], HE-AAC [27], or AAC-LC emission [27]. Although 2.0 stereo sound is possible, Foundation UHD content is considered premium content, and it is therefore recommended that productions provide at least 5.1 channels.

Foundation UHD, production practices for closed captions and subtitles are also similar to those of HD content creation. Closed captions and subtitles follow workflows established for CTA- 608/CTA-708, ETSI 300 743, ETSI 300 472, SCTE-27, or IMSC1 formats.

The remainder of this section will focus on mechanisms for producing the video components of the content.

As content is produced, it is useful to know in advance for which service mode(s) the content is intended. Equally, service providers planning to deliver Foundation UHD content need to have an understanding of the formats in which the content will be supplied.

The following is a diagram providing an overview of the content production and distribution workflow for Real-time Program Services and potentially capturing Live content for later distribution via an On-Demand Service.

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

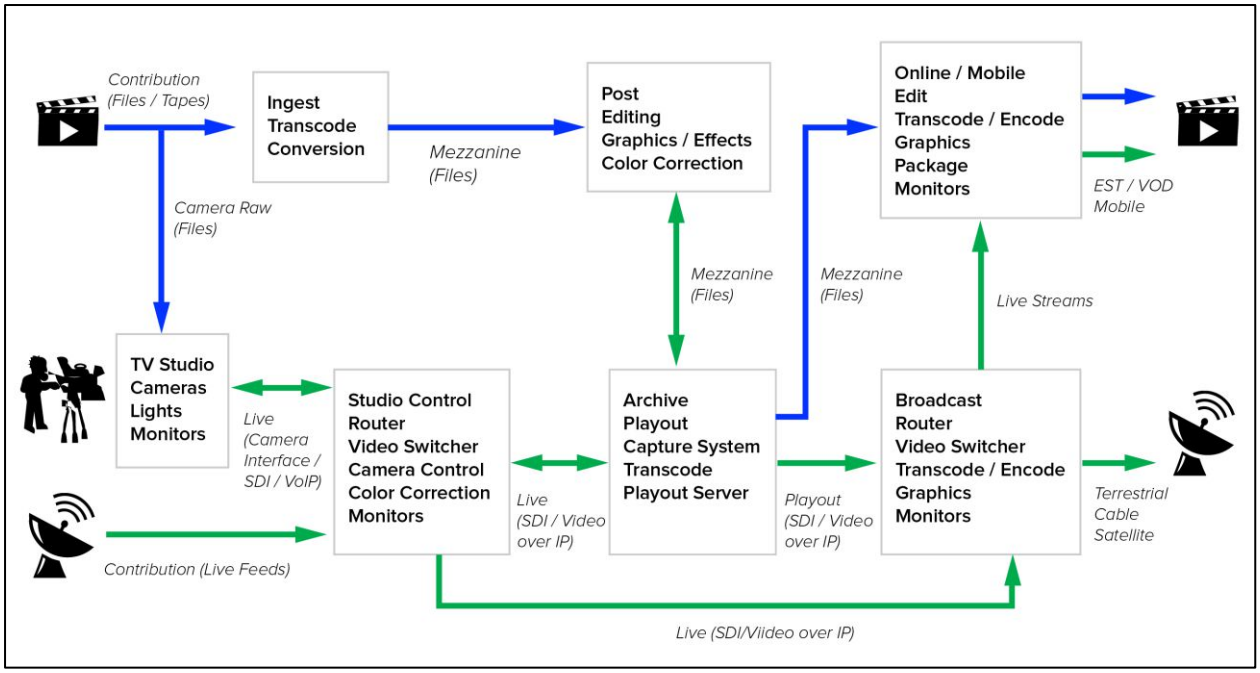


Figure 1 Content Production and Distribution Workflow

6.1 HDR/WCG Technologies

There are many terms in use in the field of HDR television. This section explains the existing terminology and how terms are used in these Guidelines.

Note that currently some UHD displays are capable of accepting BT.2020 [3] content, but as of this publication, no direct view display is available that is capable of rendering the full gamut of colors in the BT.2020 [3] system colorimetry. It is assumed that in these cases, the device employs "best effort" gamut mapping tailored to its particular display characteristics, and thus these devices are considered BT.2020 [3]-compatible.

6.1.1 Perceptual Quantization (PQ) and PQ10

One HDR transfer function defined for use in television is the "Perceptual Quantization" (PQ) developed by Dolby. PQ is defined as a reference display transfer curve, or EOTF, that is employed on the mastering monitor used in the reference viewing environment. The curve is designed to minimize visibility of banding on a display over the brightness range of 0 to 10,000 cd/m^2. SMPTE standardized the PQ curve in ST 2084 [9] and the ITU-R standardized it in BT.2100 [5]. ST 2084 specifies the shape of the curve over a range of 0 to 1 but does not specify SDI code values for the 0 and 1 values. Table 9 in BT.2100 [5] describes the code mapping for both narrow range and full range code values.[8]

The Ultra HD Forum has defined the term PQ10 to refer to the "base package" of parameters that PQ content typically has; i.e., the Perceptual Quantization EOTF described in BT.2100 [5] together with BT.2020 [3] system colorimetry and 10-bit depth. As such, PQ10 content or systems may or may not be metadata-capable. For example, HDR10 is a Foundation

[8] CTA 861-G [31] offers additional information about code ranges.

UHD PQ10-based format that has the capability of including certain static metadata (see Section 6.1.5).

The PQ signal is "display-referred", meaning that the pixel-encoded values represent specific values of luminance for displayed pixels. The intent is that only the luminance values near the minimum or maximum luminance capability of a display are necessarily adjusted to utilize the available dynamic range of the display. Some implementations may apply a "knee" at the compensation points in order to provide a smoother transition from the coded values to the display capabilities; e.g., to avoid "clipping".

When default display settings are engaged, PQ enables pixel values in the mid-range, including skin tones, to be rendered on a display at the same (absolute) luminance level that was determined at production.

For example, if a scene was graded on a 1000 cd/m^2 grading monitor and then displayed on a 4000 cd/m^2 display, the skin tones can be rendered at the same luminance values on the 4000-nit display as on the 1000-nit monitor per the grader's intent, while the speculars and darker tones can be smoothly extended to take full advantage of the 4000-nit display. See also Section 6.1.10.

The Ultra HD Forum InterOp testing has shown that consumer displays vary significantly, especially when different viewing "modes" are selected (e.g., "sports mode" or "movie mode"). Research has also shown that ambient light in the viewing room can have an impact on humans' perception of a displayed image. Displays used in non-reference viewing environments may employ adjustments to the PQ curve in order to provide some compensation for the difference between the actual viewing environment and the reference viewing environment.

PQ content requires a down-mapping step in order to provide acceptable SDR quality. See Section 10.4 below for a deeper discussion of backward compatibility, including the pros, cons and open questions that apply to various possible methods.

6.1.2 Hybrid Log-Gamma (HLG) and HLG10

Another HDR transfer function defined for use in television is "Hybrid Log-Gamma" (HLG) developed by the BBC and NHK. This is defined as a camera capture transfer curve, or OETF. This curve was designed to provide HDR while maintaining a degree of backward compatibility with SDR/BT.2020 displays. The HLG curve has been specified in BT.2100 [5].

The Ultra HD Forum has defined the term HLG10 as HLG with 10-bit representation, black at code 64, and nominal peak at code 940[9].

HLG is a "scene-referenced" HDR technology that uses pixel-encoded values that are intended to represent picture luminance levels relative to each other in a given scene. The intent is that all the displayed pixel luminance values may be adjusted in a defined manner to compensate for specific display capabilities (e.g., peak luminance) and viewing environments in such a way that the values retain their perceptual appearance relative to one another.

As the eye's sensitivity to light intensity is approximately logarithmic, a power (or "gamma") law is applied to the HLG relative scene-light pixel values, to scale them to span the luminance range of the display; thereby approximating the same relative subjective brightness values that were determined at production. Since no one luminance range is "fixed", the (absolute) displayed luminance of mid-range values, including skin tones, will increase or decrease to scale along with all other values in order to better preserve the appearance of the relative brightness values of the pixels in the scene.

[9] CTA 861-G [31] offers additional information about code ranges.

For example, if a scene was graded on a 1000 cd/m^2 grading monitor and then displayed on a 4000 cd/m^2 display, the skin tones and other scene elements will be brighter on the 4000-nit display than on the 1000-nit monitor. However, given that all pixel values are adjusted through the "gamma law", the overall image remains perceptually similar to the original "look" selected by the grader. See also Section 6.1.10[10].

As mentioned above, Ultra HD Forum InterOp testing has shown that consumer displays vary significantly, especially when different viewing "modes" are selected (e.g., "sports mode" or "movie mode"). Research has also shown that ambient light in the viewing room can have an impact on humans' perception of a displayed image. Section 6.2 of ITU-R report BT.2390 [6] describes how the display gamma may be adjusted to provide some compensation.

Content produced using HLG can be displayed on SDR/WCG devices with a degree of compatibility that may be judged acceptable for programs and services according to Report ITU-R BT.2390 [6], and subjective tests performed by the EBU, RAI, IRT and Orange Labs [88]. Backward-compatible HLG is only intended to support SDR/BT.2020(WCG) displays and not intended for displays which only support SDR/BT.709. See Section 10.4 below for a deeper discussion of backward compatibility, including the pros, cons and open questions that apply to various possible methods.

6.1.3 Recommendation ITU-R BT.2100

In July 2016 ITU-R Study Group 6 approved the publication of a Recommendation on HDR for use in production and international program exchange and is known as ITU_R BT.2100 [5]. This recommendation includes the following specifications:

- Spatial resolutions: 1080p, 2160p, 4320p
- Frame rates: 24/1.001, 24, 25, 30/1.001, 30, 50, 60/1.001, 60, 100, 120/1.001, 120
- System Colorimetry: Same as BT.2020 [3]
- Reference Viewing Environment: 5 cd/m^2 background and less than or equal to 5 cd/m^2 surround
- Reference non-linear transfer functions EOTF, OOTF, OETF: PQ and HLG
- Signal Format: $Y'C'_BC'_R$, IC_TC_P, and RGB.
- Color sub-sampling: same alignment as specified in BT.2020 [3]
- Bit depth value ranges:
  - 10-bit, Narrow (64-940) and Full (0-1023) ranges
  - 12-bit, Narrow (256-3760) and Full (0-4095)
- Floating point signal representation: Linear RGB, 16-bit floating point

BT.2100 [5] is the primary reference document on HDR for use in production and international program exchange. Note that a full signal specification will need to include the following attributes: spatial resolution, frame rate, transfer function (PQ or HLG), color signal format, integer (10 or 12 bits, narrow or full range) or floating point.

[10] It is well known that changes in brightness of a display may effect the perception of image color (Hunt Effect) [100] as well as image contrast (Stevens) [101]. Research on the Color Appearance Model (CAM), where brightness exceeds 700 cd/m^2, agrees with these observations [101]; however the BBC has found no evidence of significant changes to the perception of color or contrast in their research using HLG with displays of different peak brightness [102]. Additional research is on-going on these effects with respect to HDR imagery.

Not all of the parameters listed above were deployed as early as 2016, but they are included as informative details.

6.1.4 Static Metadata – SMPTE ST 2086, MaxFALL, MaxCLL

SMPTE has specified a set of static metadata in the ST 2086 [10] Mastering Display Color Volume Metadata Supporting High Luminance and Wide Color Gamut Images standard. Parameters included indicate the characteristics of the mastering display monitor. The mastering display metadata indicates that the creative intent was established on a monitor having the described characteristics. If provided, the implication is that the artistic intent of the content is within the subset of the overall container per the metadata values. The mastering display characteristics include the display primaries and white point as x,y chromaticity coordinates, and the maximum and minimum display luminance. For example, the metadata may indicate that the system colorimetry of the mastering display is the DCI-P3 gamut in the BT.2020 [3] container- and the luminance range is a subset of the 0 to 10,000 cd/m^2 provided by PQ.

The Blu-ray Disc Association and DECE groups have defined carriage of two additional metadata items:

- MaxFALL – Maximum Frame Average Light Level; this is the largest average pixel light value of any video frame in the program
- MaxCLL – Maximum Content Light Level: this is the largest individual pixel light value of any video frame in the program

Static metadata may be used by displays to control color volume transform of the received program to better reproduce the creative intent as shown on the mastering display, given the capabilities of the display device. However, the use of MaxFALL and MaxCLL static metadata have limitations for use with Live broadcasts since it is difficult to determine a program's maximum pixel light values during a Live production. According to the UHD Alliance, today's mastering practices may generate outlier values unintentionally, causing the content's associated MaxCLL value to be higher than expected. As a response to that, some content providers use statistical analysis to calculate a MaxCLL value that is more representative of the statistically significant brightest pixels contained in the image sequence. The Ultra HD Forum further suggests that such statistical methodology may also apply to MaxFALL.

It is worth noting that code levels below minimum mastering display luminance and code levels above maximum mastering display luminance were likely clipped on a professional reference monitor and therefore any detail in pictures utilizing code levels outside this range were likely not seen in the content production process.

6.1.5 HDR10

HDR10 is a PQ10-based HDR format. The term "HDR10" is in widespread use and has been formally and consistently defined by several groups, including DECE and BDA, as:

- Transfer curve: BT.2100 [5] (PQ)
- System Colorimetry: BT.2020 [3]
- Bit depth: 10 bits
- Metadata: ST 2086, MaxFALL, MaxCLL

Several delivery formats (e.g. Ultra HD Blu-ray™ and CTA's HDR10 profile) have specified delivery using the above parameters with the metadata mandatory but are still considered to be using HDR10 for the purposes of this document. While ATSC A/341 does not

use the term "HDR10," when the PQ transfer curve is used, the static metadata video parameters described therein are consistent with HDR10 as defined herein.

Note that some displays ignore some or all static metadata (i.e., ST 2086, MaxFALL, and MaxCLL); however, HDR10 distribution systems must deliver the static metadata, when present.

6.1.6 Foundation UHD HDR Technologies

The following HDR/WCG "packages" are recommended for Foundation UHD due to their conformance with the criteria listed in Section 4 above. The two technologies are:

- HLG10 – inclusive of HLG OETF per BT.2100 [5], BT.2020 [3] system colorimetry, and 10-bit sample depth. (See also Section 10.1 for information about carriage of HLG transfer function signaling over HDMI interfaces.)
- PQ10 – inclusive of PQ EOTF per BT.2100 [5], Rec. ITU-R BT. 2020 [3] system colorimetry, and 10-bit sample depth
- HDR10 – inclusive of PQ EOTF per BT.2100 [5] with BT.2020 [3] system colorimetry, 10-bit depth, and ST 2086 [10] static metadata, and the MaxCLL and MaxFALL static metadata [26]; i.e., an HDR10 system or format is capable of handling these static metadata when present

It should be noted that Real-time Program Services are typically comprised of both Live and Pre-recorded content, and it is not recommended that service providers alternate between SDR and HDR signal formats or mix different HDR formats. See Section8.1 for details.

6.1.7 HDR10 Metadata Generation

HDR10 includes the static metadata described in Section 6.1.5. MaxFALL and MaxCLL metadata could be generated by the color grading software or other video analysis software. In Live content production, MaxFALL or MaxCLL metadata is not generated during the production process. By definition, it is not possible to generate MaxFALL or MaxCLL for a Live program because these cannot be known until the entire program is produced, i.e., after the program is over. The special values of '0' (meaning, "unknown") are allowed for MaxFALL and MaxCLL. It may be possible to set limits on the output and thus pre-determine MaxFALL and MaxCLL even for Live content production. For example, if MaxFALL and MaxCLL metadata values are provided, a video processor could be inserted in order to reduce brightness of any video frames that would exceed the indicated values (similar to the way audio processors are often inserted to enforce audio loudness and peak levels).

If it is desired to use HDR10 in Live content production, but the production facility does not support carriage of the metadata, then default values for ST 2086 [10], MaxFALL and MaxCLL should be determined and entered directly into the encoder via a UI or a file. SMPTE 2086 metadata could be set to values that represent the monitors used for grading during production of the content.

6.1.8 HDR10 Metadata Carriage

When the Ultra HD Forum Guidelines were originally published, standards did not exist to carry HDR10 static metadata over an SDI interface. The metadata had to be either embedded within the content intended for encoding (for Pre-recorded content) or supplied to the encoder

via a file or UI. However with the publication of SMPTE ST 2108-1 [48], there are now a variety of static and dynamic color transform metadata items that can be carried in the Vertical Ancillary Data Space (VANC) as Ancillary Data (ANC) Messages over an SDI interface or over an SMPTE ST 2110 IP connection (per SMPTE ST 2110-40 [47], and see Section 6.1.11).

In HEVC [26] and AVC [25] bitstreams, MaxFALL and MaxCLL may be carried in the Content Light Level (CLL) static SEI message (sections D.2.35 (Syntax) and D.3.35 (Semantics), [26]). SMPTE ST 2086 metadata is carried in the Mastering Display Color Volume (MDCV) SEI, (section D.2.28 (Syntax) and D.3.28 (Semantics), [26]). Both the CLL and MDCV SEI messages can be represented by their corresponding ANC messages as specified by SMPTE ST 2108-1 [48] and accompany each video frame to which they apply. An SMPTE ST 2108-1 ANC message is a simple, bit-accurate encapsulation of the corresponding SEI message, to ease handling by an encoder or decoder; as a result, SMPTE ST 2108-1 is also able to support certain dynamic metadata technologies (see Section 12).

Note that there may be multiple encoders in the end-to-end broadcast chain. In the event that HDR10 is used and the metadata does not reach the decoder/display, it is expected that the decoder/display will employ “best effort” to render the content accurately. Note that such an end-to-end chain is operating in a manner indistinguishable from PQ10 content (i.e. PQ without metadata).

Given that the insertion, carriage and preservation of static metadata presents new challenges in the broadcast environment, the use of PQ10 or HLG10 may be a practical alternative to the use of HDR10 in some Live workflows. See also Section 6.2.5 on Live content production and Section 8.5 on content distribution for additional details.

6.1.9 Signaling Transfer Function, System Colorimetry and Matrix Coefficients

The system colorimetry, transfer function (SDR or HDR), and matrix coefficients must be known to downstream equipment ingesting or rendering content in order to preserve display intent. This is true for file transfers in file-based workflows and in linear content streams in linear workflows.

In file-based workflows, mezzanine file formats such as IMF/MXF (stored in picture essence descriptors) and QuickTime (stored in NCLC and MDCV color atoms) are often used.

In linear production workflows, these values are transmitted via VPID (Video Payload Identifier) within HD-SDI [99]. In linear compressed transmission workflows, these values are typically signaled in the VUI of an H.265/HEVC or H.264/MPEG-4 AVC bitstream. Details on SEI and VUI messaging are available in the HEVC specification [26], in particular, Appendix D (SEI) and Appendix E (VUI).

The tables below summarize HEVC Main10 Profile bitstream SDR, PQ and HLG indicators. (In HEVC and AVC specifications, the bitstream elements are bolded and italicized to distinguish them from temporary variables and labels). As shown in the table below, there are two methods of signaling the HLG transfer function.

Table 7 File-Based Signaling for SDR/BT.709

	System Identifier	BT. 709 YCC	BT. 709 RGB	Full-Range 709 RGB	BT. 601 525	BT. 601 625
Color properties	Color primaries	BT.709	BT.709	BT.709	BT.601	BT.601
	Transfer Characteristics	BT.709	BT.709	BT.709	BT.709	BT.709
	Signal Format	Y′CbCr	R′G′B′	R’G’B’	Y′CbCr	Y′CbCr
Other	Full/narrow range	Narrow	Narrow	Full	Narrow	Narrow
	4:2:0 chroma sample location alignment	Interstitial	N/A	N/A	Interstitial	Interstitial
CICP parameters Rec. ITU-T H.273 ISO/IEC 23001-8 (QuickTime/ HE/AVC)*	ColorPrimaries	1	1	1	6	5
	TransferCharacteristics	1	1	1	6	6
	MatrixCoefficients	1	0	0	6	5
	VideoFullRangeFlag	0	0	1	0	0
SMPTE MXF parameters SMPTE ST 2067-21**	Color primaries	06.0E.2B.34.04.01.01.06.04.01.01.01.03.03.00.00			06.0E.2B.34.04.01.01.06.04.01.01.01.03.01.00.00	06.0E.2B.34.04.01.01.06.04.01.01.01.03.02.00.00
	Transfer Characteristic	06.0E.2B.34.04.01.01.01.04.01.01.01.01.02.00.00				
	Coding Equations	06.0E.2B.34.04.01.01.01.04.01.01.01.02.02.00.00	N/R	N/R	06.0E.2B.34.04.01.01.01.04.01.01.01.02.01.00.00	
	Full/narrow level range	Inferred (indicated in black reference level, white reference level, Color range)				
	4:2:0 chroma sample location alignment	Inferred (ChromaLocType=0)	N/A	N/A	Inferred (ChromaLocType=0)	Inferred (ChromaLocType=0)

Notes:

*QuickTime/MP4/HEVC/AVC: ISO ISO/IEC DIS 23091-2 Information technology - Coding-independent code points - Part 2: Video

**SMPTE UL’s available: https://registry.smpte-ra.org/view/published/labels_view.html

Table 8 File-Based Signaling for SDR/BT.2020

	System Identifier	BT. 2020 YCC NCL	BT. 2020 RGB NCL
Color properties	**Color Primaries**	BT.2020	BT.2020
	Transfer Characteristics	BT.2020	BT.2020
	Signal Format	Y′CbCr	R′G′B′
Other	**Full/narrow range**	Narrow	Narrow
	4:2:0 chroma sample location alignment	Co-sited	N/A
CICP parameters Rec. ITU-T H.273 | ISO/IEC 23001-8 **(QuickTime/ HEVC/AVC)***	**Color Primaries**	9	9
	TransferCharacteristics	14	14
	MatrixCoefficients	9	0
	VideoFullRangeFlag	0	0
SMPTE MXF parameters SMPTE ST 2067-21**	**Color Primaries**	06.0E.2B.34.04.01.01.0D.04.01.01.01.03.04.00.0	
	Transfer Characteristic	06.0E.2B.34.04.01.01.0E.04.01.01.01.01.09.00.00	
	Coding Equations	06.0E.2B.34.04.01.01.0D.04.01.01.01.02.06.00.00	N/R
	Full/narrow level range	Inferred (indicated in black reference level, white reference level, color range)	
	4:2:0 chroma sample location alignment	Inferred (ChromaLoc Type = 2)	N/A

Notes:

*QuickTime/MP4/HEVC/AVC: ISO ISO/IEC DIS 23091-2 Information technology - Coding-independent code points - Part 2: Video

**SMPTE UL's available: https://registry.smpte-ra.org/view/published/labels_view.html

Table 9 File-based Signaling for HDR/BT.2020

	System Identifier	BT2100 PQ YCC	BT. 2100 HLG YCC	BT.2100 PQ $ICTCP$	BT.2100 PQ RGB	BT.2100 HLG RGB
Color properties	**Color primaries**	BT.2020 / BT.2100	BT.2020 / BT.2100	BT.2100	BT.2020 / BT.2100	BT.2020 / BT.2100
	Transfer Characteristics	BT.2100 PQ	BT.2100 HLG	BT.2100 PQ	BT.2100 PQ	BT.2100 HLG
	Signal Format	Y′CbCr	Y′CbCr	$ICTCP$	R′G′B′	R′G′B′
Other	**Full/narrow range**	Narrow	Narrow	Narrow	Narrow	Narrow
	4:2:0 chroma sample location alignment	Co-sited	Co-sited	Co-sited	N/A	N/A
CICP parameters Rec. ITU-T H.273 ISO/IEC 23001-8 (QuickTime/ HEVC/ AVC)*	**Color Primaries**	9	9	9	9	9
	Transfer Characteristics	16	18	16	16	18
	MatrixCoefficients	9	9	14	0	0
	Video Full-Range Flag	0	0	0	0	0
SMPTE MXF parameters SMPTE ST 2067-21**	**Color Primaries**	06.0E.2B.34.04.01.01.0D.04.01.01.01.03.04.00.00				
	Transfer Characteristic	06.0E.2B.34.04.01.01.0D.04.01.01.01.01.01.0A.00.00	06.0E.2B.34.04.01.01.0D.04.01.01.01.01.01.0B.00.00	06.0E.2B.34.04.01.01.0D.04.01.01.01.01.01.0A.00.00	06.0E.2B.34.04.01.01.0D.04.01.01.01.01.0A.00.00	06.0E.2B.34.04.01.01.0D.04.01.01.01.01.01.0B.00.00
	Coding Equations	06.0E.2B.34.04.01.01.0D.04.01.01.01.02.06.00.00		06.0E.2B.34.04.01.01.0D.04.01.01.01.02.07.00.00	N/R	N/R
	Full/narrow level range	Inferred (indicated in black reference level, white reference level, Color range)				
	4:2:0 chroma sample location alignment	Inferred (ChromaLoc Type = 2)	Inferred (ChromaLoc Type = 2)	unknown	N/A	N/A

Notes:

*QuickTime/MP4/HEVC/AVC: ISO ISO/IEC DIS 23091-2 Information technology - Coding-independent code points - Part 2: Video

**SMPTE UL's available: https://registry.smpte-ra.org/view/published/labels_view.html

In one method, the SDR transfer function indicator is signaled in the VUI and the HLG transfer function indicator is transmitted using an alternative transfer characteristics SEI message embedded in the bitstream. In this way, an "HLG aware" STB or decoder/display

would recognize that the bitstream refers to content coded with HLG (since it is indicated by the preferred_transfer_characteristics syntax element of the SEI). If an "HLG aware" STB is connected to a TV that does not support HLG, the STB would transmit the SDR indicator over HDMI to the TV. If it is connected to a TV that supports HLG, the STB would copy the transfer function value in the SEI (to indicate HLG) and transmit this over HDMI to the TV.

In the other method, the HLG transfer function indicator is directly signaled in the VUI in the same way PQ or SDR would be signaled.

In theory it is possible to achieve a lossless conversion between the two methods of signaling HLG by flipping the VUI transfer function characteristics indicator value and inserting or removing the alternative transfer characteristic SEI.

Using the first method (i.e., including the SDR transfer function indicator in the VUI and the HLG transfer function indicator in the SEI) enables backward compatibility with SDR/WCG displays. Service providers may also deem results to be acceptable on SDR/WCG displays using the second method (i.e., including the HLG transfer function indicator in the VUI). Service providers may wish to test both methods.

Additional use cases for production-level signaling wrappers include Apple QuickTime™ Wrapper NCLC (Non-Constant Luminance Coding) and MDCV (Mastering Display Color Volume) color atoms using MPEG CICP (Coding Independent Code Point) and MXF Image Essence Descriptors from ST.2067:21:2016 [38].

As described in Section 8.1, service providers should convert or remap all content into a single, consistent system colorimetry and transfer function. Setting the initial values in the encoder should be adequate assuming the encoder is dedicated to a single format.

As of 2019, while there are methods for signaling system colorimetry, transfer function, matrix coefficients and HDR-related metadata through the end-to-end supply chain however due to devices with limited support for these capabilities it is extremely difficult to successfully ensure that these signals and metadata survive through the entire broadcast linear production chain. Because gaps exist with legacy devices using SDI interfaces, as well as HDMI interfaces and software that don't consistently support file-based signaling in mezzanine file wrappers, verification of proper signaling is recommended. The Ultra HD Forum observes that standards bodies have attempted to address these issues and provided documentation, much of which is referenced herein, but some of which is still under development.

.

6.1.10 Peak Brightness: Production, Ref. Monitors, Consumer Displays and Archives

Luminance values can be coded as "absolute values" on a scale of 0-10,000 nits (e.g., PQ format) or as "relative values" (e.g., RAW, gamma, S-log or HLG). The optimal coding method depends on the application and other considerations.

Cameras can be capable of capturing a higher contrast range than reference monitors and consumer displays are capable of rendering. When content is graded, the grader makes judgments according to what he or she sees within the luminance range of the reference monitor, which may have a peak brightness of 1,000 or 2,000 nits, for example. In theory, if the consumer display characteristics match that of the reference monitor, the consumer will see what the grader intended. In practice of course, consumer HDR displays have varying peak brightness performance, black level and color gamut, so it is up to the display (or source device) to map the color volume of the transmitted content to best match the capabilities of the particular display.

As a PQ signal may carry pixel values as high as 10,000 nits, it is helpful to the display to indicate the actual maximum pixel value to be expected that are creatively pertinent as a basis for color volume transform. HDR10 systems are capable of providing static metadata to assist mapping the transmitted content to the consumer display, enabling a given display to optimize the color volume mapping based on its capability. The use of HDR10 metadata may help to optimize display mapping, but as the metadata may not be practical in Live content, the use of PQ10 may be preferred for Live content. Although a system objective is to preserve creative intent with limited perceptual alteration all the way to the consumer display, some perceptual alteration will occur because of different display capabilities both in peak brightness and black levels.

HLG is not specified for use with metadata, and instead has a specified relationship between overall system gamma (implemented as part of the display EOTF) and nominal peak display luminance. An overall system gamma of 1.2 is specified for HLG when displayed on a 1,000 nit monitor. BT.2390 [6] states that the artistic intent of HLG content can be preserved when displaying that content on screens of different peak luminance (even when that display is brighter than the mastering display), through changing the system gamma. BT.2100 [5] provides a formula to determine overall system gamma based on the desired display peak luminance in the reference viewing environment. Section 6.2 of BT.2390 [6] further specifies how the system gamma should be adapted for darker or brighter viewing environments. To preserve creative intent, these formulae are recommended.

ITU-R BT.2408 [8] recommends the reference level for “HDR Reference White” be set to 75% HLG or 58% PQ (equivalent to 203 cd/m^2 on a 1,000 cd/m^2 reference display). These values were chosen to provide sufficient headroom for “specular highlights” while allowing comfortable viewing when HLG content is shown on HDR/WCG and SDR/WCG displays. Some experts have found that setting HDR Reference White at this level and using it as an anchor for mapping SDR content into an HDR container can produce a darker SDR image; however be advised that compensating for this during conversion from SDR to HDR by remapping levels to higher brightness values may cause clipping when converting back to SDR. It has been suggested that dynamic conversion can potentially improve this.

Note that future displays may become available with higher peak brightness capability compared with those available today. Content that is expected to be of interest to consumers for many years to come may benefit from retaining an archive copy coded in “absolute values” (e.g., PQ) or original camera capture format (e.g., RAW, log) so that future grades of higher luminance range can be produced and delivered to viewers.

6.1.11 Studio Video over IP

In 2017, the Society of Motion Picture and Television Engineers (SMPTE) published the first of its SMPTE ST 2110 suite of standards for “Professional Media Over Managed IP Networks”, which exploits IP networking equipment for transport of video, audio, and metadata within television facilities, in addition to or in lieu of traditional serial digital interfaces (SDI). The SMPTE ST 2110 suite is built upon the Realtime Transport Protocol (RTP) as described in RFC 3550 [42], with the suite providing sufficient constraints such that best-effort IP connections can succeed where historically, bandwidth and synchronization was provided by SDI connections, which offered assured point-to-point latencies.

The SMPTE ST 2110 suite currently comprises five documents: SMPTE ST 2110-10 [43] specifies the network interface requirements, overall system timing model, and session description protocols. SMPTE ST 2110-20 [44] specifies the format for uncompressed video essence, while SMPTE ST 2110-21 [45] identifies timing characteristics for these video

streams and requirements for compliant senders and receivers. SMPTE ST 2110-30 [46] specifies the format for uncompressed digital audio. Lastly, SMPTE ST 2110-40 [47], published in 2018, specifies carriage of SMPTE ST 291-1 Ancillary Data (i.e., ANC messages). Each of the video, audio, and metadata essences described in SMPTE ST 2110-20, -30, and -40 are individual streams that are synchronized in accordance with SMPTE ST 2110-10.

6.1.12 Adding Dynamic HDR Metadata to Foundation UHD

For PQ HDR content[11], as described in Section 6.1.5, HDR10 provides the static metadata elements in a PQ10-based HDR format as specified by ST2086, 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.

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.

[11] 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 Section 6.1.2. 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.

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 12.1 and SL-HDR2 described in Section 12.4.

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 12.3.

6.2 Production for Pre-recorded Content

This section focuses on the creation and distribution of pre-recorded content intended for inclusion in a Real-time Program Service. This includes all content filmed, captured digitally, or rendered in CGI and delivered via a file format to a service provider. Real-time Program Services may be delivered via MVPD (satellite, Cable, or IPTV), OTT and DTT. See Section 6.2.5 for production of Live content that is recorded for subsequent re-showing.

The following diagram depicts the interfaces and primary functions within the pre-recorded content creation process.

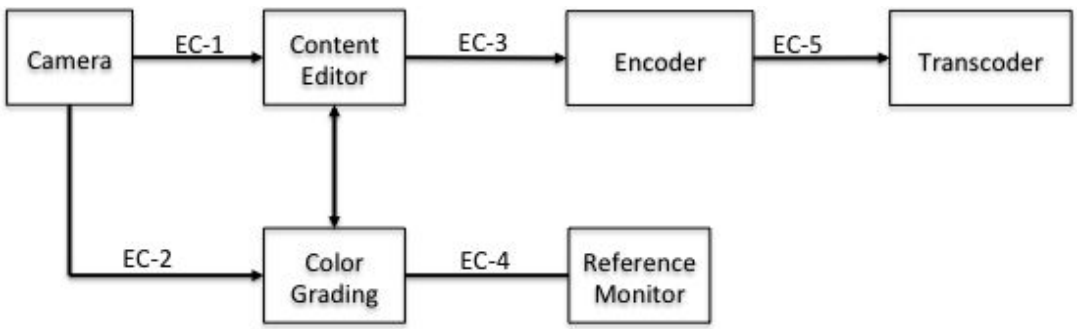


Figure 2 Pre-recorded Content Production Workflow and Interfaces

The scope of this guideline includes definition of six functions and five interfaces. The functional descriptions are described within sub-sections below, and the interface descriptions are described at a high level in the following table.

Table 10 Pre-recorded Content Interface Descriptions

Reference Point	Content Creation Functions	Reference Point Description
EC-1	Camera – Content Editor	Raw or log camera footage is ingested into the editing program. This interface may optionally contain metadata from the camera used by the editing program to provide guidance to timeline creation parameters.
EC-2	Camera – Color Grading	Raw or log camera footage is ingested into the color grading solution. This interface may optionally provide camera metadata describing the dynamic range resolution or other camera specific information from the captured footage.
EC-3	Content Editor – Encoding	Edited and color graded content is exported to an encoder for master and/or mezzanine level creation. This includes audio, video, essence, timing and CG content. Often the encoding function is included as part of the Content Editor.
EC-4	Color Grading – Professional Reference Monitor	This interface varies based on the content grading environment. It can include SDI, HDMI 2.0, Thunderbolt, DisplayPort and DVI.
EC-5	Encoder – Transcoder	This interface includes all aspects of a finished video delivered in file format for ingest to a transcoder. The transcoder uses this information to create distribution formats. If the transcoder is receiving uncompressed assets, it may also be an encoder.

6.2.1 Camera Requirements

6.2.1.1 High Dynamic Range and Wide Color Gamut Considerations

Material should be acquired at the highest resolution, dynamic range and color gamut of the camera in an operational data format best optimized for the image acquisition application in use. Although the primary consideration of this document is Foundation UHD, capturing in the native camera format allows for future content grades being used for distribution of additional UHD technologies.

There are numerous production examples that require the use of the camera specific raw or log data format from the camera sensing device. This is due to the requirement to acquire the highest level of picture quality to sustain the levels of post-production image processing usually encountered in high-end movie or a ‘made for TV’ drama or documentary with use of color correction, visual effects and the expected high levels of overall image quality.

There are equally numerous applications for episodic and live capture that employ camera-specific logarithmic (log) curves designed to optimize the mapping of the dynamic range capability of the camera sensors to the 10-bit digital interface infrastructure of production and TV studio operations. Such log 10-bit signals are often of the 4:2:2 color-sampling form and are either generated internally in the camera head or locally created in the camera control unit. These images are usually recorded utilizing high-quality mastering codecs for further post-

production of the captured scenes or, in the case of Live transmissions, minimally processed for real-time transmission to viewers.

Camera native log curves are typically designed by camera manufacturers to match the performance characteristics of the imaging sensor, such as sensitivity, noise, native dynamic range, color spectral characteristics, and response to extreme illumination artifacts. It should be noted that data describing the gamut and transfer function characteristics encoded at the camera must be passed down the production chain, particularly in cases where not all cameras used are operated with the same parameters. In post-produced content, such information is typically carried in file metadata, but such metadata is not embedded in the video signal.

6.2.1.2 Imaging Devices: Resolution, Dynamic Range and Spectral Characteristics

In the creation of Foundation UHD content signals, ideally, the image sensing devices should have a sensor resolution and dynamic range equal to or greater than a pixel count commensurate to the signal format.

In the area of spectral characteristics, the more advanced sensing devices will exhibit characteristics approaching the system colorimetry of BT.2020 [3], while more typical devices will produce color performance approximating the DCI-P3 gamut or just beyond the gamut of BT.709 [2].

Additional considerations for 2160p capture include:

- Not all lenses are suitable for capturing 2160p resolution and the MTF of the lens must be sufficient to allow 2160p capture on the sensor.
- Nyquist theory applies to camera sensors and care may be needed in selecting cameras that achieve target 2160p spatial resolution performance.
- When transmitting film content, the 16:9 aspect ratio of Foundation UHD does not correspond to the wider aspect ratio of some movies. The two alternative methods of re-formatting (full width or full height) represent the same compromises that exist in HD transmission.
- Content originating from 35mm film will translate to Foundation UHD differently than digital sources, e.g. film grain, achievable resolution.

6.2.2 Reference Monitor

Use of 2160p, HDR and WCG imply the need for reference monitors that will allow production staff to accurately adjust devices or modify content. Previously in SD and HD TV, there were accepted working practices using professional monitors and standard test signals, such as color bars, PLUGE, sweep, etc. Digital Cinema employs slightly different techniques, using calibrated monitors and LUTs to replicate the viewing characteristics of different consumer viewing conditions.

The recommendation for Foundation UHD is to agree on practical standardized methods for use of Reference Monitors for Real-time Program Service production.

Note that Live content, such as sports events using trucks and Live production switching, do not have a post-production stage; operation is fully real-time. In these environments, unlike post-production suites, there is a limited amount of opportunity for human intervention. Any human intervention happens in real-time by direct interaction with the acquisition device.

For Foundation UHD, a reference monitor can ideally render at least the following: resolutions up to 3840x2160, frame rates up to 60p, BT.2020 [3] system colorimetry (ideally at least the P3 gamut), and HDR (i.e., greater than or equal to the contrast ratio that could be derived from 13 f-stops of dynamic range). It should be noted that as with HD, consumer display technology is likely to progress beyond the capabilities of current generation

professional displays. As such, instruments such as waveform monitors and histogram displays are essential tools to ensure optimum Foundation UHD delivery.

6.2.3 On-Set / Near-Set Monitoring

Viewing conditions for HDR monitoring:

While it is difficult to establish parameters for viewing conditions for on-set monitoring, it is advisable to follow the recommendations for setup of an on-set monitor as described in BT.814 [14] or the latest version of an equivalent standard. BT.2100 [5] contains some specifications on reference viewing conditions (e.g. 5 nit surround).

Dynamic range of on-set monitor:

It is recommended to have display devices for on-set monitoring capable of at least 800 nits of peak brightness. Some RGB OLED mastering monitors are capable of 1,000 nits of peak brightness. Note that future HDR content delivered to the consumer may be intended for substantially greater than 1,000 nits peak display brightness.

6.2.4 Color Grading

6.2.4.1 Grading Work Flow

Professional color grading should take place in a controlled environment on a professional monitor whose capability is known, stable and can be used to define the parameters of the ST 2086 [10] Mastering Display Color Volume Metadata.

Current industry workflows such as ACES [50] are recommended to be used where the operator grades behind a rendering and HDR viewing transform, viewing the content much as a consumer would. The work would result in a graded master that, when rendered with the same transformation, will result in a deliverable distribution master. (See also Annex in Section 18.)

6.2.4.2 Grading Room Configuration

When there is a need to prepare both HDR and SDR video productions, which share the same physical environment and require mixing segments of different dynamic range characteristics in the same master, it is important to ensure the use of equivalent illumination levels encountered in a conventional grading environment. This is because it is important to review both the SDR and HDR rendering of the images to guarantee a level of consistency between them. The black and dark tones for both the HDR and SDR video pictures are a particular concern as the ability to discriminate the information they contain is highly affected by viewing conditions.

Secondary monitors should be turned off so as to not impact room brightness during grading or the client viewing monitor must be positioned so as to not impact the bias lighting observed by the colorist.

6.2.4.3 Grading Monitor Configuration

A professional mastering monitor should be specified and configured according to the required deliverable: System Colorimetry, Transfer Matrix, EOTF, White Point and RGB level range (Full or Narrow).

6.2.4.4 Grading System Configuration

As of 2019, no *de facto* standards emerged which define best practices in configuring color grading systems for 2160p/HDR/WCG grading and rendering. A full discussion of this topic

is beyond the scope of this document and it is recommended that the reader consult with the manufacturer of their specific tool. Annex A may provide some insight into this topic.

6.2.5 Channel-based Immersive Audio Post Production

This section describes one example of creating content with channel-based Immersive Audio containing height-based sound elements. This example is for post-produced content (file-based workflow) which can then be used in a real-time program assembly system and distributed in a linear television distribution system. One commercially deployed Immersive Audio system is Dolby Atmos using E-AC-3+JOC [35]. A high-level diagram of a post-production mixing environment for this format is shown below:

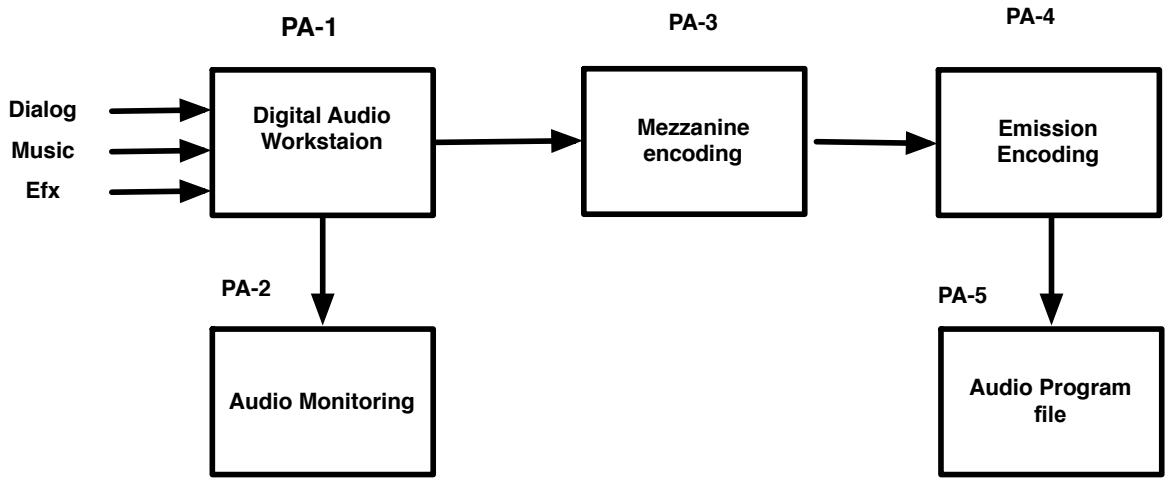


Figure 3 Channel-based Immersive Audio Post-Production

Reference Point	Content Creation Functions	Reference Point Description
PA-1	Digital Audio Workstation	A mixer uses a digital audio workstation to create groups of audio tracks consisting dialog, music and effects. Each group consists of channel-based audio beds (e.g. 5.1, 7.1.4).
PA-2	Audio Monitoring	The Immersive Audio mix is rendered to multiple different speaker layouts including 7.1.4 and 5.1 using hardware film-mastering style tools or commercially available software tools
PA-3	Mezzanine Encoding	Immersive Audio and 5.1 channel-based renders are encapsulated into a mezzanine file.
PA-4	Emission Encoding	Immersive Audio is encoded into formats such as E-AC-3+JOC[12] for final emission delivery
PA-5	Audio Program File	SMPTE ST 337 [36] embedded in a delivery file.

[12] See Section 11.5 regarding backward compatibility.

6.2.5.1 Monitor Environment

Immersive Audio monitoring should take place in a controlled environment consisting of a treated listening room and a calibrated speaker environment. The recommended speaker configuration for immersive mixing and monitoring is 7.1.4 as per Recommendation ITU-R BS.2051[13] (System G minus the Left and Right screen loudspeakers) calibrated to a reference sound pressure level appropriate for the size of the mix room. Typical levels range from 76 dB SPL C-weighted, slow response for a room less than 1,500 cubic feet or smaller (i.e., many outside broadcast trucks) to 85 dB SPL, C-weighted, slow response for rooms 20,000 cubic feet or larger (i.e., many film stages).

6.2.5.2 Immersive Audio Mastering

The channel audio data is recorded during the mastering session to create a 'print master' which includes an immersive audio mix and any other rendered deliverables (such as a 5.1 surround mix). Program loudness is measured and corrected using ITU-R BS.1770-4 [37] methodology.

6.2.6 Additional UHD Technologies beyond Foundation UHD – NGA

Next Generation Audio (NGA) offers new tools for content creators and new experiences for consumers. Immersive audio offers a sense of being completely surrounded by the aural experience: above, below, and all around the listener. NGAs can not only make these experiences possible in a home theater environment with optimally placed speakers, but also via headphones, soundbars, or even sub-optimally placed speakers to an extent. NGA can also allow the consumer to personalize the experience, such as increasing the dialog level relative to the music and effects or choosing a particular dialog track such as alternate language or commentary. See Section 14 for details about MPEG-H (Section 14.2) Dolby AC-4 (Section 14.3) and DTS-UHD (Section 14.4) NGA systems.

Maybe the most important production aspect is that NGA systems bring new paradigms in authoring and mixing the audio content requiring additional training for sound engineers. For delivery of NGA immersive and personalized experiences, authoring of metadata plays an essential role in production. Using plugins for DAW (Digital Audio Workstations) in post-production or metadata authoring and monitoring units for live production, the metadata can be authored together with the audio data delivered to the final audio encoders. In post-production the metadata can be stored as an XML file according to the ITU-R Audio Definition Model (ADM) [72]. It is anticipated that ADM profiles with tailored feature sets will be necessary at each stage of the NGA broadcast and content production chain to ensure full interoperability and quality of service.

6.3 Production for Live Content

The Ultra HD Forum's objectives include understanding the implications of creating and delivering Foundation UHD content all the way through the production chain. It is important that the whole system is understood as technical decisions to deploy a particular Ultra HD production approach upstream may have implications for downstream delivery equipment. The reverse is also true that downstream delivery constraints may affect production techniques.

[13] System G minus the Left Screen and Right Screen loudspeakers would be used for 7.1.4, while System D could be chosen for 5.1.4.

Live content examples include sports, news, award shows, etc. Pre-recorded content is captured and produced in advance (see Section 6.2). Examples include soap operas, sitcoms, drama series, etc. Real-time Program Services may include Live programs and Pre-recorded programs, and Pre-recorded content – such as advertising – may be inserted within Live programs. Real-time Program Services may be delivered via MVPD (Satellite, Cable, or IPTV), OTT and DTT and are delivered on a schedule determined by the service provider.

Unlike Cinema, Ultra HD Blu-ray™ discs or On Demand, Real-time Program Services involve performing complex manipulation of images and audio in real-time at multiple points in the delivery chain, including graphics, captions, virtual studios, Live sound mixing, logos, chroma-keying, DVE moves, transitions between sources, encoding, decoding, transcoding, ad insertion, voice over production, monitoring, conditioning, rendering, closed captioning, standards conversion, etc.

6.3.1 Live Production in Trucks or Studio Galleries

At the live event venue or studio, the action is captured and prepared in near real-time. This section addresses the processes that may be required to assemble the live show prior to compression for distribution to a “headend” or other central facility. Such video processes may include:

- Synchronous live baseband inputs from cameras to production switcher
- Racking of cameras (for chroma/luma/black balance)
- Transitions between cameras (mix, wipe, DVE, etc.)
- Keying (chroma key, linear key, alpha channel, difference key, etc.)
- Overlay of graphics, text, etc.
- Slow motion and freeze frame
- Editing (for action replay/highlights)
- Use of virtual sets

Simultaneous production of both 2160p/HDR and HD/SDR output may be employed, and assets may be shared, e.g., HD graphics or 2160p graphics for both outputs. It is recommended to maintain a single HDR transfer function and system colorimetry. See section 7.2 for details on converting assets from SDR to HDR or vice versa.

Performing the above functions on Foundation UHD content may be different than for HD/SDR content. Foundation UHD content may involve higher spatial resolution, frame rates up to 60p, and HDR/WCG.

- Graphics created for gamma transfer function may need re-mapping for an HDR transfer function depending on creative intent.
- HLG has the characteristic of providing a degree of backward compatibility with legacy SDR devices, both professional and consumer. However, care must be taken to ensure the signal being displayed is encoded with the appropriate system colorimetry preset in the display. For instance, for HLG10, a displayed picture on a legacy display that uses BT.1886 [4] / BT.709 [2] system colorimetry will be incorrect. See Section 10.4 for details and caveats.

6.3.2 Production with encodings other than PQ and HLG

For episodic television, as with features, colorists typically work in an underlying grading space adopted by the project. Usually, this grading space comprises a log encoding, often with

a color gamut native to a principle camera. The content master is created and (ideally) preserved in this project grading space. Conversion to a distribution encoding, e.g., PQ, is performed by an output viewing transform and viewed on a specific mastering display during grading, whereby the grade targets a particular peak luminance.

Similarly, live productions establish a project grading space, often choosing one that is optimized to their cameras and other equipment (e.g., SLog3 with BT.2020 for Sony equipment). Grading and mixing occurs within that selected space.

When contributions having different transfer functions and/or color gamuts are to be combined, there is a concern about cumulative quantization effects. One way to address this is to defer conversion away from the native system colorimetry as long as possible. A more general solution is to transform the various contributions to a common linear encoding as a grading space. In either case, the produced result is converted to the required delivery format, suffering the quantization of conversion only once.

In either case, a PQ master can be obtained with a corresponding output viewing transformation. Likewise, to obtain an HLG master, an appropriate output viewing transformation is used. When both PQ and HLG masters target the same peak brightness, they are typically expected to exactly match each other when viewed on the same mastering display in the respective corresponding mode. In some practices, if the PQ and HLG masters are checked in this way and they do not appear identical to each other, the overall grade is rejected.

Other workflows may grade in a distribution space, e.g., PQ or HLG, which provides one distribution master directly. Other distribution masters are obtained by performing a conversion to the target space, however such conversions are sometimes difficult to match well. Further, some distribution spaces impose limitations that compromise whether a content archive is future proof.

6.3.3 Channel-based Immersive Audio Production

For a live event, Immersive Audio can be created using existing mixing consoles and microphones used during the event. Using additional console busses, height-based ambience and effects can be added to a traditional 5.1 or 7.1 channel mix which can then be encoded by an ETSI TS 103 420 compliant E-AC-3 encoder. A local confidence decoder can be employed to check typical downmixes, including the backwards compatible 5.1 channel render described in ETSI TS 103 420. During normal mixing, this confidence decoder can serve as a useful continuity check and display (i.e. to make sure the mix is still “on-air”), though due to normal latencies it will likely be found to be impractical to be kept in the monitor path full time. A high-level diagram of a live mixing environment recently used at a major televised event using Dolby Atmos using E-AC-3+JOC [35] is shown in Figure 4 below.

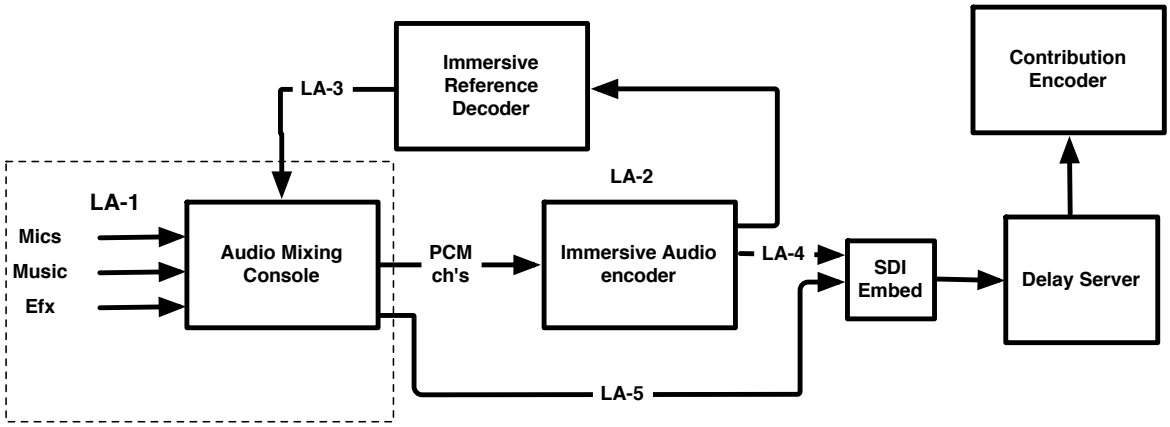


Figure 4 Channel-based Immersive Audio Live Production

Reference Point	Content Creation Functions	Reference Point Description
LA-1	Audio Capture and mixing	Microphones are placed throughout the live venue. Microphone feeds are brought into the audio mixing console in the same fashion as 5.1 production. Sound elements are mixed into a final immersive program.
LA-2	Immersive Audio Authoring	Audio groups are created in the mixing console representing the immersive mix as channel-based (e.g. 5.1.2, 5.1.4, 7.1.4) audio[14].
LA-3	Immersive Audio Monitoring	The Next-Generation Audio Processor renders the audio to 5.1.2, 5.1.4 or 7.1.4 speaker feeds and sends these feeds back to the audio mixing console for monitoring in the live mix environment.
LA-4	Immersive Audio Encoding	Atmos immersive program, encoded as E-AC-3+JOC [35] is delivered as a 5.1 channel bitstream + parametric side-data (steering data) to the contribution encoder[15] transported over either a MADI or SDI link.
LA-5	Legacy Audio Delivery	Stereo or 5.1 complete mixes may be created at the audio mixing console and delivered via traditional means.

In this channel-based immersive audio example, a Dolby Atmos enabled E-AC-3+JOC [35] encoder generates a compressed bitstream containing 5.1 channels of audio that is backwards compatible with legacy E-AC-3 decoders. In parallel, the encoder generates an additional set of parameters (specified in ETSI TS 103 420 [35]) that are carried in the bitstream, along with the 5.1 audio, for use by a Dolby Atmos E-AC-3+JOC [35] decoder. The full Atmos decode process reconstructs the original channel-based immersive audio source from the 5.1 backwards compatible representation and the additional parameters. A typical channel-based

[14] The downstream Atmos Channel-Based Immersive emissions encoder, using E-AC-3 + JOC [35] will render a legacy 5.1 audio program. It is recommended to verify the rendered 5.1 audio program using a suitable E-AC-3 decoder in the monitor chain.

[15] See Section 11.5 regarding backward compatibility.

immersive program can be encoded, for example, at 384-640 kbps total, thus fitting into existing emission scenarios.

6.3.4 Additional UHD Technologies beyond Foundation UHD – NGA

A key consideration of implementing NGA for live production is the new requirements for dynamic metadata authoring and mixing the audio content requires additional training for sound engineers. For delivery of NGA immersive and personalized experiences, authoring and carriage of dynamic metadata plays an essential role in production.

Live productions using 3G or 12G SDI interfaces with a minimum of 16 PCM audio tracks are sufficient for delivery of immersive audio programs, with 4 height channels and several commentary tracks as objects. Dynamic audio objects can be added to the program mix, when the production tools have features such as dynamic spatial panning. Additional mechanisms for efficient and secured metadata delivered over an SDI interface are provided for NGA systems. According to SMPTE, this system will be described in ST 2109 Audio Metadata over AES3 (pending publication by the SMPTE), and there is work to extend this to SMPTE ST 2110-40 [47] (metadata over IP).

From the many trials conducted in France, Germany and U.S. and the commercial broadcasts in South Korea using NGA, it has been found that integration of authoring and monitoring units in common SDI infrastructures (including remote facilities) is straightforward and sound engineers become familiar with mixing immersive content and authoring metadata relatively quickly. For example, Korean broadcaster SBS produced the 2018 Football World Cup using MPEG-H with immersive and interactive audio services (Korean commentary, English commentary, and stadium atmosphere alone). Additional trials were conducted in 2018 using object-based audio and interactive audio services (including separate language tracks in some cases): by NBC during Winter Olympics using E-AC-3 plus JOC [35], by France Television during Roland Garros Tennis Open using both MPEG-H [70] and AC-4, [65] as well as by the EBU during the European Athletics Championships using both MPEG-H and AC-4.

7. Security

7.1 Content Encryption

Digital content has always been exposed to illegal reproduction and illegal distribution. Various content protection technologies have been developed, involving different scrambling algorithms. Some of these algorithms, still in use, were designed more than 20 years ago and are no longer resistant to sophisticated attacks.

As a general rule, it should be considered that scrambling algorithms with key size less than 64 bits are not resistant to sophisticated attacks; in fact, the time needed to succeed in such attacks is measured in seconds not hours.

A well-known algorithm is Data Encryption Standard (DES), designed in the 1970's, and referred as FIPS 46-3. It was withdrawn from the list of algorithms recommended by the US National Institute of Standards and Technologies (NIST) in May 2005.

Another well-known algorithm is DVB CSA, approved by DVB in 1994, and described in ETSI ETR 289. Its design was initially secret but was subsequently disclosed in 2002. Since then, many publications describe various attacks. In particular, the publication "Breaking DVB-CSA", by Tews, Wälde and Weiner, Technische Universität Darmstadt, 2011 describes an implementation of such attack. This paper reveals that with very reasonable investment, DVB-CSA with a 48-bit key size (also known as DVB-CSA1) could be reversed in real-time.

Fortunately, DVB-CSA with a 64-bit key size (also known as DVB-CSA2) is still resistant against these attacks and is likely to remain so for another few years.

Content protection technologies are now using more recent scrambling algorithms with a larger key size, usually a minimum of 128 bits.

The algorithms that have been standardized for protection of digital content are:

- AES, Advanced Encryption Standard, published in FIPS-197, NIST, 2001,
- DVB-CSA3 published in 2011 by DVB in ETSI TS 100 289 V 1.1.1,
- DVB-CISSA published in 2013 by DVB and described in ETSI TS 103 127 V1.1.1

The Ultra HD Forum recommends the following regarding content security:

- UHD content should not be scrambled with DVB-CSA1, nor with DES scrambling algorithms.
- UHD content should be scrambled with AES or DVB-CSA3, using a minimum of 128 bits key size.
- DVB-compliant service providers should use DVB-CSA3 or DVB-CISSA when transmitting Live or linear UHD content.
- In the case where DVB-CSA3 is still not deployed, it is acceptable to use DVB-CSA2 with a crypto-period between 10 and 30 seconds, during the time needed to upgrade the equipment to DVB-CSA3.

7.2 Forensic Watermarking

7.2.1 Introduction

Forensic Watermarking complements content protection technologies such as Digital Rights Management (DRM) and Conditional Access Systems (CAS) by providing a means to deter piracy. In this case, Forensic Watermarking is used to generate individualized copies of a video asset thereby allowing the recovery of forensic information to address security breaches. For instance, such forensic information could identify the user receiving the content or the device receiving or displaying the content, its model ID, or other information that can help identify a piracy source.

A watermarking technology used for Forensic Watermarking is characterized by a number of properties, including:

- Fidelity: the modifications made to the content to introduce the Watermark Identifier shall remain imperceptible for a human being;
- Robustness: the Watermark Identifier shall be detectable after post-processing operations that may be performed by pirates, e.g., re-compression, camcording, screencasting, etc.
- Payload length: the size of the Watermark Identifier expressed in number of bits that can be inserted.
- Granularity: the duration of multimedia content that is needed for detecting a Watermark Identifier, usually expressed in seconds and dependent on the payload length.
- Security: the Watermark Identifier shall withstand targeted attacks from adversaries that know the Forensic Watermarking technology in use as well as its technical details.

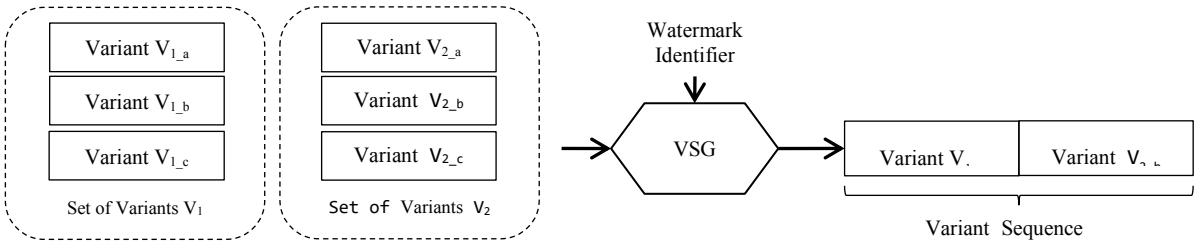


Figure 5 Illustration of Watermark Identifier

The following sections provide:

- Information about watermarking terminology, use cases and applications.
- Consideration for performance, feasibility, privacy, and other aspects.
- High level information flow for some use cases.
- Communication specification for integration of Forensic Watermarking.

7.2.2 Use Cases

Forensic Watermarking is routinely used nowadays in professional environments, e.g., for distributing movies prior to theatrical releases and in Digital Cinemas. In 2014, MovieLabs released a specification [39] that mandates the use of Forensic Watermarking on UHD content. In this context, it is currently being considered at different stages of the content distribution

pipeline. This section provides a comprehensive overview of these different stages and how different watermarks contribute to the forensic analysis. However, the remaining sections will only focus on the use case "Forensic Watermarking Identifying Consumers Devices."

7.2.2.1 Forensic Watermark Identifying Operators

It is common practice for content owners to incorporate a master Watermark Identifier in the copy of the movies that they are shipping to operators. It provides them with means to track which resellers have been pirated in order to ask them to adopt relevant security countermeasures. Such master Watermark Identifiers are embedded in very high-quality content and fidelity is therefore of paramount importance. In this case, fast turnaround processing time is not as critical as in other application use cases. The watermark embedding process can be performed offline and watermark detection results can be returned after a few days.

7.2.2.2 Forensic Watermark Identifying Redistribution Networks

An operator may distribute content through several channels: satellite, terrestrial broadcast, over-the-top, etc. These redistribution channels can be part of the operator's own redistribution network. They may also be operated by some affiliate networks, and in such a situation, when piracy is detected to originate from an operator, there is some uncertainty about the source of the leak. To disambiguate the situation, alternate Watermark Identifiers can be embedded for the different redistribution channels. In this context, fast detection is usually not required. On the other hand, watermark embedding may need to be performed on-the-fly to accommodate redistribution of some video content, e.g., live broadcast.

7.2.2.3 Forensic Watermark Identifying Consumers Devices

The MovieLabs specification [39] requires identification to stop leakage on the level of device or device type. This can provide means to analyze piracy at the finest granularity in order to deploy targeted anti-piracy measures accordingly. Based on this requirement for premium content, it is likely that Forensic Watermarking could be useful to help deter piracy of live events such as sports and music concerts.

The live service application use case has stronger requirements, as the watermark embedding operation must not introduce significant delay in the content transmission chain. In addition, it may be even more critical to quickly identify the source of an illegal retransmission in order to stop it, ideally before the end of the pirated live program such as a sporting event. On the other hand, the watermark fidelity constraint might be relaxed on some occasions to speed up the detection process. Finally, while it is important to secure the Watermark Identifiers against malicious attacks, the attacker will have more time to execute attacks on VOD content than live content, which quickly expires in value even in pirated form.

7.2.3 Distribution

For Forensic Watermarking that is identifying consumers' devices, different methods of media distribution may result in different workflow and process optimizations.

- Physical and Broadcast Distribution: The content is distributed either via physical media such as a Blu-ray disc or via broadcast.
- File Distribution: A single file is made available for playback, used in environments with guaranteed bandwidth streaming, or (progressive) download.
- ABR Streaming: Streaming without guaranteed bandwidth but adaptation to the available bandwidth is done based on several available bandwidth options made available in a manifest or playlist.

The distribution mechanism has impacts on the integration of the watermarking system. Sections 7.2.4 and 7.2.5 present in detail two major approaches, namely those that operate in a single step (one-step) and those that require a preprocessing step (two-step). In addition to the distribution mechanism, the selection of one-step versus two-step watermarking is guided by other aspects such as the integration complexity, the availability of client-side marking capabilities, the ability to modify components on the head-end side, etc.

7.2.4 One-Step Watermarking

This approach to create forensically watermarked content is to mark decompressed baseband video content in a single step:

- During encode on the server, or
- During playback in the secure video pipeline.

One-step watermarking on the distribution server side requires delivering individualized watermarked content to every device requesting it. As such, it may not be suitable for serving a large number of recipients due to scalability constraints when encoding a stream for each recipient.

One-step watermarking is therefore usually applied on the client side, where the watermark is enabled. It typically involves communication between the Conditional Access (CA) and the watermarking (WM) modules as shown in Figure 5.

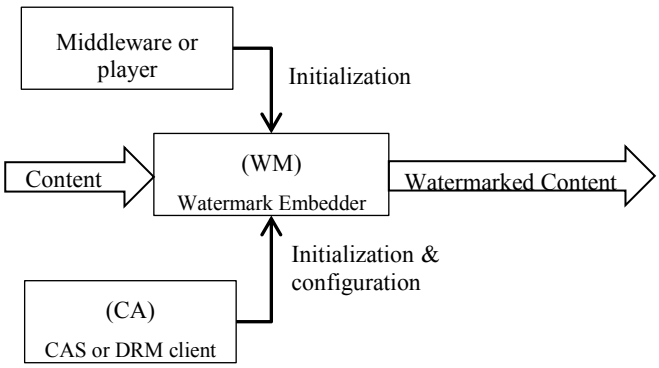


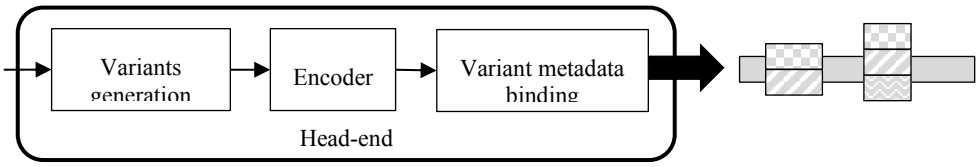
Figure 6 One-Step Watermark Performed on the Client Side

7.2.5 Two-Step Watermarking Integration

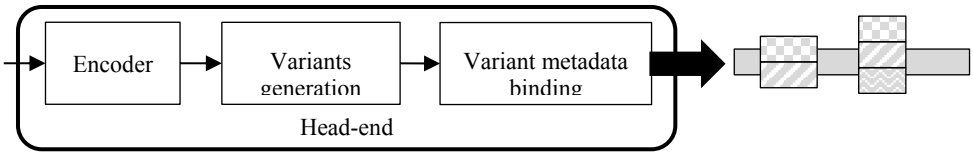
Two-step watermark integration requires (i) integration of the preprocessing step at the head-end, (ii) integration of the Individualization step somewhere in the distribution pipeline, and (iii) defining a mechanism to transport the Variant metadata along the video.

7.2.5.1 Step One: Content Variants Preparation

The generation of Variants is performed at the head-end so that distributed video can be forensically watermarked easily further along the distribution pipeline.



(a) Variant generation in the baseband domain



(b) Variant generation in the encoded domain

Figure 7 Two Examples of Two-step Watermarking Systems

The resulting integration with encoders will depend on whether the generation of Variants is performed in the baseband domain or in the encoded domain.

- **Variant generation in the baseband domain:** In this case, the preprocessing module is fed with baseband video content in input and generates uncompressed Variants. The resulting Sets of Variants are then forwarded to the encoder to produce encoded Variants.
- **Variant generation in the encoded domain:** In this case, the preprocessing module is fed with encoded video content produced by the encoder and generates Variants of encoded content. Such a preprocessing module can operate independently of the encoder, however, for latency-constrained applications (e.g., live broadcast). it is desirable to have a more intimate integration to minimize processing delay.

7.2.5.2 Step Two: Individualization

Further along the video distribution pipeline, an individualization agent produces the forensically watermarked videos. This individualization agent has typically access to a WM ID and comprises:

- A Variant Sequence Generator (VSG) that receives Sets of Variants in a relevant container, selects a single Variant from each Set of Variants, and thus produces a Variant Sequence that encodes the desired WM ID, and
- A Content Assembler that merges the Variant Sequence into the compressed video bitstream in order to obtain the forensically watermarked video encoding the desired WM ID. If Variants are associated to independent encryption blocks, the Assembler may operate in the encrypted domain without knowledge of the decryption key.

The resulting forensically watermarked video can be decoded for rendering, stored for later use (PVR) or forwarded further along the transmission pipeline.

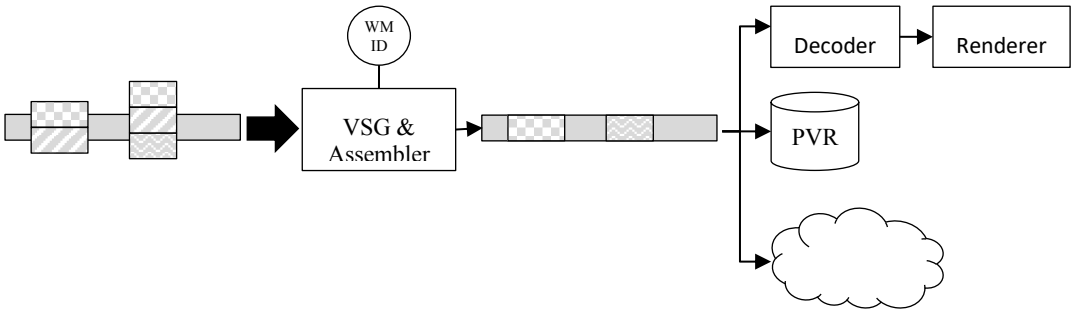


Figure 8 Watermark Embedding Using a Unique Variant Sequence

Deployment scenarios differ with respect to the time and location of the individualization agent, which themselves are dependent on integration preferences and network types. Individualization can be performed on the server side. It creates requirements on the server such as VOD server or CDN edge, but it is transparent to the client. Alternatively, individualization can be done on the client side before decoding:

- **Just-in-Time Server Individualization:** The delivery server hosts the Sets of Variants computed in the preprocessing step and possibly (parts of) the original content. When a client requests the content, the server identifies the session or client using known authentication methods such as HTTP cookies, tokens or authenticated sessions and associates a WM ID to the request. The individualization agent can then operate and deliver forensically watermarked video to the requesting client.
  The storage overhead on the server amounts to the Variants data. The smaller is the Variants data compared to the original content, the less overhead is induced by the Forensic Watermarking technology.
- **Prepared Server Individualization:** In this case, the delivery server hosts pre-individualized video segments to lower the complexity of delivering forensically watermarked content.
  In the most extreme case, the whole content is individualized using a pool of WM IDs resulting in a collection of forensically watermarked contents that are placed in a stack. When a client requests the content, the server delivers the next pre-watermarked content in the stack and records in a database the link between the session and the WM ID. This approach does not scale well though as it induces significant storage and caching overhead.
  In practice, it is usually more efficient to pre-individualize video content at a finer granularity. For instance, a watermark vendor may use several Sets of Variants to encode a single WM ID symbol. The baseline idea is to then pre-compute alternate versions of a video segment, each version encoding one of the possible symbols. The delivery server then only has to forward one of these pre-computed versions for each video segment based on the client's WM ID.
  When using a binary alphabet, this solution amounts to having two versions of the content being stored and transmitted downstream. The boundaries of the pre-computed video segments can be made to align with the boundaries of the transmitted video packets, e.g. a MPEG TS packet or an ABR chunk. This further lower the complexity of the delivery server. A typical example is A/B watermarking routinely used in ABR and described in detail in Section 7.2.6.
- Client Device Individualization: In some application uses cases, individual stream delivery is not feasible, e.g. one-way distribution such as broadcast or physical media. In this case, the individualization process needs to be performed on the client

side.

The client device therefore receives the Variant metadata along with the encoded video content. Using the Watermark Identifier of the device, the VSG produces a Variant Sequence that is then used by the Assembler to produce a watermarked video bitstream to be forwarded to the video decoder. To secure the watermarking operation, the individualization process shall be performed in a secure environment e.g. along the secure video path under the TEE control. Another security mechanism includes controlling access to Variants with decryption keys.

This approach is useful for scenarios where individual stream delivery is not feasible, such as one-way distribution, including physical media.

7.2.5.3 Variants Transmission Mechanisms

The steps of pre-processing and watermark embedding operations are unlikely to be co-located. It is therefore necessary to define a mechanism to transport the Sets of Variants along the content to be watermarked. A first challenge to address is the temporal synchronization between the video content and the Variants metadata e.g. to support skip modes. Depending on the implementations, the transport mechanism may also define which portion of the Watermark Identifier applies to a given Set of Variants:

Transport at the Media Layer

Variants metadata can be incorporated at the media layer. For instance, MPEG standards indicate that any proprietary information can be placed in the video bitstream as dedicated NALUs referred to as Supplemental Enhancement Information (SEI). Variants SEI NALUs can be signaled using a dedicated identifier [25], [26].

The advantage of such a low-level transport mechanism is twofold. First, the Variants metadata is finely interleaved with media essence, thereby providing the necessary temporal synchronization between the content and the Variants metadata. Second, Variants metadata can inherit the protection provided by CAS or DRM systems. On the other hand, such low level signaling may induce integration along the secure video path and/or the trusted execution environment, which requires collaboration from chip vendors for adoption.

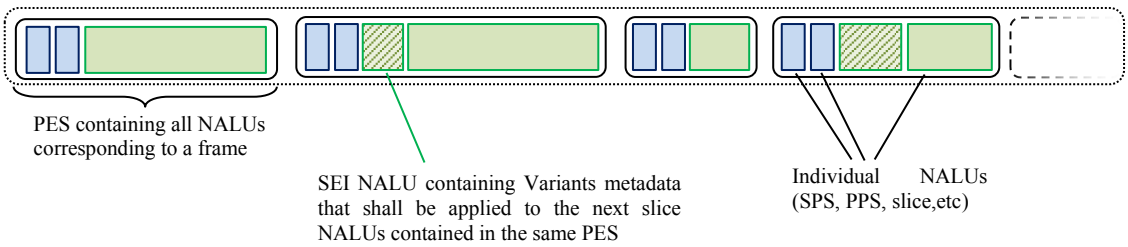


Figure 9 Transport at the Media Layer Using MPEG SEI NALUs

Transport at the Container Layer

Variants metadata can be transmitted in an alternate transmission channel next to the video content at the container layer. For instance, Variants metadata could be placed within a MPEG2-TS bitstream using a dedicated PID and using the PCR to synchronize the video and metadata channels. Alternately, Variants metadata could be incorporated as an extra track in an ISOBMFF file. In that case, synchronization can be achieved by aligning samples across different tracks. When Variants metadata is handled as a component separate from the video, proper care shall be taken to guarantee its protection if needed with relevant content protection techniques.

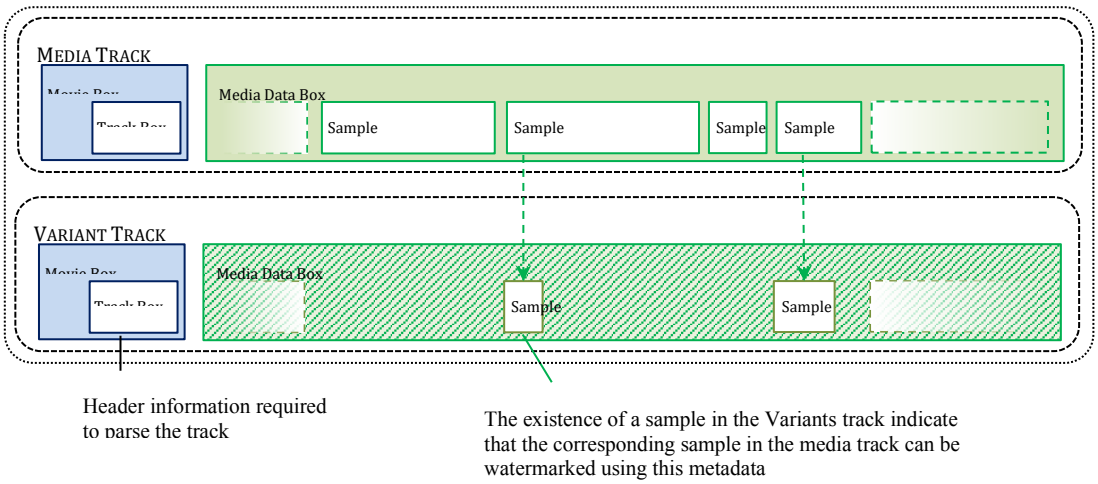


Figure 10 Transport at the Container Layer Using a Track in an ISOBMFF File

An example of how to transmit Variants metadata as an extra track in an ISOBMFF file is described in [40][16]. This standard applies to file-based delivery, e.g. on physical media with embedding on the client side. The baseline principle is to define a dedicated track in an ISOBMFF file that describes how to construct watermarked video samples. For instance, a constructor for a sample indicates which portion of the video track shall be kept and which portions shall be replaced by a Variant available in the variant track. Access to the MPEG variant constructors is subject to cryptographic keys. Different users/devices will have a different set of keys and thereby been able only to produce different watermarked video samples using different constructors. Moreover, the Variants are double-encrypted to serve as an anti-bypass mechanism. A player that would not perform the watermark embedding operation would not be able to display a good-quality video since some segments of the video would still be encrypted. The strong link between encryption and watermarking requires collaboration between CAS/DRM and watermarking systems e.g. for key management and provisioning. The virtue of this design is that it enables a secure integration of the watermark embedding module on open devices outside of the secure video path or trusted execution environment.

Note: The embed location may be communicated out of band, in a common format or can be embedding as extra data during the encoder and is then removed during use by the packetizer, e.g. using a format like EBP [49].

Out-of-Band Transport

Transporting Variants metadata at the media as well as container layer requires some level of video parsing. However, in some integration scenarios, such a requirement may not be acceptable, e.g. for scalability reasons or because parsing is impossible at the integration point due to encryption. In such a situation, it is desirable to have a direct access to the Variants metadata either with a separate file containing the metadata itself or with a separate file containing relevant pre-computed parsing information to access the Variants metadata directly

[16] It shall be noted that the terminology “variants” is slightly different in the MPEG standard and these UHD guidelines. In the MPEG standard, a variant is a full MPEG sample composed of parts of the original bitstream and parts of the Variants, as defined in this document i.e. segments of bitstream that can be used interchangeably at a given location in the bitstream.

in the video file without performing the actual parsing. This is typically the case on a CDN Edge server.

A CDN Edge server is serving video chunks on request to clients, among other things. To guarantee the scalability of the service, it is of the utmost importance to reduce parsing to a minimum. A possible solution is to store Sets of Variants in the container in such a way that one of the Variants can be selected simply by skipping the data of the other Variants. The indices of where the Variants are located in the stream can then be stored in a separate file which can be loaded in memory by the CDN edge server. As a result, the server can perform the watermark embedding operation directly without any video parsing.

Another use case of the out-of-band transport is when the video stream is being broadcasted but the CDN Edge server is not modifiable for the watermark embedding operation. In such scenario, broadcaster can supply two streams, the video stream and the Variants metadata stream, to the client devices using the conventional CDN. The VSG at the client side can directly access to the broadcasted Variants and can perform the client device individualization. This approach demands additional bandwidth but is applicable when the service provider is not able to control network configurations between streaming servers and end-user clients.

7.2.6 Use Case: ABR VOD

The characteristics of this use case requires that segment boundaries are known before the start of the watermark processing.

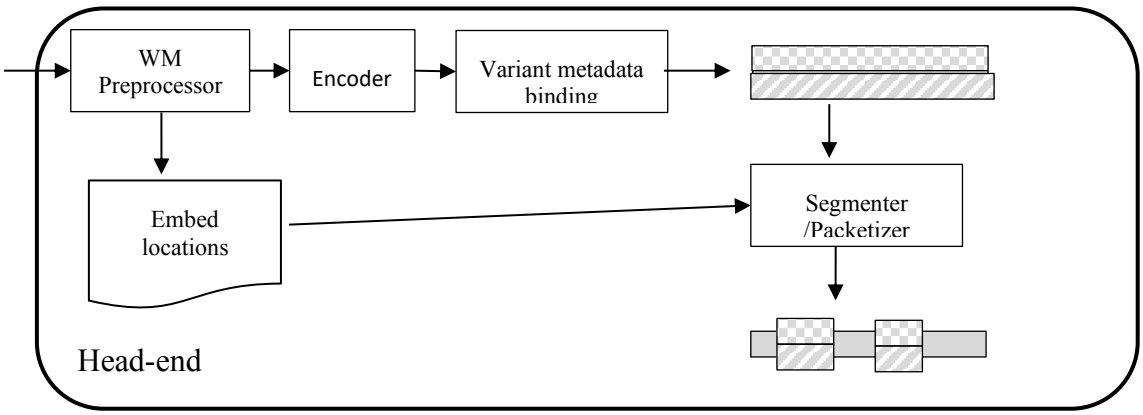


Figure 11 Pre-processing in the Baseband Domain for Two-step Watermarking

In one scenario, video segments are pre-serialized to lower the complexity of delivering forensically watermarked content. The baseline idea is to prepare serialized versions of video segments comprising Sets of Variants for a single WM ID symbol. A typical example is for instance to associate all the Sets of Variants for an ABR fragment to a single WM ID bit. The server can then host two versions of the ABR fragments, corresponding to the WM ID bit equal to 0 or 1. When a client requests an ABR fragment, the server only has to deliver the version of the fragment corresponding to the WM ID of the client.

In the most extreme case, the whole content is pre-serialized using a pool of WM IDs resulting in a collection of forensically watermarked contents that are placed in a stack. When a client requests the content, the server delivers the next pre-watermarked content in the stack and records in a database the link between the session and the WM ID.

This strategy implies duplication of content at a bigger granularity than in the previous case. The resulting storage is therefore usually more significant and since files are individual per user, caching is not applicable.

In another more scalable approach, server hosts pre-serialized ABR fragments. However, the server has no notion of session and is therefore unable to individualize content delivery based on some WM ID.

Forensic Watermarking is achieved by providing individualized manifests/playlists to requesting clients. In ABR, the manifest essentially declares the list of addresses where video fragments can be acquired. The playlist server therefore personalizes the manifest so that each client only 'sees' the video fragments that encode its WM ID. When consuming the playlist, the client requests pre-serialized ABR fragments (encoding its WM ID) that are served by the streaming server.

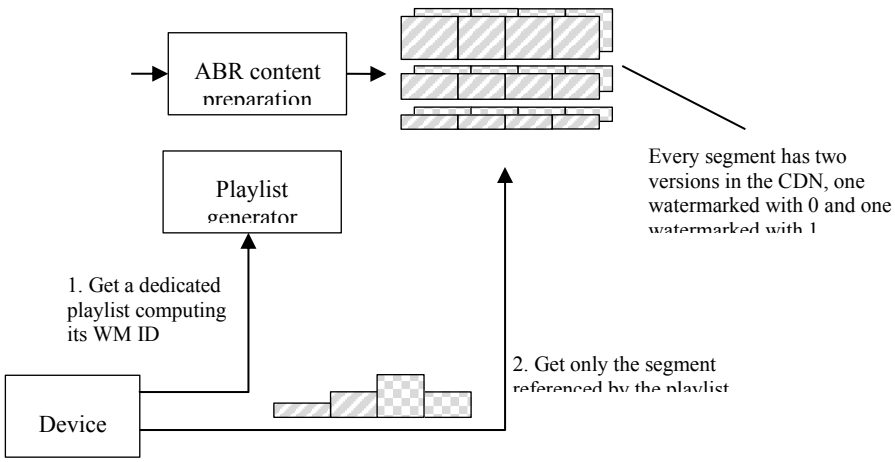


Figure 12 ABR Playlist Serialization

8. Real-time Program Service Assembly

Real-time Program Services consist of a linear, pre-scheduled stream of content that is assembled in real-time for distribution to consumers such as a broadcast television channel, a cable network, etc. Real-time Program Services are comprised of Live and/or Pre-recorded content and may also include graphic overlays, such as station logos, emergency text crawls, etc.

A Real-time Program Service may be assembled at one or more locations. At Production, graphics and slow-motion replays may be added to Live content. At the Broadcast Center (see Table 11) and service provider, interstitials, logos, news or emergency alert crawls, etc. may be combined to produce the final product delivered to the consumer. It is also possible that assembly can occur at the consumer device, such as for targeted advertisements and rendering closed captions.

8.1 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[17]. 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 remap SDR/BT.709 content into HLG10 or PQ10, to up-convert SDR/BT.709 content to HLG10 or PQ10 and vice versa, and to convert content from PQ10 to HLG10 or vice versa. The following subsections offer guidelines for normalizing content in the headend for this purpose. See also Section 10.4 for conversion possibilities in consumer equipment for backward compatibility.

8.2 Conversion from SDR/BT.709 to PQ10/HLG10

In multiple areas of the production chain, it is anticipated that a requirement exists to “mix” or “switch” SDR/BT.709 sources and PQ10/HLG10 sources when constructing a Real-time Program Service. Mixing of this kind may take many forms as it does today in SDR only environments (practicing BT.1886 [4] gamma and BT.709 [2] system colorimetry). To make it possible to combine such elements, SDR/BT.709 sources ***must*** be converted into the PQ10/HLG10 domain with respect to both an industry compliant transfer function (e.g., HLG

[17] See “Guidelines for Implementation: DASH-IF Interoperability Points v4.3”, Section 6.2.5, found at https://dashif.org/docs/DASH-IF-IOP-v4.3.pdf for reference.

or PQ per BT.2100 [5]) and system colorimetry (i.e., BT.709 [2] primaries "mapped" into the BT.2020 [3] container). Such conversions can utilize:

- Direct Mapping: SDR/BT.709 content is decoded and repackaged as PQ10 or HLG10 containers, but while changing the system colorimetry, remapping does not change the color gamut or the dynamic range of the content; the content is simply mapped across to the equivalent color and brightness values.
    - When remapping SDR to HDR for PQ and HLG, the level of reference white should be considered. Reference white in SDR content is typically about 642.5mV (92% of SDR Peak white level. 700mV) leaving very little headroom for speculars. In the case of HLG, the BBC and NHK recommend the reference level for HDR graphics white (aka "reference white") be set to 75% (equivalent to 203 cd/m^2 on a 1,000 cd/m^2 reference display, or 343 cd/m^2 on a 2000 cd/m^2 reference display). This was chosen as it leaves sufficient headroom for "specular highlights" and allows comfortable viewing when HLG content is shown on HDR/WCG and SDR/WCG displays.
- Inverse Tone Mapping: SDR/BT.709 is decoded and then enhanced/modified to emulate PQ10/HLG10 and repackaged as above. Blind (not supervised) ITM (up-mapping) can lead to undesirable results depending on the technology so care should be used when converting SDR to HDR. Conversion algorithms or equipment should be carefully evaluated prior to their use. ITU-R BT.2446-0 describes two techniques for SDR to HDR conversion.

Each method has particular advantages and disadvantages, and one or the other may be preferable under different circumstances. If the service provider intends to simulcast the Foundation UHD service in SDR/BT.709 for backward compatibility, then remapping may be preferable, because content that was originally SDR/BT.709 will remain exactly as intended in the legacy version of the service. Conversely, the service provider may prefer that all the segments of the Real-time Program Service look as uniform as possible to ensure a consistent consumer experience, and thus up-mapping may be appropriate. For example, up-mapping may be preferred for Live production mixing SDR/BT.709 and PQ10/HLG10 cameras, high-quality SDR content, and legacy content like advertisements.

Under some circumstances, (e.g. viewing in a darkened room) HDR displays and content can cause discomfort if consideration is not given to the viewer's vision adapting to the average light level on the screen at any given moment. For instance, if a feature program has a low average light level such as a night scene and the picture is abruptly cut to a high average luminance scene in an interstitial, the viewer may experience discomfort similar to that experienced with changing light conditions in nature. When advertisements are inserted into content, consideration should be given with respect to transitions from dark to light.

The described SDR to HDR conversions typically will be performed by dedicated devices using a combination of 1D and 3D LUTs or other appropriate algorithms or technology. Devices of this type may be used for both SDR/BT.709 to PQ10/HLG10 or vice versa as the production and the capability of the equipment in use requires.

A real-time dedicated conversion device is essential for some use cases, which may be encountered throughout the production chain, such as:

- Mix of SDR/BT.709 and PQ10/HLG10 live sources
  Broadcasting live events (typically sports) in PQ10/HLG10 may require a relatively high number of cameras and it is probable that all these cameras will not be HDR-capable. In that situation, SDR/BT.709 cameras can be utilized if the conversion

process is implemented either at the output of the SDR/BT.709 camera or at the input of the mixer.

- SDR/BT.709 interstitials in a PQ10/HLG10 program
  With SDR/BT.709 interstitials, the interstitial content will likely come from a playout server. In this case the conversion process has to be implemented either at ingest, at the output of the playout server, or at the input of the mixer.
- Use of SDR/BT.709 content
  Extensive libraries of SDR/BT.709 content may be used, for example a live sports match production that includes archive footage from previous matches; such material needs to be converted to PQ10/HLG10 (using a near real-time file-to-file converter) before entering the video pipeline

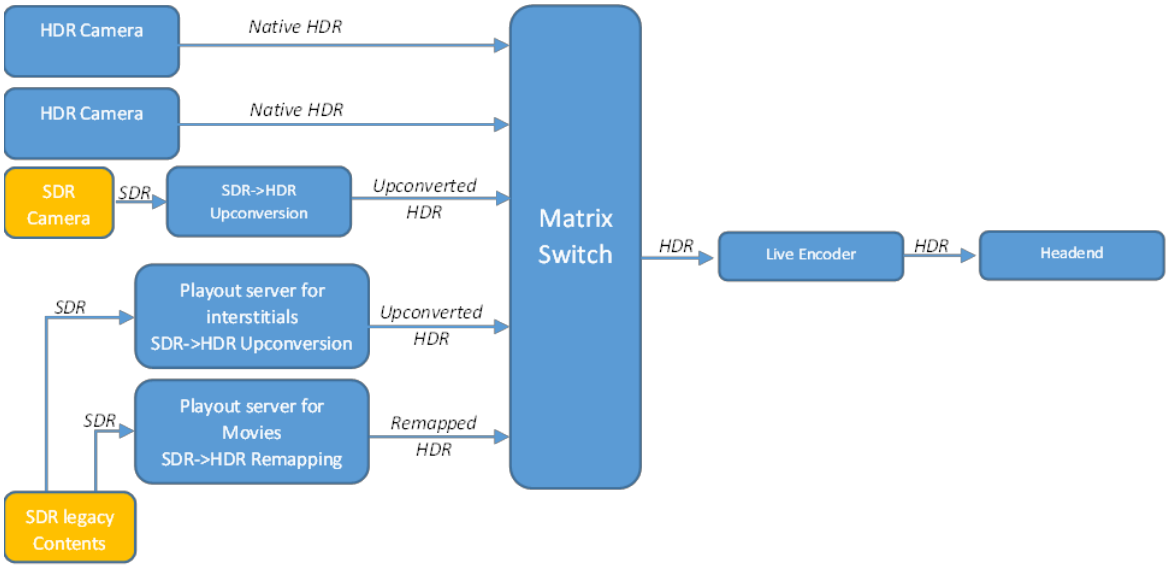


Figure 13 Sample Live Workflow with Mixed Format Source Content

Converting content to a common set of HDR and WCG technologies may occur at different points in the workflow from Production to the Consumer premises.

At the Production stage, multiple versions of the content may be produced. In the case of Pre-recorded content, producers have the opportunity to create multiple versions of the content applying creative judgment to each version. This is recommended for all Pre-recorded content and especially for interstitial material that may be inserted into a variety of Real-time Program Services, which may be operating with different HDR/WCG configurations. Live content may also be output in multiple formats; however, time constraints may prevent highly detailed artistic input. Automated conversion technologies can be "tuned" by the content creator to make the conversion the best it can be for the given program.

At the Broadcast Center or service provider stage, content may be converted to the provider's chosen HDR/WCG configuration using automated tools. This may not include program-specific creative input; however, professional equipment may produce results that the service provider deems acceptable.

At the consumer premises, color volume transform may be possible for the purpose of backward compatibility (with limitations, see Sections 10 and 10.4). Also, display devices may "map" content internally to best suit the display characteristics. Both of those processes operate on a content stream with one, constant dynamic range and system colorimetry. Real-time Program Service providers should not expect consumer equipment to render seamless

transitions between segments of content that have different transfer function or system colorimetry.

8.3 Conversion between Transfer Functions

A receiver may not be able to switch seamlessly between HDR transfer functions; therefore, it is recommended that only one transfer function be used for a given Real-time Program Service in Foundation UHD. This section offers guidelines to service providers to convert HDR content from PQ10 to HLG10 or vice versa in order to maintain a single transfer function in the service. Equations and best-practices for these conversions can be found in ITU-R BT.2390 [6].

8.3.1.1 PQ10 to HLG10

It is possible that Pre-recorded programs are produced in PQ10 in Foundation UHD; however, some service providers may favor HLG10 delivery for its degree of backward compatibility. It is possible to perform the required conversion with a LUT. As the PQ program may have pixel values as high as 10,000 nits, some color volume transform of highlights may occur in content converted from PQ10 to HLG10 and then rendered on a lower peak luminance display.

Note that in an HEVC program stream, it is possible to deliver PQ-related metadata (ST 2086, CLL) in the SEI even when HLG is signaled as the transfer function. At this time, that practice is not referenced in any current standard, thus it should be avoided or employed with caution.

8.3.1.2 HLG10 to PQ10

It is possible that Live programs are produced in HLG10 in Foundation UHD; however, some service providers may favor PQ10 delivery making it necessary to convert HLG10 encoded content into PQ10. This conversion can be accomplished with a static or programmable LUT prior to encoding. It must be considered that this conversion involves translation from a scene-referred relative signal to a display-referred absolute signal. As such, a target peak luminance needs to be preset as part of the LUT conversion function. At the time of this writing, grading monitors typically have a peak brightness of 1,000 nits so that value may be used; i.e., the maximum value of the HLG signal would map to the PQ value of 1,000 nits.

8.4 Conversion from PQ10/HLG10 to SDR/BT.709

This operation is needed for PQ10 and HLG10. This method may be employed at the headend prior to final distribution to provide a backward compatible feed to legacy 2160p/SDR/BT.709 TVs[18] and legacy HD networks. (See also Section 10.4 on Format Interoperability including STB conversions.)

It is possible to do this conversion using a 3D LUT mechanism as described in Section 8.2. Another method is an invertible down-mapping process described in ETSI TS 103 433 [33], in which HDR/WCG content is down-mapped in real time to SDR/BT.709 at or prior to the

[18] Although HLG offers a degree of backward compatibility with SDR/WCG displays, there is no evidence that HLG offers backwards compatibility for BT.709 [2] displays without additional processing as described in this section. However, both DVB and ARIB require support of BT.2020 system colorimetry (and HDMI 2.0) in legacy 2160p/SDR TVs, so it may be reasonable to expect that many of these units are BT.2020-compatible, and thus able to render HLG content.

emission encoder. As a commercially deployed system in 2016, this down-mapping process is considered to be a Foundation UHD technology.

8.5 Avoiding Image Retention on Professional and Consumer Displays

8.5.1 Background

As per EBU recommendation R 129 "Advice to Broadcasters on Avoiding Image Retention on TV Production Displays" [89], to minimize the risk of static image retention or premature ageing of displays, broadcasters and other content providers should take note of the following.

Broadcasters encounter 'static' images in a number of situations, including on-screen channel identifications, interactive application flags, banner displays, screens displayed when radio services are being received, program guides as well as longer-term text inserts such as sports scores.

This type of content is likely to remain an editorial feature of many broadcasts. It should also be noted that if image retention issues occur in production or broadcaster environments, it can also occur on viewers' television screens.

8.5.2 Definition of Static Images

For the purpose of this document an image is deemed to be static if any part of the screen is occupied by any part of the image for more than a total of six hours in any 12-hour period on more than one occasion in a seven-day period.

If an image is not static the risk of a retained image being formed from it is low. To assist in ensuring that images are not static, certain specific practices might be considered, including:

- Moving the position of images on the screen from time to time in order that the definition of 'static' is not met.
- Instigating a time-out of static images where appropriate.

8.5.3 Recommendations

The luminance or fully saturated chrominance (in the case of high color static images) value of any static image should be restricted to a value equal to the average picture level of the screen in order to minimize the risk of forming a retained image.

Two alternative methods of achieving this are:

- To use a technique known as 'Linear–key mixing' that overlays the static image as a partly transparent image over the picture content. The 'added image volume' level that sets the apparent transparency should not be set any higher than a level necessary to make the added image acceptably visible.
- To limit the signal level of the static image to no more than:
    - 40% of peak white for standard dynamic range (SDR) static images
    - 47% of reference white1 (i.e. 35% of peak white signal level) for Hybrid Log-Gamma (HLG) high dynamic range (HDR) static images
    - 37% of peak white for ST.2084 Perceptual Quantizer (PQ) high dynamic range (HDR) static images

Further, it is recommended that the use of saturated color static images be avoided wherever possible and particularly where one is laid over the other.

8.6 Additional UHD Technologies beyond Foundation UHD

Methods and challenges for adding non-Foundation UHD technologies into real-time program assembly vary from technology to technology. Further, an enhancement UHD technology might be continually incorporated into the program stream (i.e., part of the “house format”) or could be incorporated intermittently. Ultra HD Forum will continue to study these cases and intends provide further information in future versions of these Guidelines.

9. Distribution

This section describes the various stages of distribution for a Foundation UHD workflow. It shows the primary nodes and distribution paths in a typical workflow and describes each node and interface in the chain.

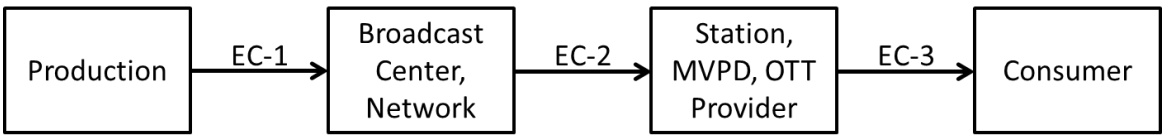


Figure 14 Distribution Nodes and Interfaces

Table 11 Compression and Distribution Nodes and Interfaces

Production	Content production location, e.g., studio, live venue truck, etc.
EC-1	Compression/distribution from Production to a central location
Broadcast Center, Network	A central location such as a broadcast network where content is aggregated for further distribution to station affiliates or service providers; this node is often used in distribution of Live content
EC-2	Compression/distribution to the final point before distribution to consumers
Station, MVPD, OTT Provider	A service provider that distributes content to consumers, e.g., a local television station affiliate, an MVPD or an OTT service provider; this node also often produces Live content, such as news programming
EC-3	Compression/distribution to the consumer device
Consumer	The viewer device, such as a television set, tablet, PC, STB connected to a display, etc.

Some workflows may be simpler, with only three nodes (production, service provider and consumer) while others may be more complex with more than 4 nodes in the chain. The concepts that encompass a four-node workflow can largely be applied to both simpler and more complex scenarios.

The workflows described include those available since 2016. The workflows described are able to carry metadata except where noted otherwise. The workflows apply to Real-time Program Services and to On Demand content that was originally offered as Live content.

Typical distribution practices involve decoding, modification and re-encoding the content as it flows from production to consumer. Carriage of transfer, color container, color matrix, and optional HDR10 static metadata is possible in production over both SDI and IP (see Section 6.1.11), in contribution feeds using AVC or HEVC, and in distribution using HEVC. When content is decoded at a node, modified or otherwise, and then re-encoded, attention must be given to preserving this data at each step. Audio and caption/subtitles are similar to those used in HD content distribution, and thus do not require the same attention. For pre-recorded

content, embedded test patterns at the head or tail of the program can be useful for verifying accurate signaling.

Section 14 describes the Next-Generation Audio (NGA) workflow. For the purpose of Foundation UHD, audio follows workflows established for Dolby-E and PCM in contribution applications, and AC-3, E-AC-3, HE-AAC, and AAC-LC as the 5.1 emission codecs.

Captions and subtitles follow workflows established for CTA 708/608, ETSI 300 743, ETSI 300 472, SCTE-27, and IMSC1 formats. HEVC [26] includes provisions for carrying captions and subtitles in the VUI and SEI in a similar manner to legacy video codecs.

In Foundation UHD the production system is likely SDI-based (1x12G or 4x3G) and therefore deployment of an SDI workflow is likely. In the future, an IP-based workflow should be implemented, using techniques such as Media over IP: ST 2022-6 [82] and near-lossless compression technologies such as VC-2 (Dirac), JPEG 2000, or other vendor proprietary solution.

Methods of carrying 2160p over SDI defined by SMPTE are shown in Table 12 below.

Table 12 SDI Input Standards for 2160p Content

Interface	Standard	Details	Notes
4x 3G-SDI*	SMPTE ST 424 [79]	4 quadrants of 3G-SDI	
	SMPTE ST 425-1 [80]	3G-SDI source input format mapping (Level A and Level B)	2 options: quad split or 2 sample interleave
1x 12G-SDI	SMPTE ST 2082 [84]	12Gbps SDI native	

* For 1080p, only 1x 3G-SDI is needed.

Metadata for HDR10 can be carried over SDI in VANC, per SMPTE ST 2108-1 [47]. If HDR10 metadata is present, it can be applied at the video encoder as follows:

- In compressed format, HEVC, Main 10 Profile, Level 5/5.1 may be used for metadata signaling. The metadata is carried via VUI and SEI messages (see Section 6.1.8).
- In a “light compression” format, such as a 12G-SDI signal mapped into a 10GbE, there are multiple options including VC-2 (Dirac), JPEG 2000, and other vendor proprietary solutions.

For Foundation UHD content, only uncompressed video over SDI (4x 3G-SDI or 1x 12G-SDI) or compressed video using HEVC, Main 10 Profile Level 5.1 is recommended for Foundation UHD workflows.

9.1 Production Processing and Contribution

This section describes processes for transmitting Foundation UHD content from an event venue or production studio to a central facility such as a Broadcast Center or Network. Note that in some cases the content may be distributed to more than one Broadcast Center, and different facilities may have different standards or requirements for content acquisition. For example, an international sports event program may be transmitted to a Broadcast Center, which in turn distributes the program to broadcast stations in different countries, which may have different format requirements, e.g., frame rate.

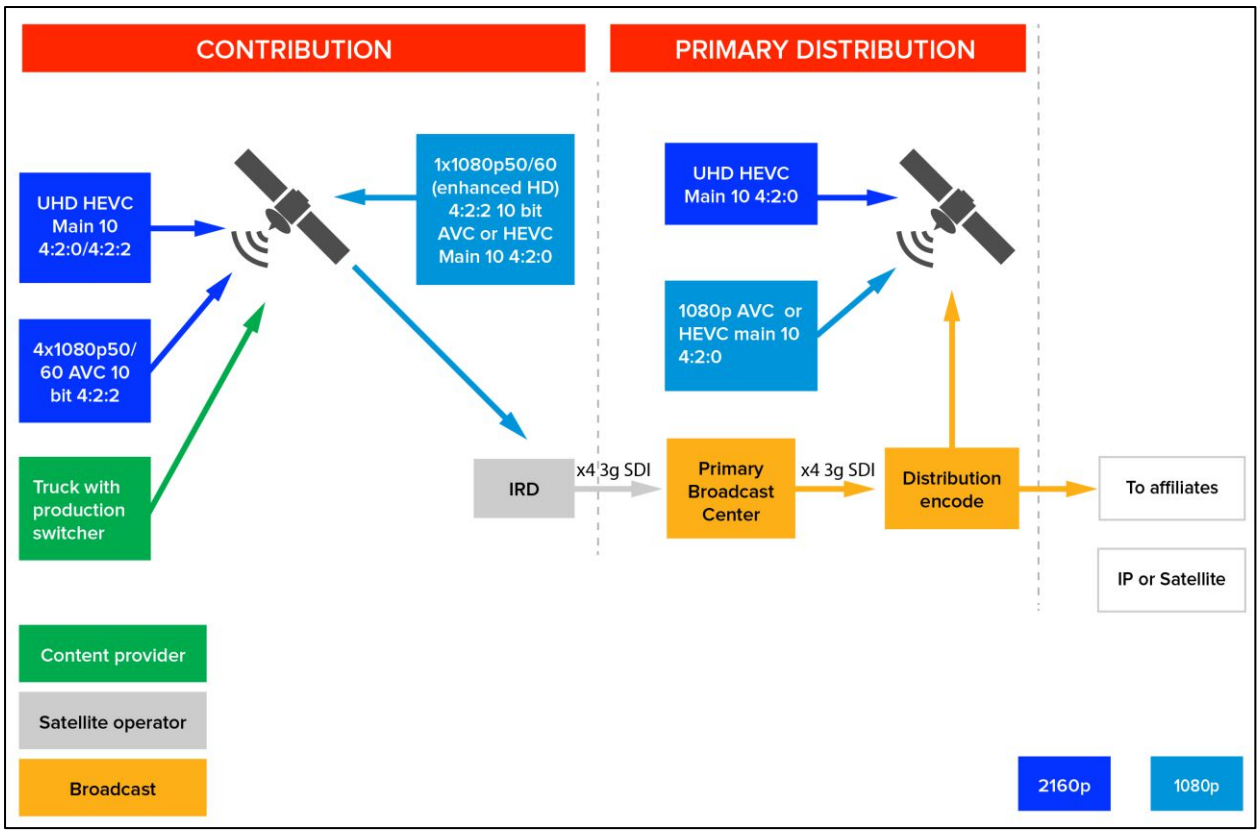


Figure 15 Contribution Workflow

In the case of Live content, the first stage is either a truck or remote production gallery, equipped with a production switcher, camera racking equipment, graphics and audio mixing. The production switcher typically will be receiving baseband signals from cameras and other sources such as digital disc recorders (DDRS) or other playback devices. The Technical Director (vision mixer) in the truck then applies transitions, lower thirds, etc., and may splice in Pre-recorded content (slow motion, interviews, match analysis, etc.). See Section 6.2.6 for details on Live content production and Section 8.2 for details on mixing source content of different formats in a Live program production.

In the case of Pre-recorded content, the studio has the opportunity to apply post-production edits. See Section 6.2 for details on Pre-recorded content production.

9.1.1 Video

It has been tested and confirmed by several Ultra HD Forum member companies that in the content production part of the chain, HLG10 and PQ10 can be used with 10-bit 4:2:2 workflow methods and equipment available since 2016.

Examples of image processing functions can include mixes, wipes, fades, chroma-keying, linear or luma-keying, motion tracking, DVE moves, slow motion, freeze frames, addition of graphics tickertapes or logos, use of virtual sets, etc. Since image processing may occur many times before final delivery to the consumer, this can be particularly important for Live programming, in which much of the workflow must be fully automated.

Once the content is ready, it is sent to a contribution encoder. Live production workflows typically feed a modulated uplink to a satellite that delivers the content to a Broadcast Center or Network. In some cases, fiber will be used as well as or instead of a satellite. For cost and bandwidth reasons, the satellite feeds will be compressed.

A 2160p feed may be compressed in HEVC, Main 10 Profile, 4:2:0 or 4:2:2 or quadrant-based in 4x1080p in AVC 10-bit 4:2:2. A 1080p HDR/WCG feed may be compressed in either HEVC, Main 10 Profile, 4:2:0 or AVC 10-bit 4:2:2. Note that when fiber links are used, intra-frame encoding and decoding, such as JPEG 2000, may be used. The HDR transfer function and system colorimetry must be predetermined and used uniformly for the production.

The satellite operator decodes the signal back to baseband using an integrated receiver decoder (IRD).

Quadrant based (4x1080p 4:2:0 or 4:2:2 10-bit) encoding/decoding is commonly used to create a 2160p image. AVC, AVC-I and JPEG 2000 all could be used for the quadrant codec. In the case that quadrant streams are sent, 4 encoders and 4 IRDs are synced. Single frame 2160p solutions using HEVC are likely to replace quadrants over time as they offer better compression efficiency (HEVC vs. AVC and single frame encoding) and are simpler to operate. This 2160p HEVC[19] contribution method is used in Foundation UHD.

AVC, HEVC and JPEG 2000 differ in the expected bitrate of the contribution file and in the mechanism for carrying HDR/WCG signaling.

Approximate examples of the contribution bandwidth required and HDR carriage signaling are shown below:

Table 13 Contribution Bitrates and Key Parameters

Source	Contribution Format	HDR/WCG Carriage Signaling	Approximate Typical Bitrate Range
1080p 50/60 fps	AVC 4:2:2 10-bit	Under definition in MPEG	20 – 50 Mbps
	HEVC, Main 10, Level 5.1 4:2:2/4:2:0	As per MPEG VUI/SEI signaling [26]	10 – 40 Mbps
	JPEG 2000	Not standardized	100 – 150 Mbps
2160p 50/60 fps	AVC 4:2:2 10-bit (4 quadrant)	Under definition in MPEG	90 – 140 Mbps total
	HEVC, Main 10, Level 5.1 4:2:2/4:2:0	As per MPEG VUI/SEI signaling [26]	50 – 80 Mbps
	JPEG 2000	Not standardized	450 – 550 Mbps

The Ultra HD Forum offers these bitrates based on the general experience of its members. It is important to note that the actual contribution bitrates can vary substantially from the figures shown depending on many factors, such as latency, quality of source content, type of source content, type of network, multi-hop contribution, etc.

In Foundation UHD, the main factor affecting contribution bitrates is the step between 1080p and 2160p spatial resolution; the addition of HDR or WCG has a much smaller impact. HDR and WCG do, however, require 10-bit encoding precision and modifications to the signal that, if not maintained, will ‘break’ the overall performance of the system resulting in an unacceptable image.

[19] For use in China, the AVS2 codec, Main10 profile, is used in lieu of HEVC. See Annex E: AVS2.

9.1.2 Audio

In Foundation UHD, production practices for audio are similar to those used in current HD content creation. In Foundation UHD audio follows multi-channel workflows established for 5.1 surround for delivery using one of the following emission codecs as appropriate for contribution applications: AC-3, E-AC-3+JOC (an instance of Dolby Atmos channel-based immersive audio), HE-AAC, or AAC-LC.

9.1.3 Closed Captions and Subtitles

Production practices for closed captions and subtitles are similar to those of HD content creation in Foundation UHD. Closed captions and subtitles follow workflows established for CTA 608/708, ETSI 300 743, ETSI 300 472, SCTE-27, or IMSC1.

9.2 Broadcast Center Processing and Primary Distribution

This section describes the processes and functions involved in Primary Distribution, i.e., transmitting content from a central facility such as a Broadcast Center or Network to a service provider such as a DTT, MVPD or OTT provider.

In the Broadcast Center, PQ10 or HLG10 signals can follow roughly similar workflow methods as those used for HD programming during image processing operations using a presentation switcher.

The output of the Primary Distribution encoder will go to a MVPD or OTT provider that will typically decode the content, modify it, and re-encode it to the final distribution format.

Figure 16 depicts different mechanisms of Primary Distribution delivery in TS formats to an MVPD.

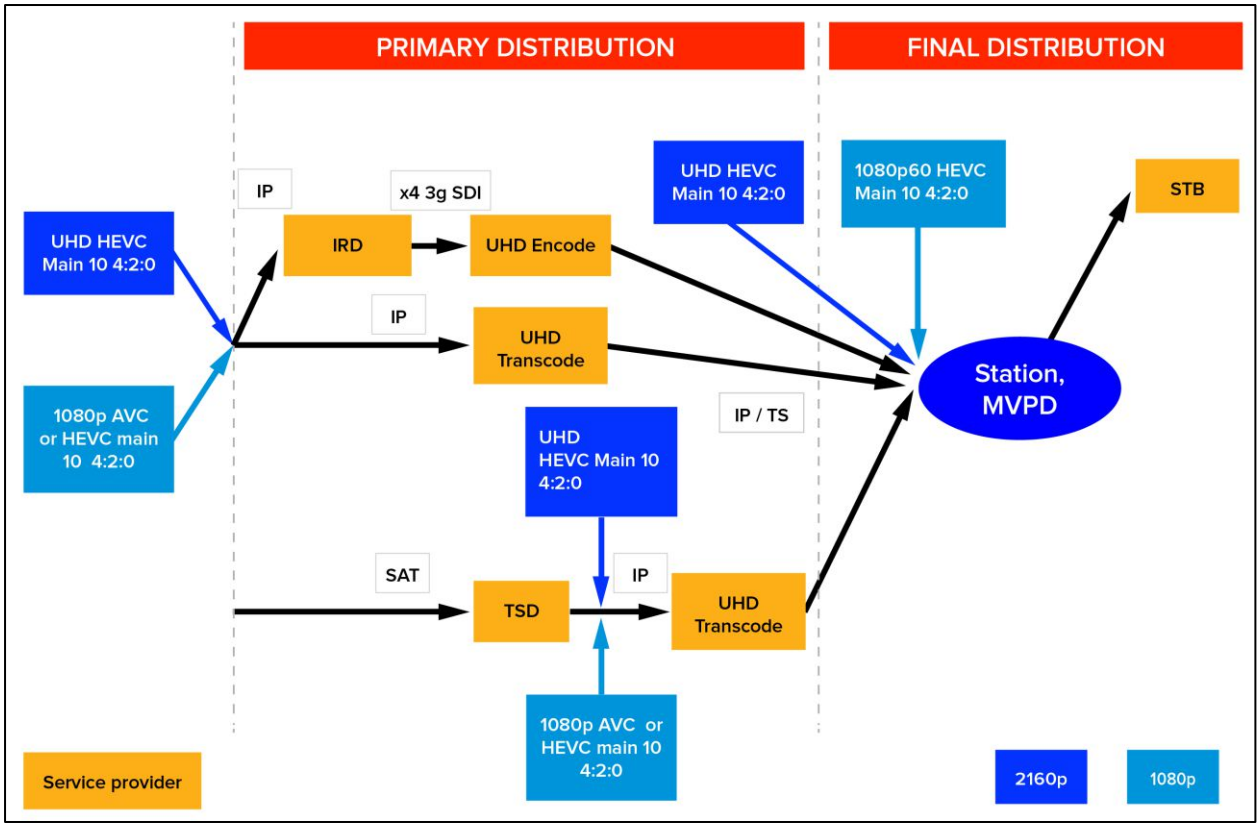


Figure 16 Primary Distribution to an MVPD or Station

Primary distribution to an OTT provider follows a similar scheme, except that the content is delivered using MPEG DASH instead of MPEG TS as shown in Figure 17 below.

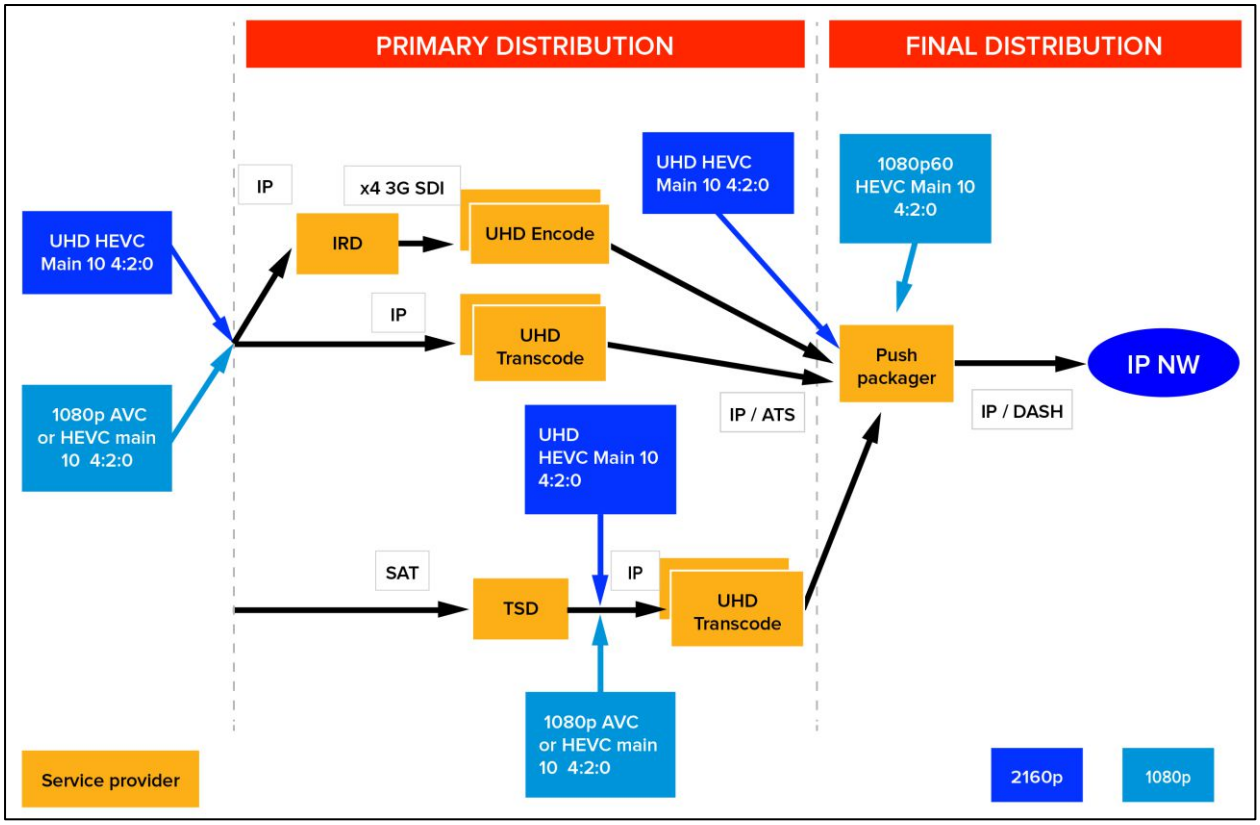


Figure 17 Primary Distribution to an OTT Provider

Primary distribution encoding to affiliates or other partner organizations is expected to be 4:2:2 or 4:2:0 10-bit. Note that frame rate (temporal) conversion may be required for international distribution.

For Live content, the production of the signal may be done via a live ingest server and playout server whose output will be SDI (1x12G or 4x3G or 1x3G in the case of 1080p content). The signal encoded by the Primary Distribution encoder should be HEVC, Main 10 Profile, 10-bit depth [26].

Table 14 describes the formats that can be used for Primary Distribution transmission.

Table 14 Primary Distribution Bitrates and Key Parameters

Spatial Resolution	Primary Distribution Format	HDR/WCG Carriage Signaling	Approximate Typical Bitrate Range
1080p	HEVC Main 10 Profile, 10-bit [26]	VUI/SEI signaling [26]	10-20 Mbps
2160p	HEVC Main 10 Profile, 10-bit [26]	VUI/SEI signaling [26]	40-50Mbps

Like contribution bitrates, the typical bitrates shown for HEVC 2160p are early approximations only based on the general experience of Ultra HD Forum members; Primary Distribution bitrates depend on many factors, such as latency, the quality of the source content, type of source content, type of network, etc.

9.3 Final Distribution from MVPD/OTT/DTT Provider Processing

This section describes image processing that may occur at the MVPD, OTT, or DTT provider and the encoding, multiplexing and packaging required to prepare content for final distribution to consumers.

The MVPD, OTT, or DTT service provider receives the Primary Distribution feed from the broadcast or network center. In many cases, the service provider will decode the signal and re-encode it for final distribution to the consumer.

9.3.1 Bit Depths

The below table illustrates bit depths currently in use:

Table 15 Existing Practices for Real-Time Program Service Distribution Formats

Case	Bit Depth	Color Gamut	Peak White at D65	Color Volume	HDR	Use Case
1	8	BT.709	100	BT.709	No	Deployed
2	10	BT.709	100	BT.709	No	Deployed
3	10	Up to BT.2020	100	BT.2020	No	DVB UHD-1, Phase 1 Scenario
4	10	Up to BT.2020	Up to 10,000	BT.2100	Yes	Ultra HD Forum Guidelines

The Ultra HD Forum finds that Case 2 has not been widely deployed and may be phased out quickly. Only Cases 3 and 4 are recommended for Foundation UHD services, and Case 4 is preferred. Cases and 1 and 2 are included for context.

In Cases 3 and 4, SMPTE ST 2086 can be used to signal Peak White. It should also be noted that in Cases 3 and 4, the color gamut can be up to BT.2020 color primaries; however, in current practice the gamut does not exceed DCI P3 primaries.

9.3.2 Video

In this final part of the chain, image manipulation may still be needed for ad insertion, ‘squeeze & tease’, channel logo insertion, etc. PQ10 or HLG10 can again follow roughly similar workflow methods as used for HD programming as they require only signaling describing the transfer function and system colorimetry.

HEVC[20] Main 10 Profile, 10-bit [26] is recommended as the only final distribution or emission codec as shown in Table 16 below, as all UHD decoders support HEVC.

[20] For use in China, the AVS2 codec, Main10 profile, is used in lieu of HEVC. See Annex E: AVS2.

Table 16 Final Distribution Bitrates and Key Parameters

Spatial Resolution	Final Distribution Format	HDR Carriage Signaling	Approximate Typical Bitrate Range
1080p	HEVC Main 10 Profile, 10-bit [26]	VUI/SEI signaling [26]	5-18 Mbps
2160p	HEVC Main 10 Profile, 10-bit [26]	VUI/SEI signaling [26]	10-40 Mbps

The Ultra HD Forum provides guidance for bitrates used throughout the final distribution chain only in the form of ranges due to the many parameters that influence the choice of the final distribution bitrate. History suggests that encoding rates improve significantly over time.

The bitrates used will depend on factors such as:

- Whether 2160p or 1080p is used
- Whether the source content is p60 or p50 (or lower) frame rate
- The quality criteria of the operator
- The performance of the encoder

The Ultra HD Forum members offer the below table of bitrates that were in use in services as early as 2016. These "real world" bitrates are intended to offer an additional benchmark for potential bitrates for Foundation UHD services. This information was provided voluntarily, and not all volunteers were able to provide all the metrics.

Table 17 Example "Real World" Bitrates as early as 2016

Delivered via	**Transfer function**	**Frame rate**	**Bit depth**	**System Colorimetry**	**Audio codec(s)**	**Bitrate**	**Notes**
Satellite	SDR	30fps				27Mbs	
Satellite	HLG or SDR	59.94fps	10-bit	BT.2020	Audio adds 5Mbps	35Mbps	Bitrate works for sports and HDR content
Satellite or IPTV	SDR	50fps	8 or 10-bit		AC-3	24-30Mbps	
	PQ	59.94fps	10-bit		AAC, AC-3	32Mbps	
IPTV	PQ	50fps	10-bit	BT.2020	MPEG2, AC-3, DD	25Mbps	Sports content
Satellite	SDR	50fps	10-bit	BT.709		30-38Mbps	Drama, movie content

Note that all the examples in Table 17 are 2160p spatial resolution and use HEVC encoding with 4:2:0 color subsampling.

9.3.3 Adaptive Bitrate (ABR) Streaming

ABR streaming works to minimize playback interruptions when content is delivered OTT through variable quality Internet connections. Multiple encodings enable switching to lower bit rates, which allow a bias towards continuous playback versus higher quality. The encodings are separated into tracks for audio, video, and subtitles. The separate tracks allow real-time multiplexing and addition of tracks as needed without changing the original tracks. Guidelines for Foundation UHD ABR for Live content and Real-time Program Service assembly are as follows:

- Use DASH-IF framework (Guidelines for Implementation: DASH-IF Interoperability Points [16]), which includes manifest format of available assets (Media Presentation Description [MPD]), container (ISO BMFF), video segmenting, security, and HEVC profile settings. DASH-IF may be extending their work to include additional Ultra HD signaling in their Guidelines for Implementation.
- It is recommended that the adaptation set include 720p, 1080p, and 2160p resolutions. However, it should be noted that DASH-IF does not have any specific recommendations for spatial resolutions of Representation in one Adaptation Set, but only specifies maximum resolution for each DASH-IF Interoperability Point (since in practice the input video may be in any arbitrary resolution).
    - Adaptation Sets should be generated such that seamless switching across all Representations in an Adaptation Set is possible. For this purpose, a Representation may be encoded in a lower resolution to provide suitable quality at lower bit rates, permitting down-switching and continuous playout. In this case it is expected that the decoder and renderer apply upscaling to the display resolution in order to ensure seamless switching.
    - Each Adaptation Set is expected to have sufficient annotation to be selected at start-up based on decoding and rendering capabilities. Typically, if one MPD targets different receiving classes, then multiple Adaptation Sets in one MPD for the media type are present. Selection is a key functionality of DASH to support backward and device-compatibility; i.e. not every player has to support all formats described in an MPD.
- Keep HDR and WCG parameters the same across the Adaptation Set resolutions[21]. Aspect ratio must also be constant across the adaptation set, and it may also be advisable to maintain the same framerate across the adaptation set in some circumstances.
- Since segment sizes for the same segment duration for 2160p will be larger than for lower resolutions, ensure the CDN is configured to optimize the larger segment sizes, or reduce segment duration to accommodate the CDN capability for segments at 2160p resolution.

Table 18 offers example bitrates that could be used for OTT services. Note that the combined rows in the table do not represent a suggested “adaptation set” of streams that are intended for seamless switching in a given program. Each row represents an independent example of possible expected bitrates.

[21] See “Guidelines for Implementation: DASH-IF Interoperability Points v4.3”, Section 6.2.5, which may be found at https://dashif.org/docs/DASH-IF-IOP-v4.3.pdf for reference.

Table 18 Example Bitrates for Video Streams

Resolution	frame rate	Approximate HEVC bitrate
3840x2160	p60/50	15-20Mbps
1920x1080	p60/50	5-10Mbps
1280x720	p60/50	2-5Mbps
1280x720	p30/25	1-2Mbps
720x404	p30/25	<1Mbps

9.3.4 Audio

In Foundation UHD, Audio may be delivered using legacy or currently employed multi-channel audio codecs, e.g. AC-3, E-AC-3, HE-AAC, and AAC-LC. E-AC-3 and HE-AAC, are considered to offer similar quality at reduced bitrates. While 2-channel stereo can be delivered, it is recommended to deliver 5.1 channel soundtracks when available for an improved sonic experience. Dolby Atmos soundtracks are available for some programs and can be delivered using the E-AC-3+JOC [35] codec.

Many broadcasters and operators are governed by regional legislation regarding managing loudness of broadcast audio. In the United States, for example, a DTT or MVPD provider is obligated to manage loudness per the CALM act, and thus should ensure the audio transmission is compliant with ATSC A/85. OTT providers in the U.S. may also wish to be compliant with CALM in the event the FCC decides to consider them as MVPD providers. Other territories should consider any local specifications or regulations regarding loudness, e.g. EBU R-128. Note that while E-AC-3+JOC [35] delivery of Atmos soundtracks using channel-based production and delivery is described as a Foundation UHD technology, E-AC-3+JOC is also capable of delivering the dynamic spatial objects that are described as an NGA feature in Section 14.3.

9.3.5 Closed Captions and Subtitles

Production practices for closed captions and subtitles are similar to those of HD content creation in Foundation UHD. Closed captions and subtitles follow workflows established for CTA 608/708, ETSI 300 743, ETSI 300 472, SCTE-27, or IMSC1. HEVC carries captions and subtitles in User data registered by Rec. ITU-T T.35 SEI defined in HEVC specification section D.2.36 (syntax) and D.3.36 (semantics) [26].

9.3.6 Considerations for UHD Technologies beyond Foundation UHD

CAE is a technology that works with OTT devices equipped with HLS or DASH- compliant video players. The encoders implementing CAE should also follow guidelines defined for content preparation for HLS [67] or DASH [16] formats respectively.

9.4 Transport

Operators deploying 2160p HDR/WCG content over MPEG-2 TS can use the DVB UHD-1 Phase 2 specification. Operators can carry Foundation UHD content over RTP/UDP/IP per DVB IPTV [41] (single bitrate only, not ABR), i.e., MPEG-2 TS over RTP/UDP/IP.

In addition to DVB UHD-1 Phase 2 specification, DTT operators can refer to ATSC A/331 [53] which specifies MPEG Media Transport (MMT) and ROUTE/DASH for carriage of UHD content.

For OTT services, MPEG DASH is used to transport Foundation UHD content as follows:

- DASH per DVB DASH specification [13] for live applications
- DASH 265 for live from DASH-IF Guidelines [59]

10. Decoding and Rendering

This section covers guidelines for implementation of decoding capabilities in the consumer player device, picture processing capability of the consumer display device as well as the interface between the consumer player device and the consumer display device. There are two possible architectures for decoding and rendering in the consumer home: 1) STB decoder connected to a display, and 2) integrated decoder/display.

The extent to which the consumer decoder or display is able to switch between SDR/BT.709SDR/BT.709 and PQ10/HLG10 content or switch between PQ10 and HLG10 seamlessly is not proven, nor is it specified by any standards. It is recommended that service providers employ conversions as needed to ensure that program content, interstitial material, and graphic overlays (bugs, crawls, etc.) within in a given program are either entirely SDR/BT.709 or entirely PQ10 or entirely HLG10, to the extent possible. Section 7.2 offers details on conversions at Production and Distribution and Section 10.4 has details on conversions in consumer equipment for backward compatibility.

This section addresses equipment that is compatible with Foundation UHD content streams. Methods for addressing Backward Compatibility for legacy decoders and displays are discussed in Section 10.4. Note that decoders that support only 8 bits are not considered Foundation UHD decoders. (These were the first generation of “UHD” decoders.)

10.1 Decoding

Foundation UHD consumer decoder device capabilities:

- Able to decode HEVC, Main 10 Profile, Level 5.1
- Able to process BT.2020 [3] system colorimetry
- Able to process PQ transfer characteristics
- Able to process HLG transfer characteristics
- Able to process HDR10 content (with or without metadata)
- For the STB-display architecture, the STB also supports the following:
    - Output Interface – at least HDMI 2.0a/b*
    - Optionally able to transmit ST 2086 [10] metadata, MaxCLL, and MaxFALL to the connected display device
- Able to decode multi-channel Dolby AC-3, E-AC-3, DTS-HD, HE-AAC and AAC-LC audio
- Able to decode closed captions and subtitles per CTA- 608708, ETSI 300 743, ETSI 300 472, SCTE-27, or IMSC1
- Able to ignore enhancement technologies that are layered upon foundation technologies

*Note that HLG transfer function requires at least HDMI 2.0b interface.

10.2 Rendering

The characteristics of Foundation UHD consumer display devices differ significantly from those of professional displays used to grade and master the content. These characteristics include luminance range, color gamut, screen size (smartphone, tablet, TV), and more. In order

to compensate for these differences, Foundation UHD consumer display devices are capable of processing incoming Foundation UHD content so that the rendered video reproduces the creative intent as optimally as possible, for example by appropriate color volume transformation of an HDR/WCG video signal to the display panel.

Foundation UHD consumer rendering device capabilities:

- Able to process PQ transfer characteristics
- Able to process HLG transfer characteristics
- Able to process HDR10 (with or without metadata)
- Able to process BT.2020 [3] system colorimetry
- Able to render 60p frame rates
- Able to render content having 2160p spatial resolution
- Able to process multi-channel 5.1 channel surround sound
- Optionally able to render Atmos immersive soundtracks delivered by E-AC-3+JOC [35]
- For STB-display architecture:
    - Input Interface – at least HDMI 2.0a/b*
        - Transmission of Extended Display Identification Data (EDID) information including peak and minimum luminance
        - Transmission of supported EOTFs
        - (Optional) Transmission of RGB primaries

*Note that HLG transfer function requires at least HDMI 2.0b or later interface.

10.3 Overlays Inserted at the Consumer Device

Closed captions, subtitles and graphic overlays may be rendered by a STB, a display connected to a STB, or an integrated decoder/display. In the case of the STB-display architecture, it is possible that both the STB and the display are rendering overlays at different times or simultaneously (e.g., STB rendering an EPG and the display rendering a volume control indicator).

The current specifications regarding closed captioning and subtitles are based on BT.709 [2] system colorimetry and SDR. When overlaying closed captions and/or subtitles onto BT.2020 [3] system colorimetry and HDR video, Foundation UHD decoders should remap RGB values of closed captions and/or subtitles as needed to ensure that color shifts do not occur while mixing two elements having different system colorimetries and/or dynamic ranges.

Similar care should be taken when displaying graphics, EPGs, user interface overlays and so on.

10.4 Considerations for UHD Technologies beyond Foundation UHD

Service delivery has to account for the format support of the decoder/display device through a variety of strategies. Foundation UHD formats are generally supported on all Foundation UHD decoder/displays but deploying additional UHD Technologies may require one or more

strategies to ensure continued operation of Foundation UHD decoder/displays. These strategies can be formalized as follows:

- Simulcast – Sending enhanced UHD technology-based service separately from a Foundation technology-based service where the content of the two streams is essentially identical, which can be used when the enhanced technology-based service offers no backwards compatibility with Foundation UHD decoder/displays
    - Example: A service that includes an NGA Audio program, and also makes available an alternate audio program using a Foundation, channel-based format, with the correct audio program selected by the client device.
- Backward compatibility – The enhanced UHD technology-based service is inherently supported by Foundation decoder/displays, normally with a graceful degradation in experience or functionality on the display
    - Example: A Linear Broadcast in HLG10 that is decoded by all displays and rendered by HDR/BT.2020 decoder/displays as HDR and by SDR/BT.2020 decoder/displays as SDR.
- Optional Capabilities – The enhanced UHD technology is delivered along with a Foundation UHD based service as an optional component that is not identifiable or usable by a Foundation UHD technology-based device
    - Example: An SL-HDR2 encoded service, which comprises a PQ10-encoded video in conjunction with ST.2094-20/ST-2094-30 metadata, which an SL-HDR2 capable decoder/display will render fully, and a PQ10 capable decoder/display will render as PQ10 without the enhancement of the dynamic metadata.
- Layering – The enhanced UHD technology-based service is delivered alongside a base layer that is UHD Foundation compliant and the decoder/display processes the required combination of layers in a way suited to the its capabilities
    - Example: Delivery of a backward compatible HFR video that is encoded using a base layer standard frame rate elementary stream and a secondary enhancement layer elementary stream that contains the additional frames to allow an HFR capable decoder/display to re-multiplex the two streams together to render the HFR video. A non-HFR capable decoder/display would identify and render only the base layer standard frame rate video, ignoring the additional frames.
- Service Provider Down-conversion – The enhanced UHD technology-based service is converted to a Foundation UHD service, generally with the loss of enhancements
    - Example: An operator receiving a SL-HDR2 encoded linear service for distribution, undertaking a down conversion to a HLG10 encoded service for distribution.
- Device Down-conversion / Up-conversion – The enhanced UHD technology-based service is down converted (or, in the case of Foundation based services on enhanced UHD based devices, up converted) to a format suited for final display, normally with loss of enhancement in down-conversion, or with a simulacrum of an improvement in up-conversion.
    - Example: An operator distributing an SL-HDR1 encoded service to an STB that can detect the capability of the display via an HDMI interface. The STB applies the SL-HDR1 metadata to supply HDR/BT.2020 to HDR/BT.2020 displays and does not apply the SL-HDR1 metadata to supply SDR/BT.709 displays.

Each additional UHD technology can make use of one or more of these strategies, which is described in the relevant section of the additional UHD technology.

11. Format Interoperability

There are a number of requirements for format interoperability, when considering the needs of broadcasters or service providers working with both Foundation UHD and HD (and even SD) content. One example of this is Backward Compatibility, i.e., the means of delivering a Foundation UHD service to a legacy consumer device, in such a way that it can be viewed at some level of acceptable quality on an SDR/BT.709 display.

Backward compatibility that conveys the full creative and artistic intent of the original Foundation UHD content is not attainable. Foundation UHD gives producers, camera operators, directors, editors, production designers, etc. more creative possibilities. Since legacy screens will not be able to display the full resolution, dynamic range and color gamut of the original production, some of the original creative intent will be lost.

Foundation UHD services are distributed via DTT, OTT or MVPD.

This section addresses Foundation UHD backward compatibility for the installed base of SDR 2160p TVs, both BT.709 [2] andBT.2020 [3] displays. Thus, not all facets of Foundation UHD content are considered for Backward Compatibility in Foundation UHD. Specifically:

- Spatial resolution down-conversion is not in scope; only 2160p-capable decoder/displays for Foundation UHD content are included
- Frame rate down-conversion is not in scope; only 50/60 Hz-capable decoder/displays for Foundation UHD content are included
- HDR and WCG are the primary parameters being considered for Foundation UHD backward compatibility

Backward compatibility for OTT and MVPD services involves either:

- For HLG10:
  - HLG10 technology is designed to produce acceptable results using the same content stream on both HDR and SDR devices, provided that the SDR device can process BT.2020 [3] system colorimetry (nb. not valid for BT.709 only devices). Validation of the quality of the SDR output has been investigated by the EBU, IRT, RAI and Orange Labs [88]. In the event that the SDR rendering of HLG10 content does not meet an operator's requirements, schemes similar to those proposed for HDR10/PQ10 may be used (see below).
- For HDR10 or PQ10
  - Simulcasting multiple broadcast streams, one in HDR10 or PQ10 and the other in SDR/BT.709 (see Section 11.2), and/or
  - Using a STB that can decode the Foundation UHD stream and deliver material suitable for an HDR/WCG, HDR/BT.709, or SDR/BT.709 display. In the case of HDR10, the STB may be able to take advantage of the HDR10 static metadata 6.1.5, when present, in creating the down-conversion. Ideally, the Foundation UHD STB is capable of serving any of these displays so that when a consumer decides to take advantage of HDR services, only a new display is needed.

Creating a backward compatible version of the content that is acceptably rendered on a 2160p SDR/BT.709 display may take place at various points in the supply chain between production and the display (see also Section 7.2):

- Content producers can generate both HDR/WCG and SDR/BT.709 versions, applying creative intent adjustments to both versions. This results in the highest

quality conversion but requires time and resources and both versions must be carried throughout the supply chain. This option may not be practical for Live content workflows.

- Professional equipment can down-convert HDR/WCG to SDR/BT.709 with or without the benefit of creative adjustments. This equipment may be sophisticated and thus may be the best option if automated conversion is necessary.
    - HDR10 static metadata 6.1.5, when present, may assist this process.
- Consumer equipment (i.e., STB) can down-convert HDR/WCG to SDR/BT.709 without the benefit of creative adjustments. This equipment is likely to be less sophisticated than professional equipment but may be a viable alternative when it is impractical to offer multiple versions of the content to the consumer premises.
    - HDR10 static metadata 6.1.5, when present, may assist this process.

11.1 Legacy Display Devices

In Foundation UHD, the Ultra HD Forum is considering legacy display devices that are connected to the MVPD STB or are receiving a suitable unicast stream from the OTT provider. In the latter case, the OTT provider offers a suitable stream, and it is up to the provider to determine which devices it can support. A STB that can ingest a Foundation UHD 2160p stream and output a stream that a legacy display device can render is considered. The variety of legacy display devices that a STB can accommodate varies by product as does the quality of the down-conversion.

A backwards compatible distribution solution or STBs capable of down-conversion can address first-generation 2160p SDR televisions, i.e., devices that can render 2160p resolution content with BT.709 [2] or BT.2020 [3] color gamut but only in SDR. In the absence of one of these solutions, a direct IP stream can be used to address HDR TVs, e.g., using an embedded HTML5 [32] or RVU client[22] that extracts the received broadcast stream and re-encapsulates it into an IP stream that can be transmitted to a TV via a Local Area IP Network. Note that currently some UHD displays are capable of accepting BT.2020 [3] content, however at this time no direct view display is available that is capable of rendering the full gamut of colors in the BT.2020 [3] system colorimetry. It is assumed that in these cases, the device employs “best effort” color volume transform tailored to its particular display characteristics, and thus these devices are considered BT.2020 [3]-compatible for the purpose of this discussion.

11.2 Down-conversion at the Service Provider

This option may be employed by OTT providers or by MVPDs. With this method, providers offer both Foundation UHD and legacy versions of the service and send the appropriate stream to devices (unicast) or simulcast both streams. In general, providers that use DASH as a transport method may use unicast and providers that use MPEG-2 TS may use simulcast. The variety of legacy devices served is a function of how many different streams a given service provider chooses to make available. This method may require content producers to deliver multiple versions of the content to the service provider and/or utilize professional conversion equipment at the headend (see Section 7.2).

The below diagram illustrates this method of backward compatibility.

[22] See http://rvualliance.org/what-rvu.

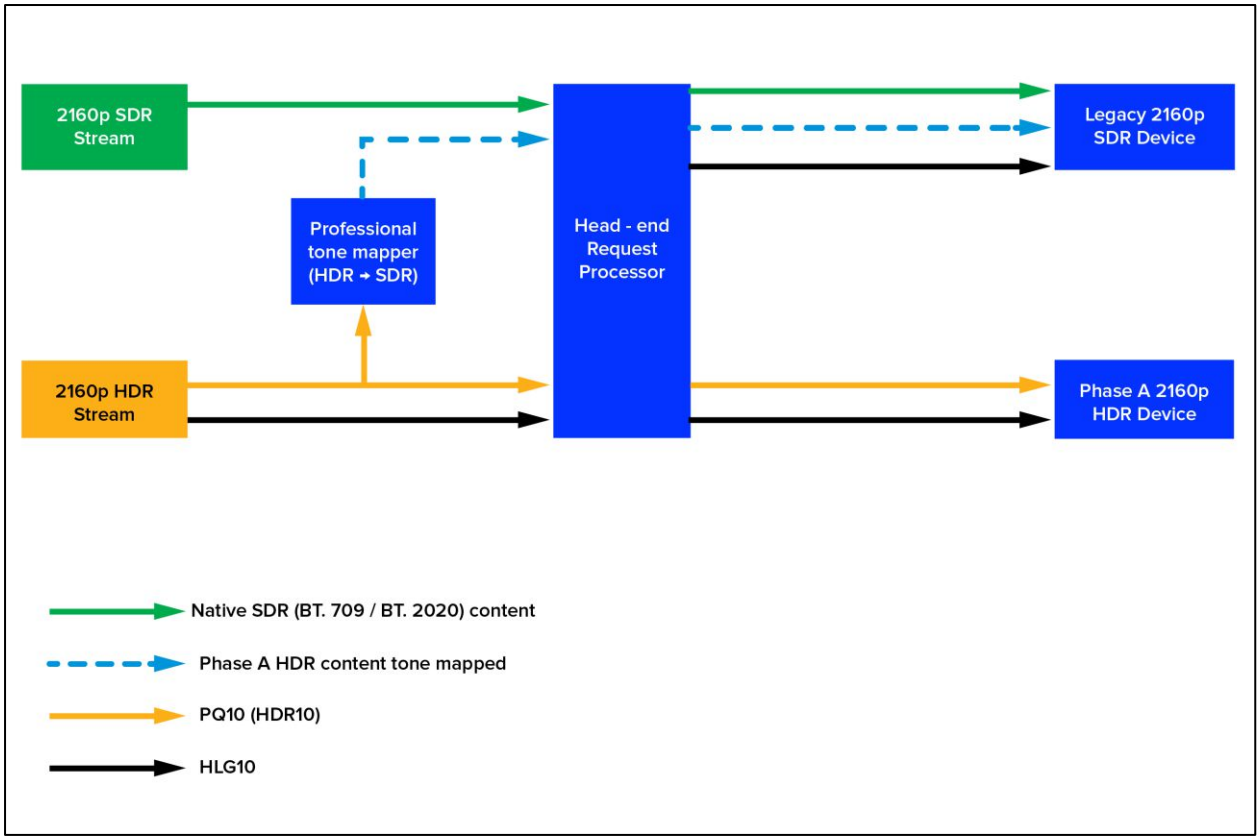


Figure 18 Down-conversion at the Headend

In the above diagram:

1. Operator receives legacy and Foundation UHD content from different content sources.
2. Operator can convert* Foundation UHD streams for legacy 2160p SDR devices.
3. Device requests content from headend based on its capabilities.
4. Headend request processor provides appropriate stream.

*Note that conversion could occur upstream of the headend; i.e., the content producer could provide the operator with both SDR and HDR versions of the content.

11.3 Down-conversion at the STB

This option may be employed in Foundation UHD by MVPDs that prefer not to use the bandwidth required for offering multiple unicast streams, such as via switched digital video technologies, or multiple simulcast streams. In this case, the STB is capable of decoding a Foundation UHD stream and is also capable of down-converting the content. As stated above, there are compromises with down-conversion of Foundation UHD content and service providers should test the quality of the output for acceptability.

Although there is no standardized method of down-converting BT.2020 [3] to BT.709 [2], it is expected that some STBs may have this capability. STBs may also have the capability of down-converting PQ10 or HDR10 to SDR. The diagram below illustrates this method.

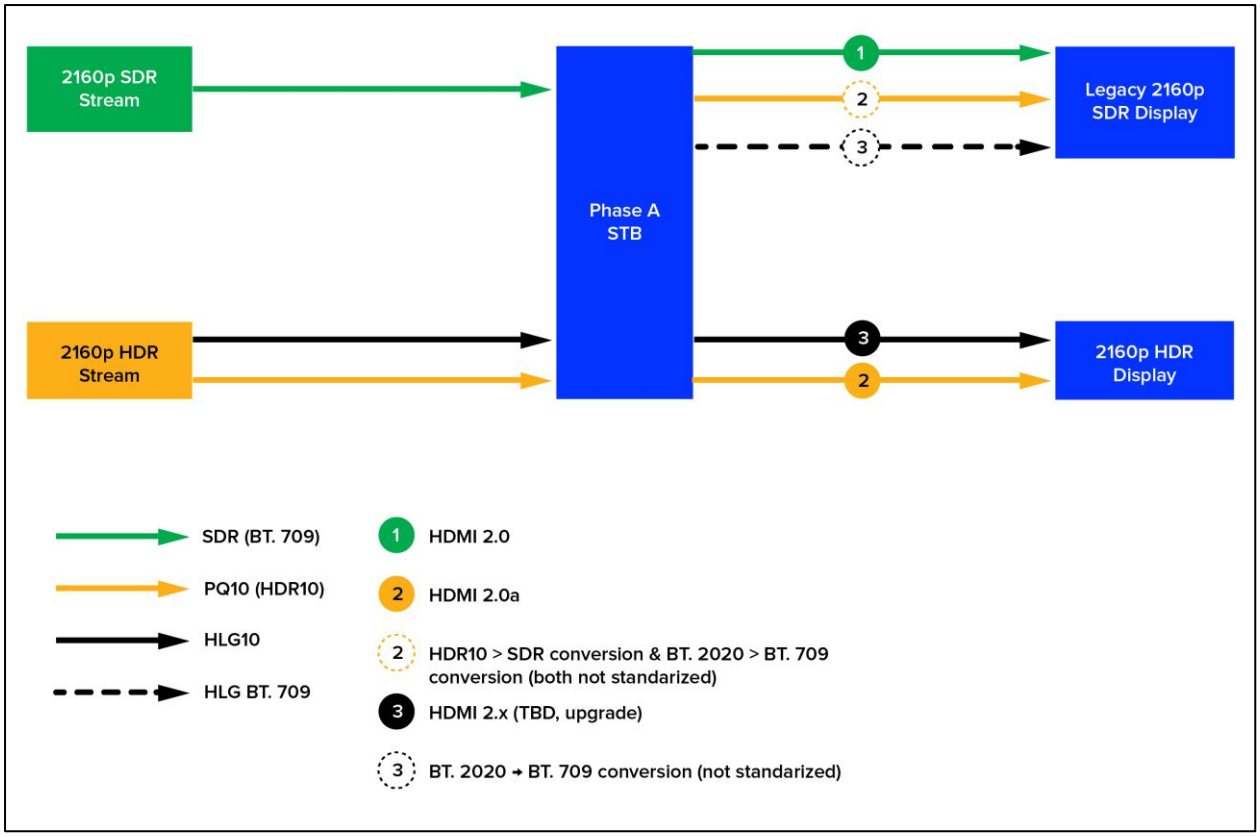


Figure 19 Down-conversion at the STB

Details:

- The Foundation UHD STB supports HEVC, Main 10 Profile, Level 5.1, BT.2020 [3], and HDMI 2.0a and optionally IP output.
- In this example, the legacy 2160p SDR display supports BT.709 [2] but does not supportBT.2020 [3].
    - Therefore, in the diagram, the Foundation UHD STB would convert the video from BT.2020 [3] to BT.709 [2] before transmitting it to the legacy 2160p SDR display.
    - Note that some legacy 2160p SDR displays may support BT.2020 [3] and for these displays, a Foundation UHD STB does not need to convert from BT.2020 [3] to BT.709 before transmitting to the TV.

11.4 Spatial Resolution Up-conversion of Legacy Services

This option may be employed in Foundation UHD by MVPDs that prefer not to use the bandwidth required for offering multiple unicast streams, and when Foundation UHD STBs are not be able to convert a Foundation UHD stream to an appropriate format and/or with sufficient quality for display on a legacy 2160p SDR display. Foundation UHD STBs (as well as legacy 2160p SDR displays) are expected to have the capability of upscaling 720p/1080i/1080p SDR channels to 2160p resolutions. This option requires simulcasting; however, the 720p/1080i/1080p SDR stream/service often already exists, e.g., during a transition period. In this case, the legacy 2160p SDR display gets the legacy stream and up-converts the spatial resolution to 2160p. Only the 2160p HDR display gets the Foundation

UHD stream. There are compromises with up-conversion of 720p/1080i/1080p content and service providers should test the quality of the output for acceptability.

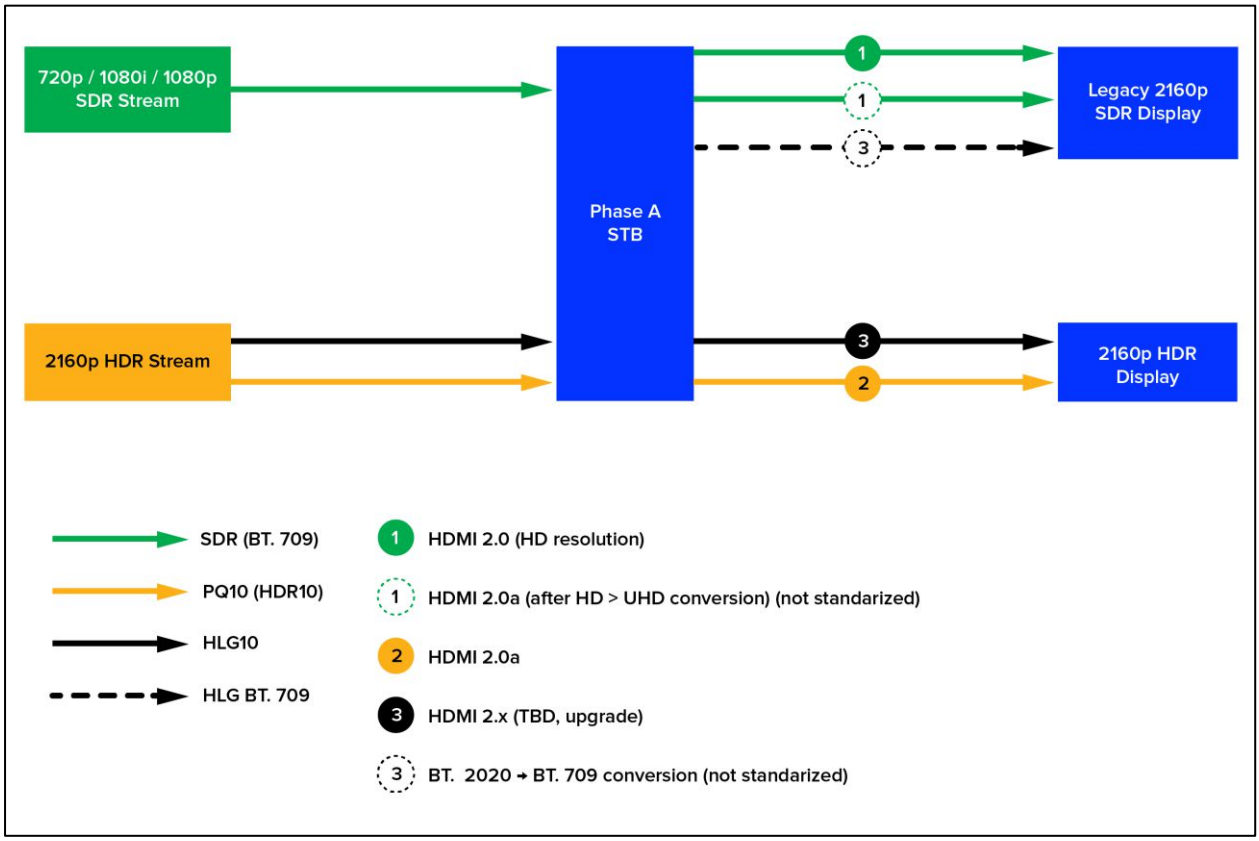


Figure 20 Spatial Resolution Up-conversion of Legacy Services

Details:

- The Foundation UHD STB decodes the 2160p HDR stream when connected to 2160p HDR displays.
- The Foundation UHD STB decodes the 720p/1080i/1080p SDR stream when connected to legacy 2160p SDR displays. The STB can either transmit the decoded 720p/1080i/1080p SDR video or convert the 720p/1080i/1080p SDR video to 2160p SDR video before transmitting it to the legacy 2160p SDR display.
- Note that SDR to HDR conversion, if needed, is best performed in the display device rather than in the decoder device.

11.5 Interoperability of Atmos Immersive Audio

For emission of Atmos Channel-based or Object-based Immersive Audio, an E-AC-3+JOC encoder with ETSI TS 103 420 [35] functionality is required. Internally, the encoder will create a backward compatible 5.1 channel version rendered from the 7.1.4 (or 7.1.2 or 5.1.2 or 5.1.4) input. This 5.1 channel render is E-AC-3 coded as normal and information about the render, as described in [35], is also carried. Legacy E-AC-3 decoders will reproduce the backward compatible base surround program while advanced E-AC-3 decoders, compliant with [35] will reproduce the full 7.1.4 (or 7.1.2 or 5.1.2 or 5.1.4) immersive audio program.

11.6 Considerations for UHD Technologies beyond Foundation UHD

Foundation UHD service formats can be rendered by Foundation UHD decoder/displays, as well as by decoder/displays that offer additional UHD Technology capabilities.

When deploying services that make use of additional UHD Technologies (beyond Foundation UHD) care needs to be taken to ensure that all devices are supported using one or more of the strategy approaches described in Section 10.4. The selection of the strategy being selected depends upon the nature of the additional UHD technology and the characteristics of the service that the provider wishes to deploy.

12. High Dynamic Range

12.1 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 [5]
- IC_TC_P: Recommendation ITU-R BT.2100 [5]
- Mastering display metadata: SMPTE ST 2086 [10] and CTA 861.4
- Dynamic metadata: SMPTE 2094-10 [86] and CTA 861.4
- MaxFall/MaxCLL: CTA 861.G

12.1.1 Dolby Vision Encoding/Decoding Overview

Figure 21 illustrates a functional block diagram of the encoding system. HDR content in PQ is presented to the encoder. The video can undergo content analysis to create the display management data at the encoder (typically for Live encoding) or the data can be received from an upstream source (typically for prerecorded content in a file-based workflow).

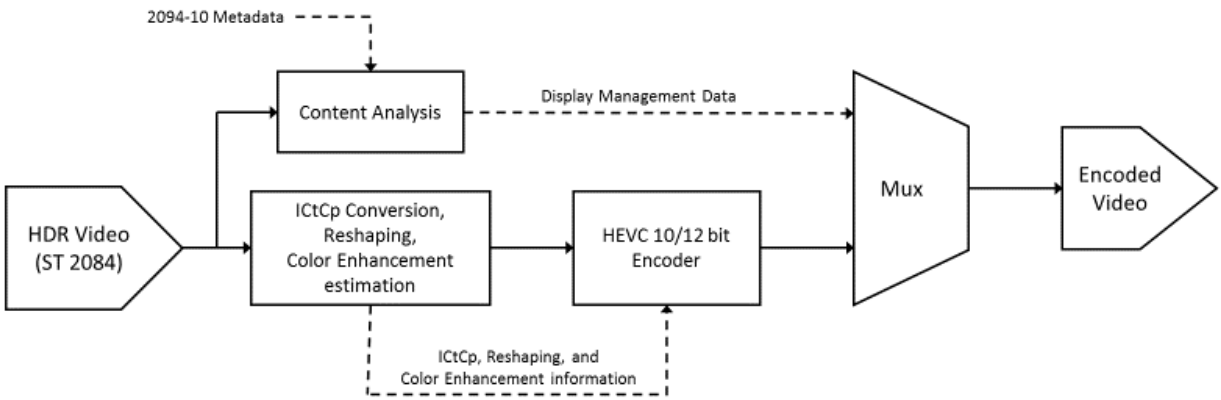


Figure 21 Encoder functional block diagram

If not natively in $ICTCP$ signal format, it may be advantageous to convert the HDR video into $ICTCP$ 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 22 illustrates the functional block diagram of the decoder. It is important to note that the system in no way alters the HEVC decoder: An off-the-shelf, unmodified HEVC decoder is used, thereby preserving the investment made by hardware vendors and owners.

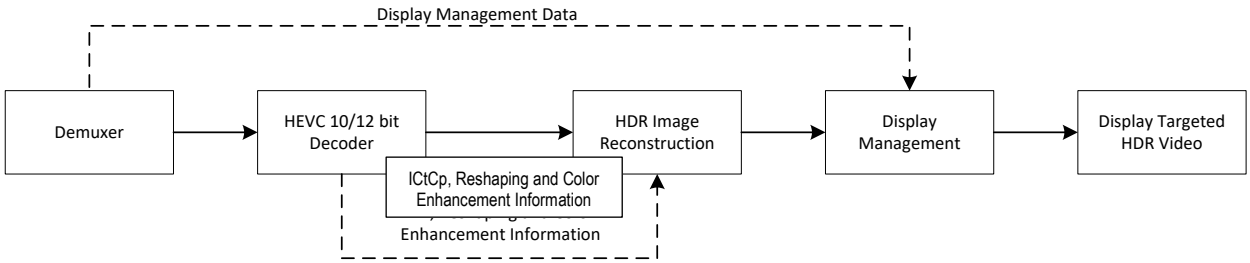


Figure 22 Decoder function block diagram

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.

12.1.2 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 12.1.4 and in ATSC A/341 [54], and when used in this method the streams are fully backwards compatible with HDR10 (assuming the underlying signal format remains YCbCr). A player receiving the stream can simply ignore the SMPTE 2094-10 dynamic metadata contained in the SEI messages and play the fully conforming HDR10 stream. In the case where the underlying signal format is ICtCp, the streams are generally not cross compatible, and the delivery system would need to deliver an alternate stream for non-Dolby Vision devices.

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

12.1.3 Dolby Vision Color Volume Mapping (Display Management)

Dolby Vision is designed to be scalable to support display of any arbitrary color volume within the BT.2100 standard [5], onto a display device of any color volume capability. The key is analysis of content on a scene-by-scene basis and the generation of metadata, which defines parameters of the source content; this metadata is then used to guide downstream color volume mapping based on the color volume of the target device. SMPTE ST 2094-10 [86] is the standardized mechanism to carry this metadata.

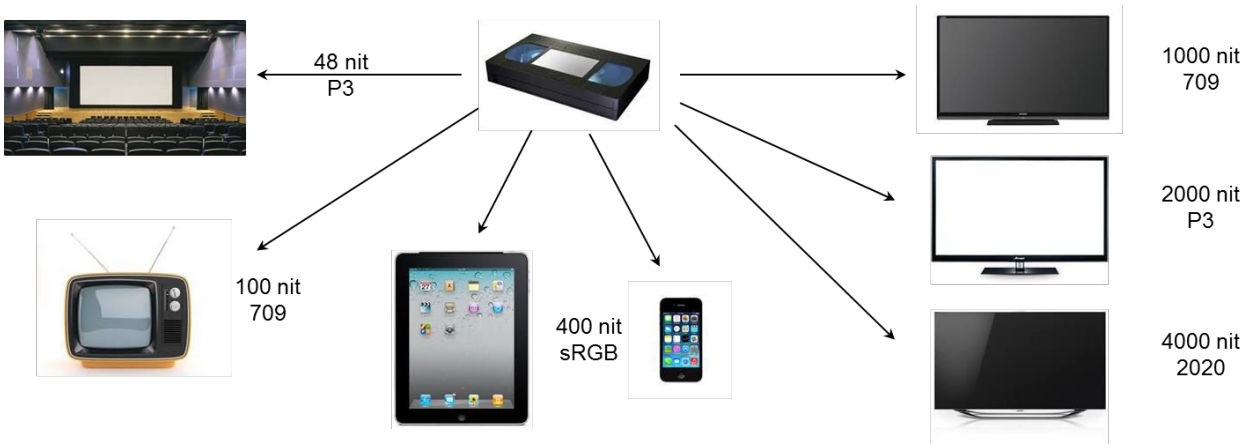


Figure 23 Example display device color volumes

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

12.1.4 Dolby Vision in Broadcast

In a production facility, the general look and feel of the programming is established in the master control suite. Figure 24 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.

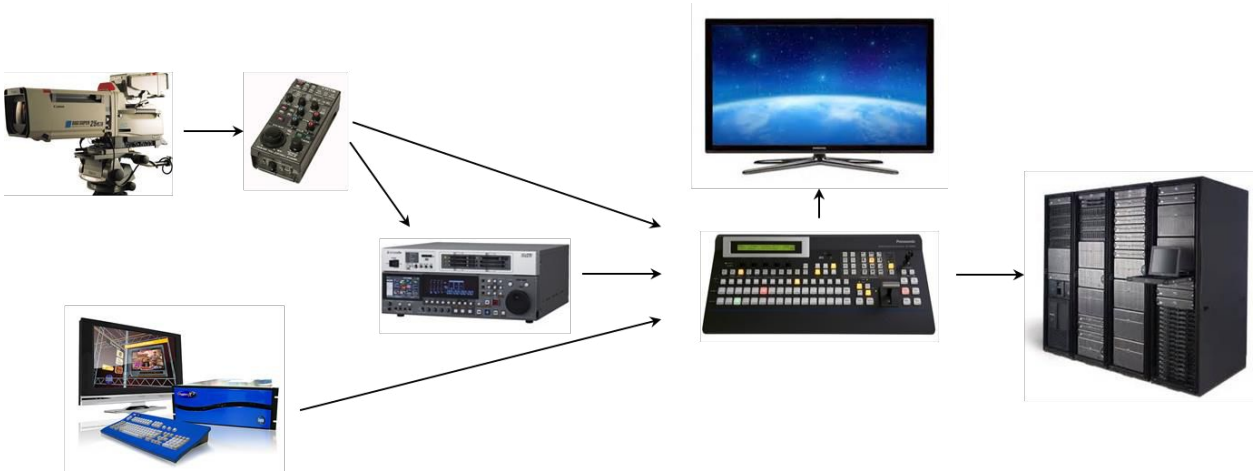


Figure 24 Example broadcast production facility components

Figure 25 shows a block diagram of the workflow in an HDR Broadcast facility using BT.2100 [5] PQ workflow. What is important to note is that in the transition phase from SDR to HDR, there will typically be a hybrid environment of both SDR and HDR devices and potentially a need to support both HDR and SDR outputs simultaneously. This is illustrated in the block diagram. In addition, because existing broadcast plants do not generally support metadata distribution today, the solution is to generate the ST 2094-10 [86] metadata in real time in just prior to, or inside of, the emission encoder as shown (block labeled “HPU” in brown in Figure 25). In the case of generation at the encoder, the display management metadata can be inserted directly into the bitstream using standardized SEI messages by the HPU. Each payload of the display management metadata message is about 500 bits. It may be sent once per scene, per GOP, or per frame. Note that the SEI message approach allows a production facility to utilize a common HDR10 bitstream, where one single stream is used for both HDR10 devices (which simply ignore the ST 2094-10 metadata) and Dolby Vision devices that correctly utilize the included metadata.

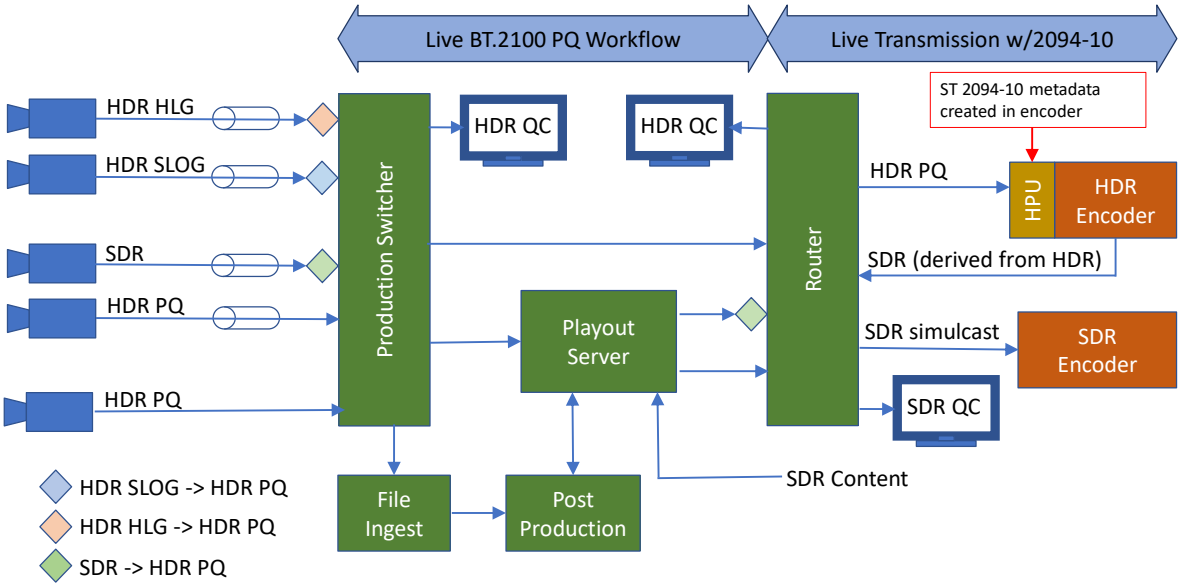


Figure 25 HDR broadcast production facility with BT.2100 PQ workflow-transition phase

SMPTE ST 2110-40 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 26 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.

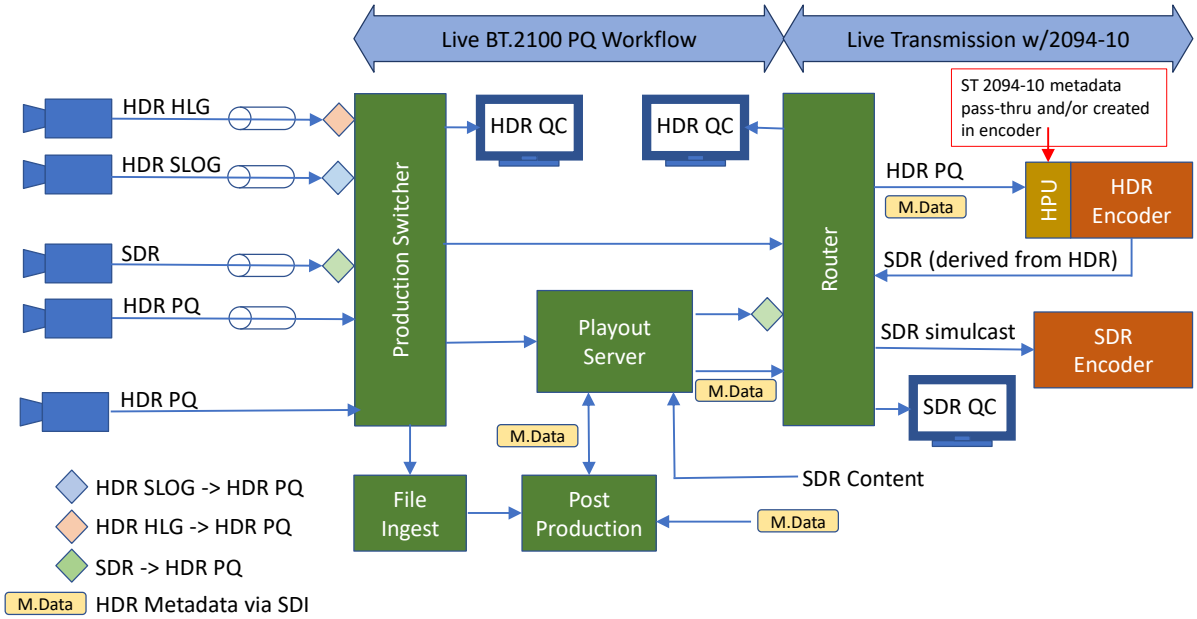


Figure 26 HDR broadcast production facility with BT.2100 PQ workflow-SDI metadata

12.2 Dual Layer HDR

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 [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 [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 in herein.

The combined BL and ELs should provide Foundation UHD 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 UHD 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 UHD content.

As an alternative, the base and ELs may be provided in an SDR format, which with metadata (see [33]) provided in either one of the two layers is decodable as HDR plus WCG, yet allows non-HDR devices to provide a picture with either just the BL, or both the base and ELs.

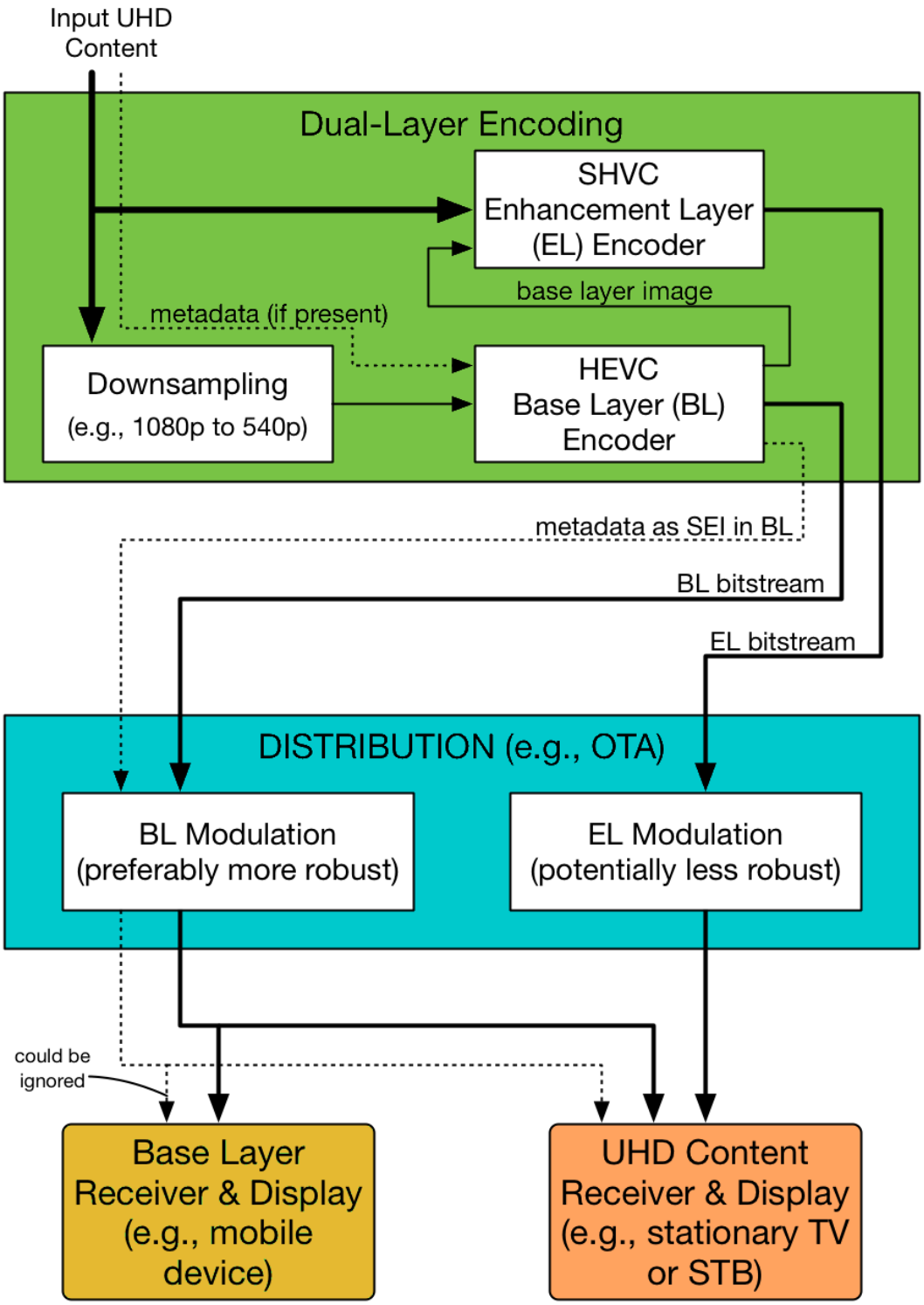


Figure 27 Example dual-layer encoding and distribution

Figure 27 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 [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.

12.3 SL-HDR1

As pointed out in Section 8.4, ETSI TS 103 433-1 [33] describes a method of down-conversion to derive an SDR/BT.709 signal from an HDR/WCG signal. The process supports PQ, HLG, and other HDR/WCG formats (see Section 6.3.2 of [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 [33]) to carry dynamic color volume transform metadata created during the down-conversion process. A receiver may use the SL-HDR information in conjunction with the SDR/BT.709 signal to reconstruct the HDR/WCG video.

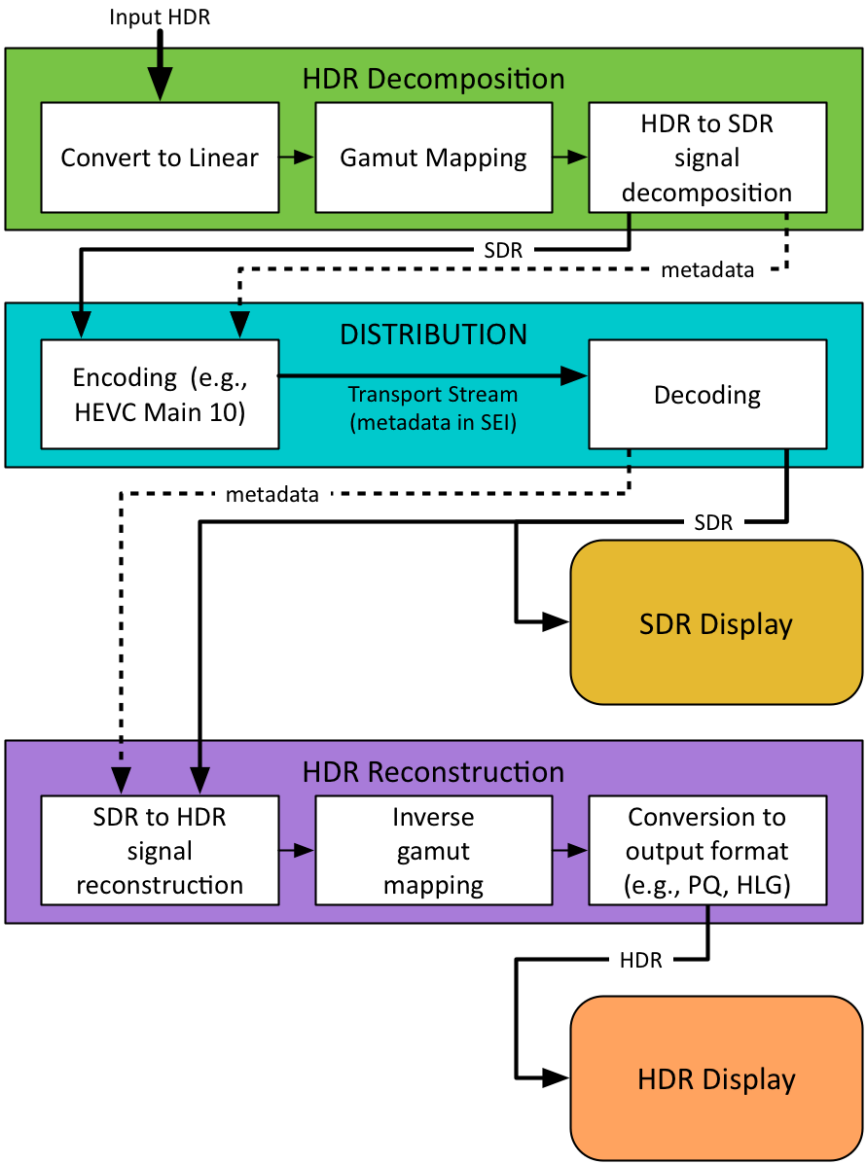


Figure 28 SL-HDR processing, distribution, reconstruction, and presentation

Figure 28 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 [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.

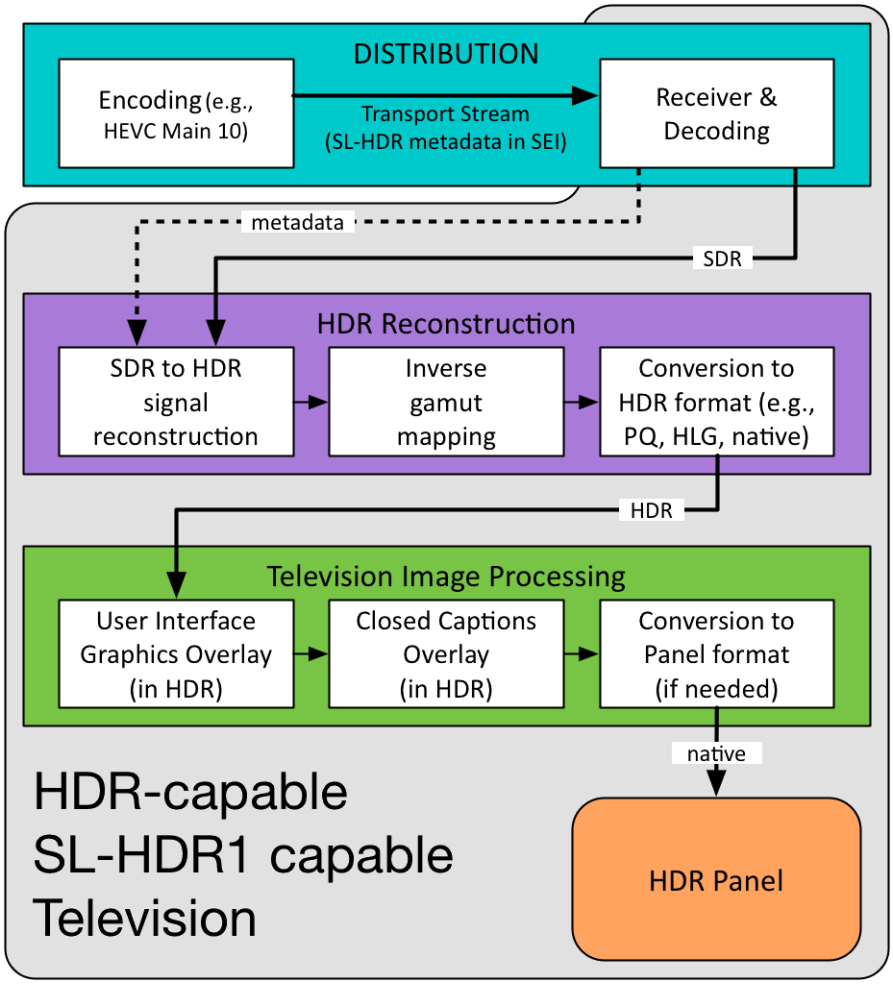


Figure 29 Direct reception of SL-HDR signal by an SL-HDR1 capable television

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 28 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 28.

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 29, the decoder recognizes metadata to be used to map the HDR/WCG video to an HDR format suitable for subsequent internal image processing (e.g., overlaying graphics and/or captions) before the images are supplied to the display panel.

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

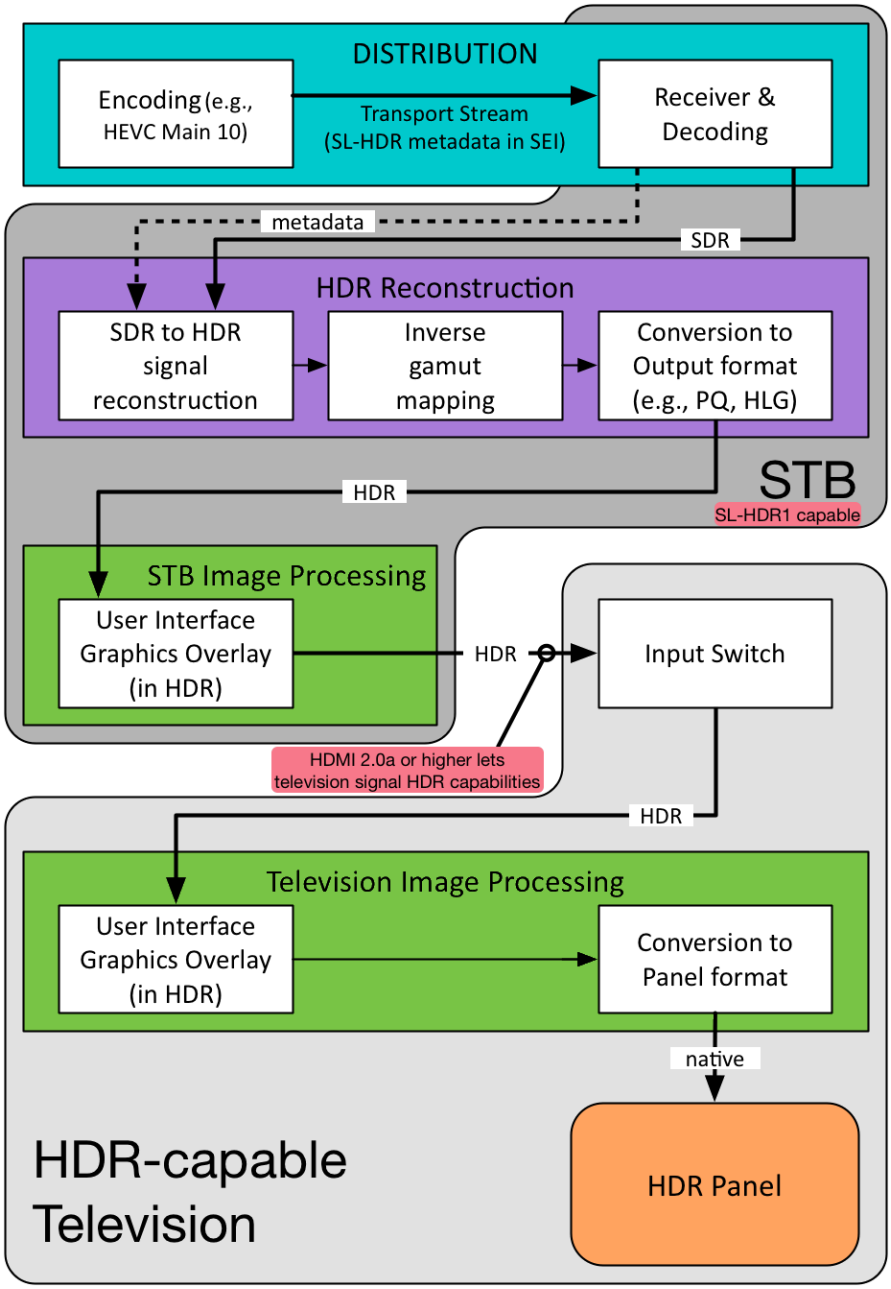


Figure 30 STB processing of SL-HDR signals for an HDR-capable television

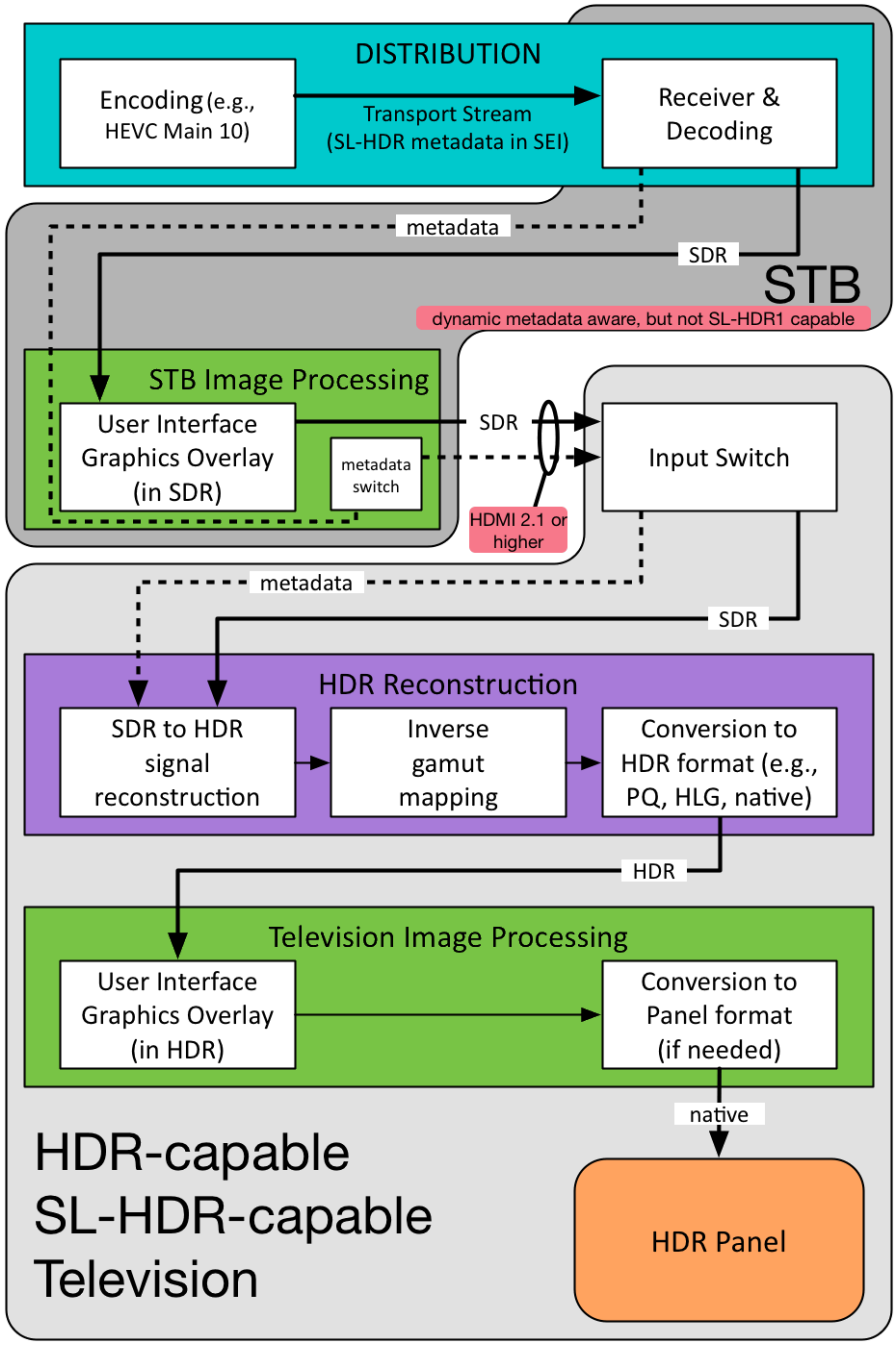


Figure 31 STB passing SL-HDR to an SL-HDR1 capable television

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 30, then the STB may query the interface with the display device (e.g., via HDMI 2.0a or higher, using the signaling described in [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 33.

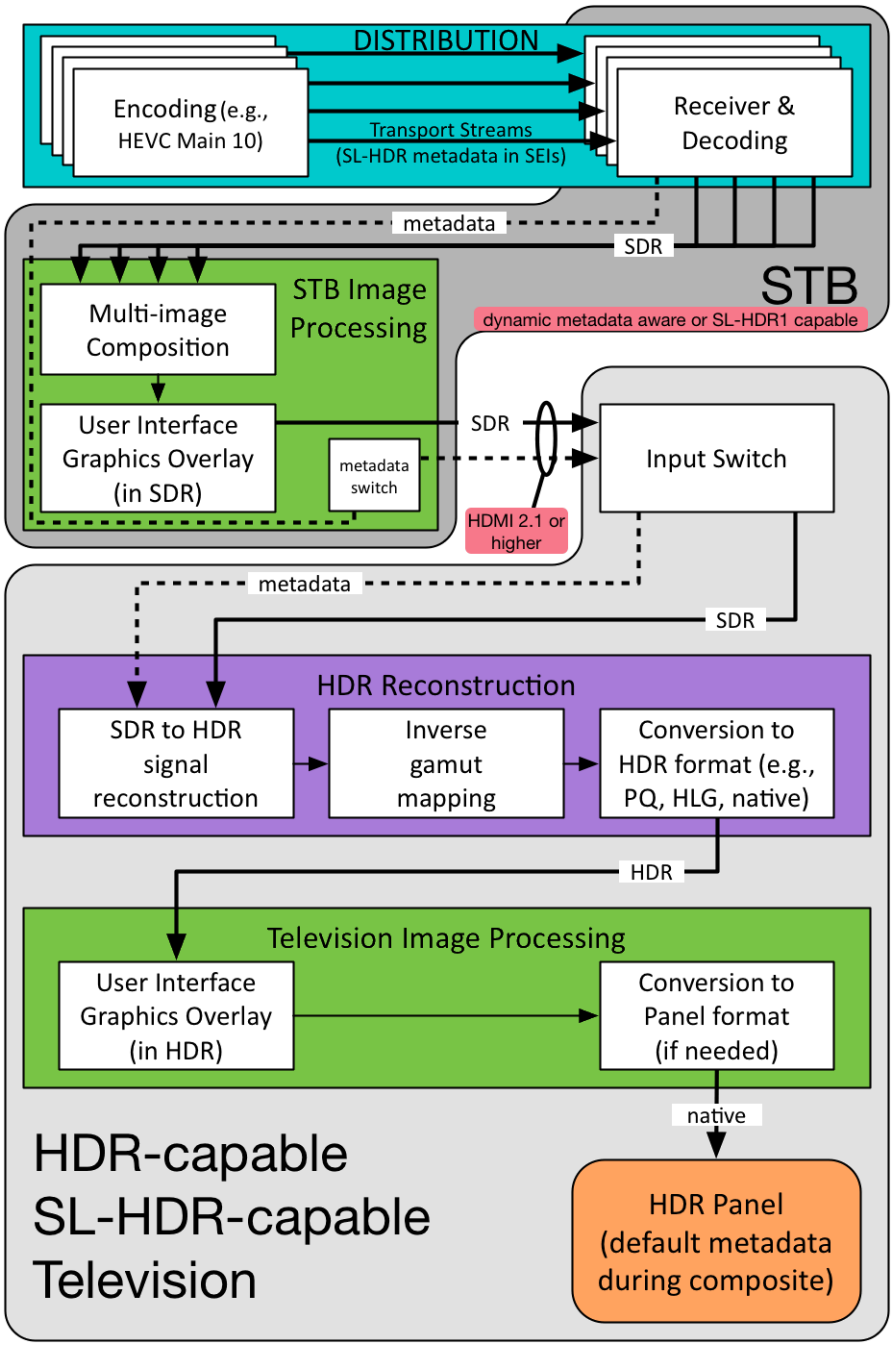


Figure 32 Multiple SL-HDR channels received and composited in SDR by an STB

If, as in Figure 31, 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 31. The first option is to retain the original SL-HDR information, which is dynamic. The second option is to revert to default values for the metadata as prescribed in Annex F of [33]. Either choice allows the display to maintain the same interface mode and does not induce a restart of the television's display processing pipeline, thereby not interrupting the user experience. The former choice, the dynamic metadata, may in rare cases produce a "breathing" effect that influences the appearance of only the STB-provided graphics. Television-supplied graphics are unaffected. Switching to the specified default values mitigates the breathing effect, yet allows the SL-HDR capable television to properly adapt the reconstructed HDR/WCG image to its display panel capabilities

Another use for the default values appears when multiple video sources are composited in an STB for multi-channel display, as when a user selects a multiple sports or news channels that all play simultaneously (though typically with audio only from one). This requires that multiple channels are received and decoded individually, but then composited into a single image, perhaps with graphics added, as seen in Figure 32. 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 28 shows HDR decomposition and encoding taking place in the broadcast facility immediately before emission. A significant benefit to this workflow is that there is no requirement for metadata to be transported throughout the broadcast facility when using the SL-HDR technique. For such facilities, the HDR decomposition is preferably integrated into the encoder fed by the HDR signal but, in the alternative, the HDR decomposition may be performed by a pre-processor from which the resulting SDR video is passed to an encoder that also accepts the SL-HDR information, carried for example as a message in SDI vertical ancillary data (as described in [86]) of the SDR video signal, for incorporation into the bitstream. Handling of such signals as contribution feeds to downstream affiliates and MVPDs is discussed below in conjunction with Figure 33 and Figure 34.

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.

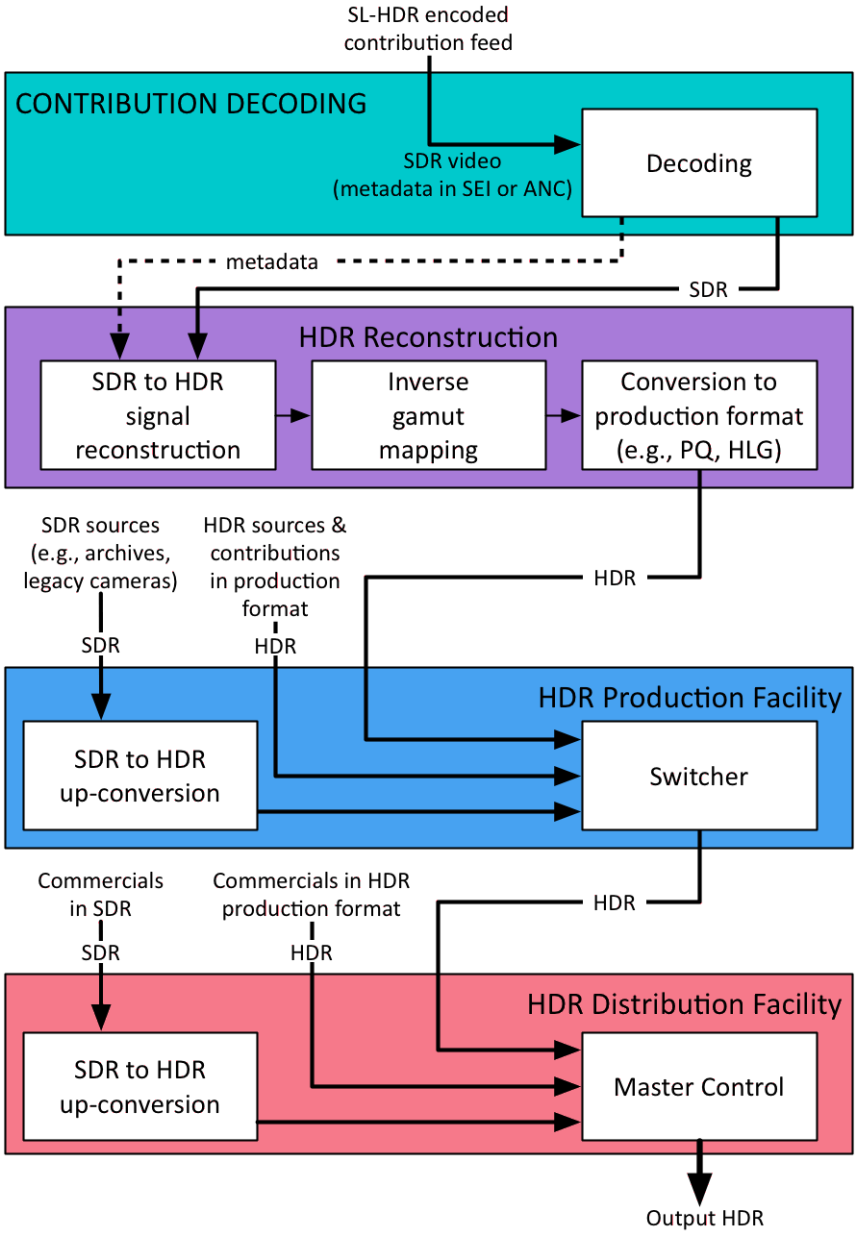


Figure 33 SL-HDR as a contribution feed to an HDR facility

The workflow for an HDR-ready affiliate receiving an SDR video with SL-HDR metadata as a contribution feed is shown in Figure 33. The decoding block and the HDR reconstruction block resemble the like-named blocks in Figure 28, with one potential exception: In Figure 33, the inverse gamut mapping block should use the invertible gamut mapping described in Annex D of [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 33, 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.

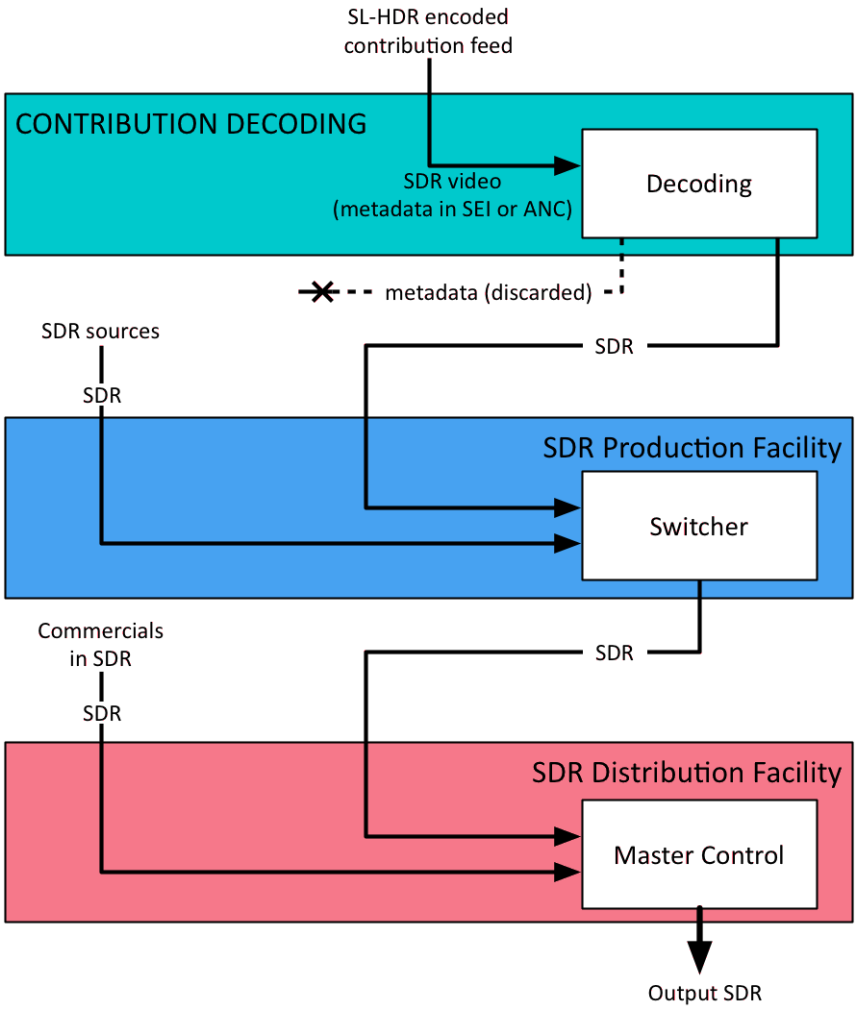


Figure 34 SL-HDR as a contribution feed to an SDR facility

In an HDR-based facility, the output HDR is complete immediately prior to the emission encode. As shown in Figure 28, 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 34 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 28 and detailed in Annex C of [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 28, by contrast, is considerably lighter weight computationally, and as such well suited to consumer premises equipment, and widely available for inclusion in hardware.

12.4 SL-HDR2

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

Generation and application of SL-HDR2 metadata is specified in ETSI TS 103 433-2 [34]. Typically, SL-HDR2 metadata is generated immediately prior to, or as a part of, distribution encoding, as shown in Figure 35, but SL-HDR2 metadata can also be generated upstream of the distribution encoder, e.g., as an encoding pre-process, and carried to the encoder as ST 2108-1 ANC messages [48] via SDI, or via IP using ST 2110-40 [47], or stored in file-based production infrastructures.

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

The SL-HDR information SEI message used to carry SL-HDR2 metadata is as specified in ETSI TS 103 433-1 (in Annex A.2 of [33]), but with the constraints specified in ETSI TS 103 433-2 [34].

Figure 35 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 [33], 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.

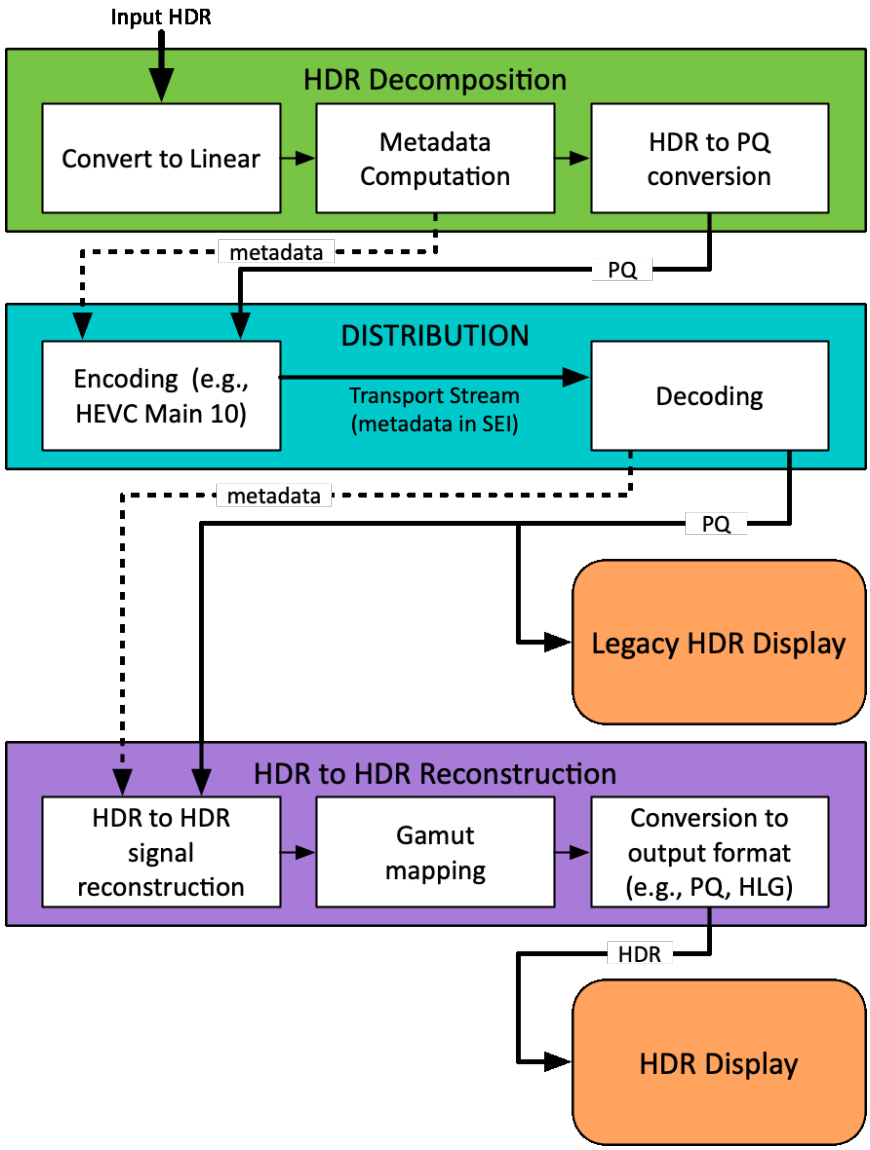


Figure 35 SL-HDR2 processing, distribution, reconstruction, for HDR presentation

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 35, 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 35 and Figure 36, respectively. An optional Gamut Mapping may be used during the reconstruction process if the presentation display is only able to support BT.709 images.

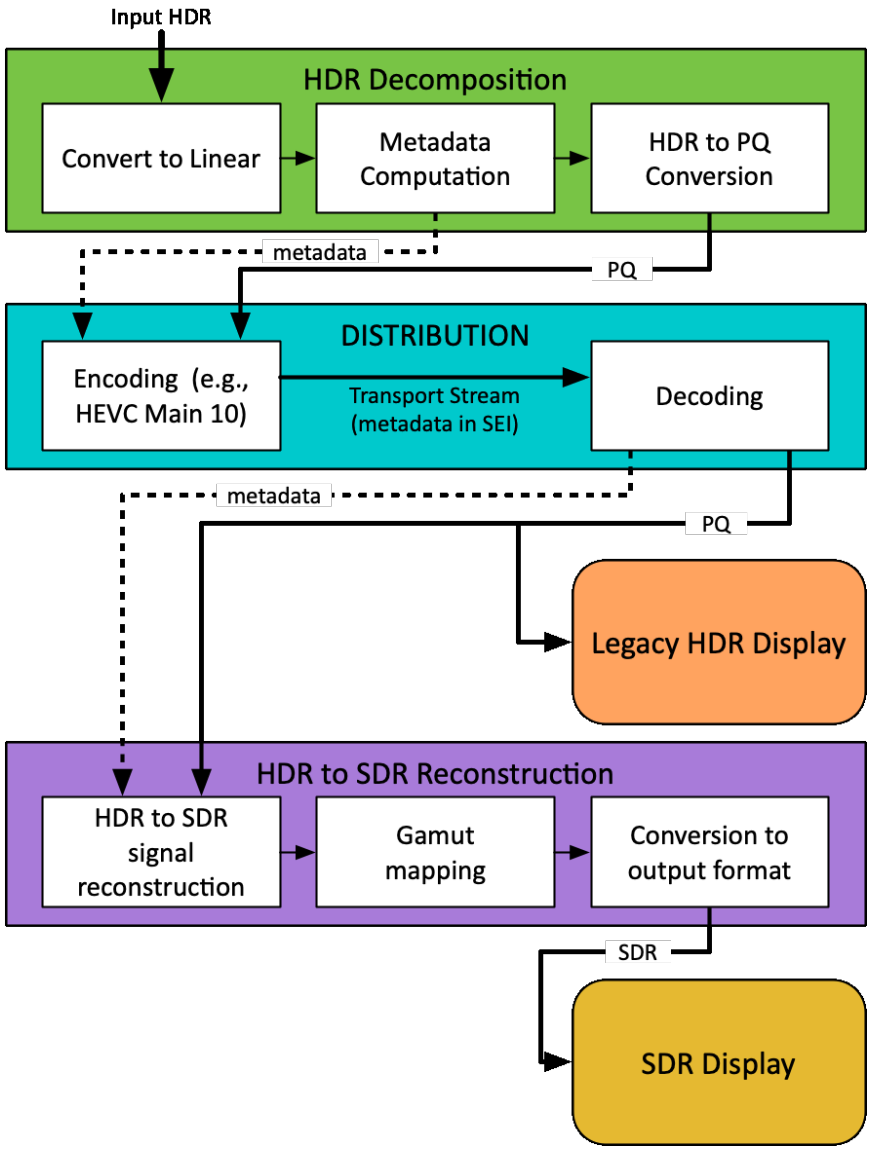


Figure 36 SL-HDR2 processing, distribution, reconstruction, and SDR presentation

The capability of this presentation display adaptation extends all the way to a downstream recipient having an SDR display, as shown in Figure 36, 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 35, and HDR to SDR Reconstruction process of Figure 36 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 37, the decoder recognizes metadata to be used to map the HDR/WCG video to an HDR format suitable for subsequent internal image processing (e.g., overlaying graphics and/or captions) before the images are supplied to the display panel.

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

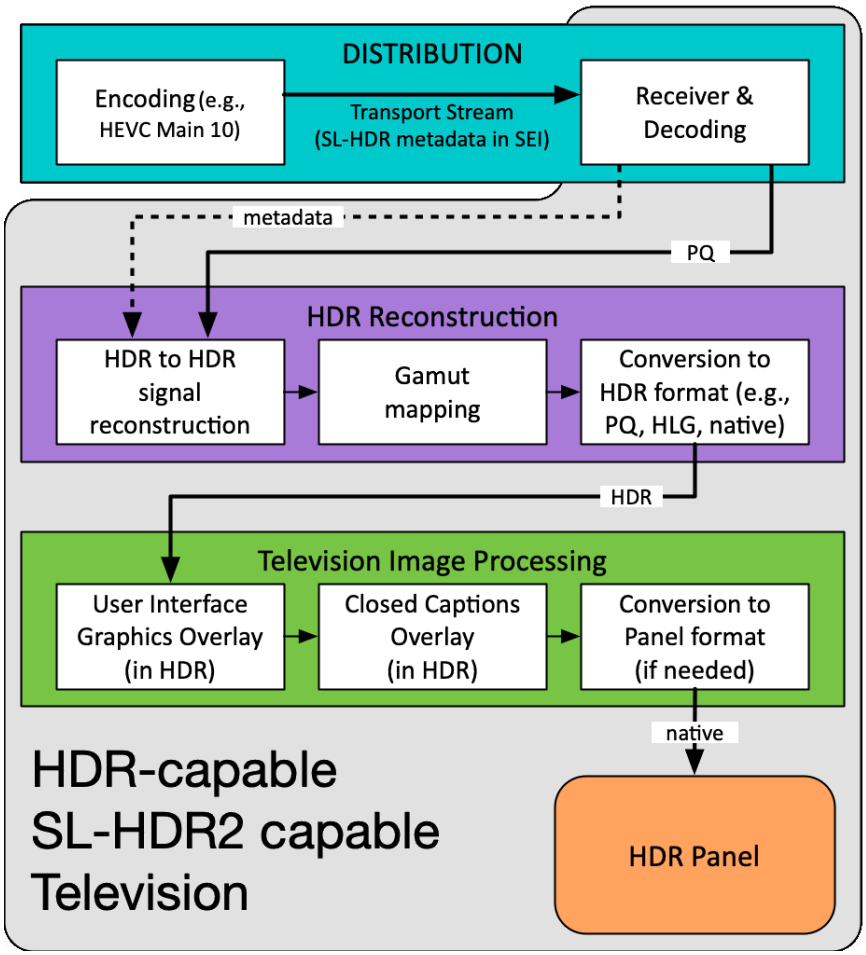


Figure 37 Direct reception of SL-HDR signal by an SL-HDR2 capable television

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 38, then the STB may query the interface with the display device (e.g., via HDMI 2.0a or higher, using the signaling described in [31]) to determine the display capabilities (HDR and corresponding peak luminance or SDR, gamut capabilities) that will serve in conjunction with the metadata to reconstruct, in an appropriate gamut and with an appropriate peak luminance, the HDR or SDR image to be passed to the display. If graphics are to be overlaid by the STB (e.g. captions, user interface menus or an EPG), the STB overlays graphics after the HDR reconstruction, such that the graphics are overlaid in the same mode that is being provided to the display.

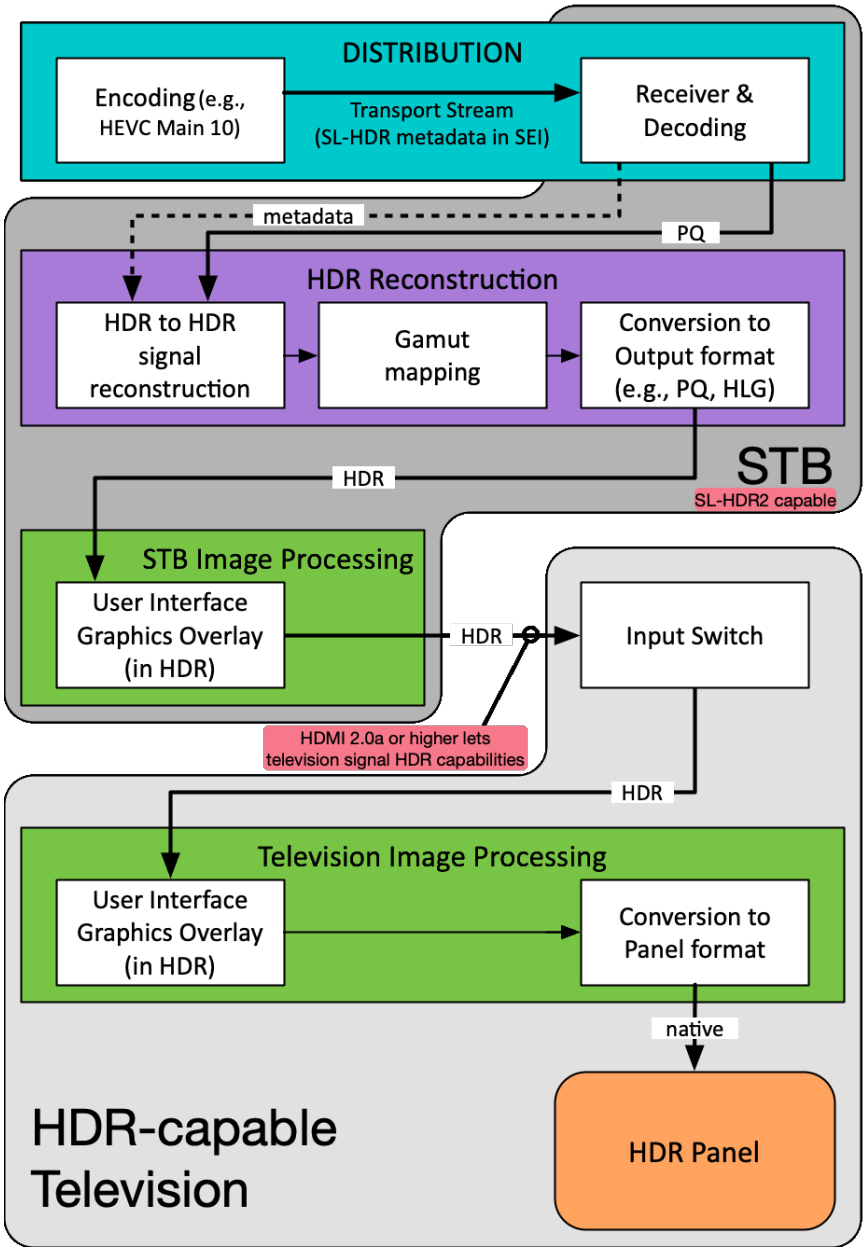


Figure 38 STB processing of SL-HDR signals for an HDR-capable television

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 41.

If, as in Figure 39, 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 [31]) that such information would be meaningful, then the STB may pass the SL-HDR information to the display in conjunction with the PQ video, enabling the television to reconstruct the HDR/WCG image.

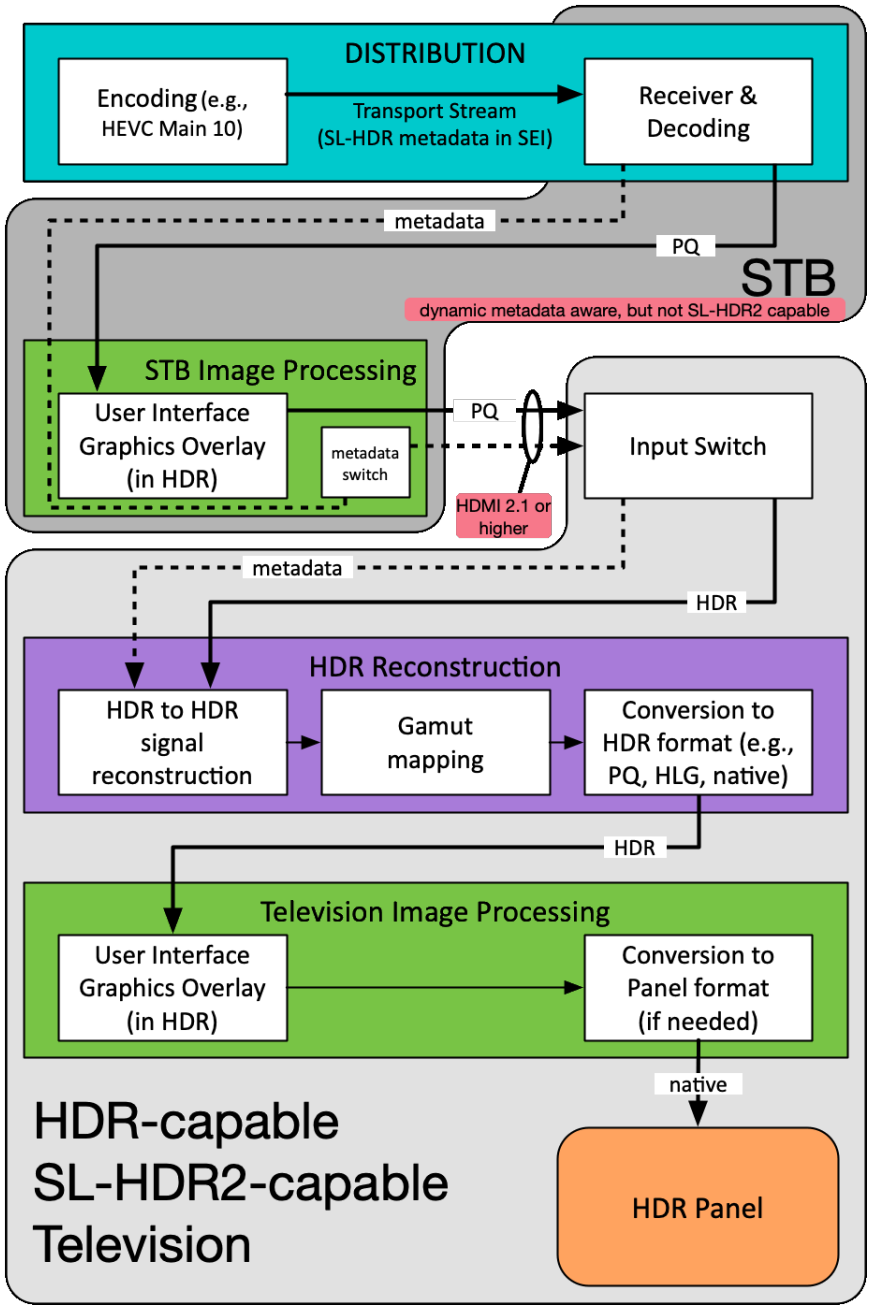


Figure 39 STB passing SL-HDR to an SL-HDR2 capable television

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 39. 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 [34]. Either choice allows the display to maintain the same interface mode and does not induce a restart of the television's display processing pipeline, thereby not interrupting the user experience. The former choice, the dynamic metadata, may in rare cases produce a "breathing" effect that influences the appearance of only the STB-provided graphics. Television-supplied graphics are unaffected. Switching to the specified default values mitigates the breathing effect, yet allows the SL-HDR capable television to properly adapt the reconstructed HDR/WCG image to its display panel capabilities

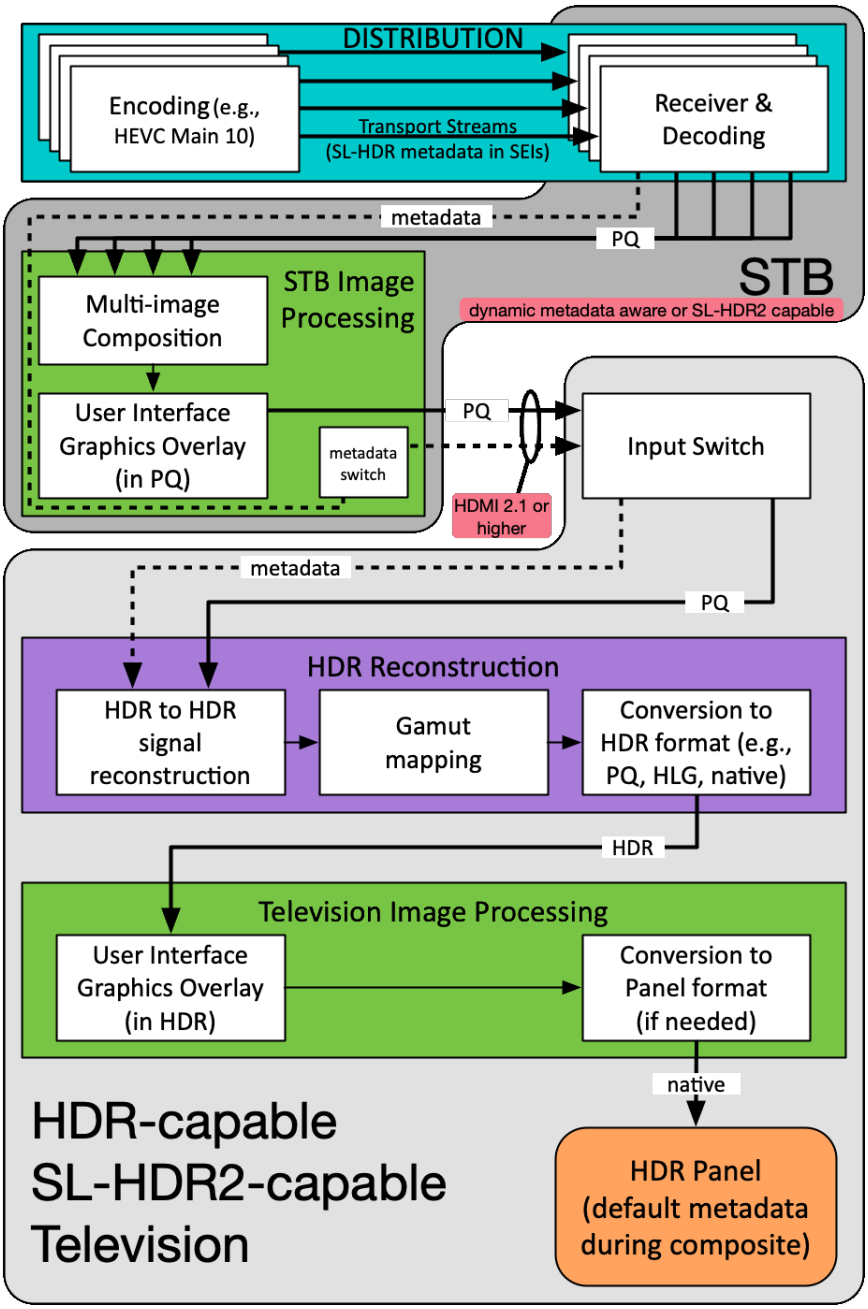


Figure 40 Multiple SL-HDR channels received and composited in HDR by an STB

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 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 40. 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 35 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 [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 41 and Figure 42.

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 nits while the original stream is 4000 nits) 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 41.

In HDR-based production and distribution facilities, such as shown in the example of Figure 41, 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.

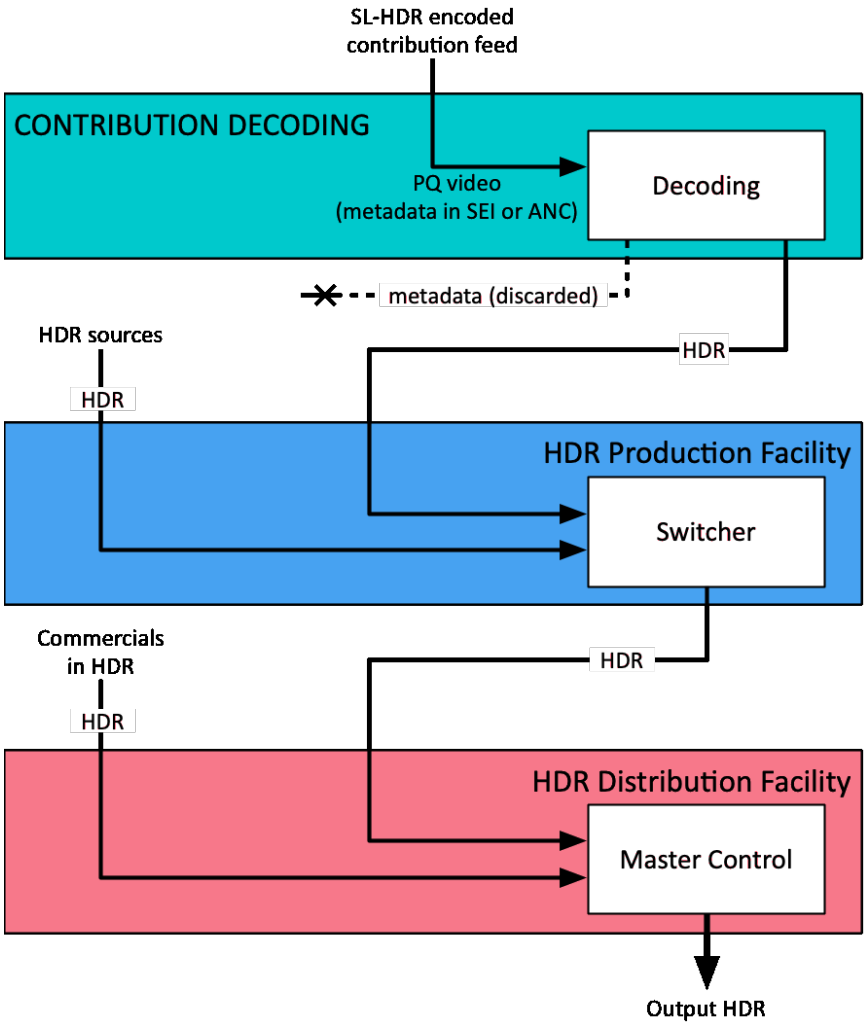


Figure 41 SL-HDR as a contribution feed to an HDR facility

In an HDR-based facility, the output HDR is complete immediately prior to the emission encode. As shown in Figure 35, 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 42 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 36, with one potential exception: In Figure 42, the Gamut mapping block should use the forward gamut mapping described in Annex D of [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 35 and detailed in Annex C of [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.

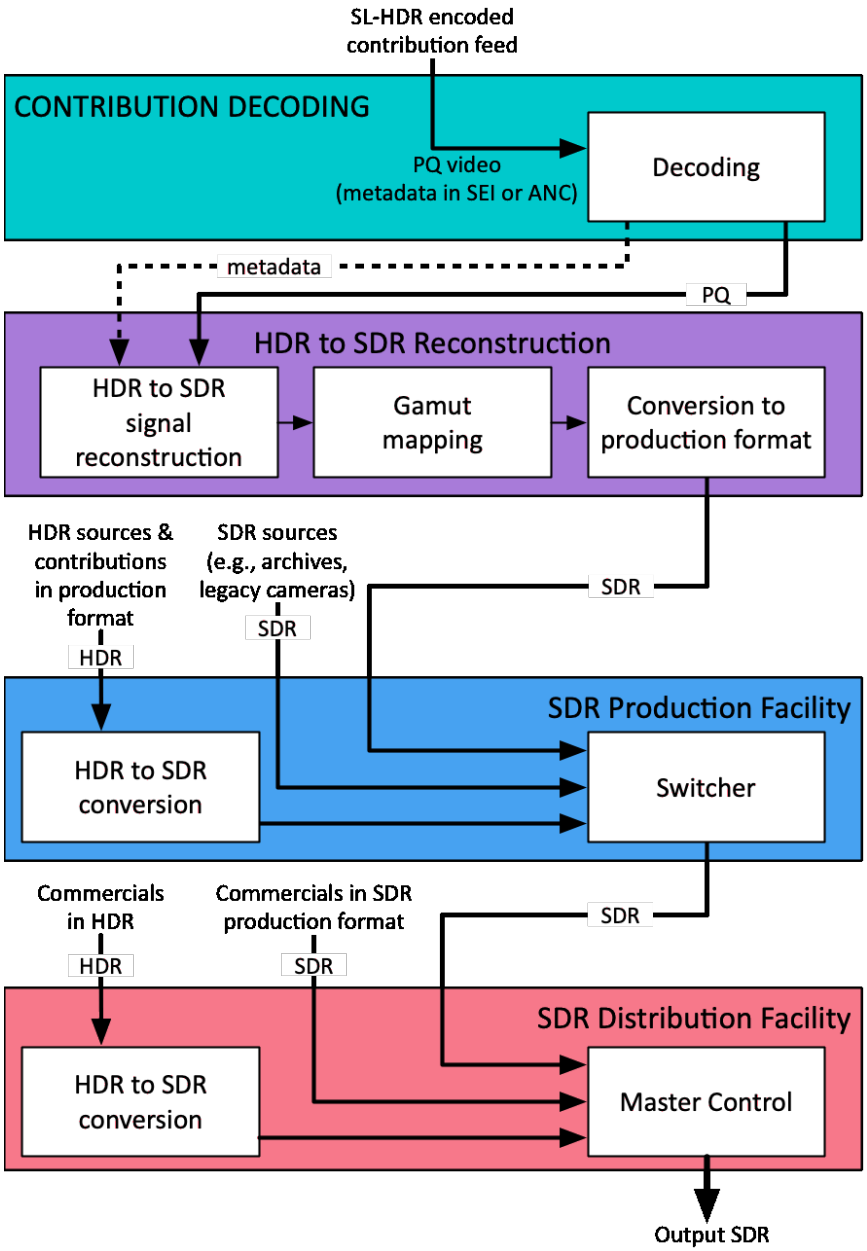


Figure 42 SL-HDR as a contribution feed to an SDR facility

13. High Frame Rate

13.1 Introduction

For the purpose of this document, High Frame Rate (HFR) refers to frame rates of 100fps or higher, including 100, 120/1.001[23] and 120, Standard Framer Rate (SFR) refers to frame rates of 60fps or lower, which are commonly used including 24/1.001, 24, 25, 30/1.001, 30, 50, 60/1.001 and 60.

According to a SMPTE/HPA paper authored by Mark Schubin[24], frame rates of 100, 120/1.001 and 120 fps add significant clarity to high motion video such as sports or action scenes. Schubin also notes that high dynamic range put new demands on temporal resolution. He notes that, "Viewers of HDR imagery sometimes report increased perception of motion judder... Increased frame rate, therefore, might be necessary to accompany HDR."

Citing Schubin again, "... the [EBU[25]] 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 43.

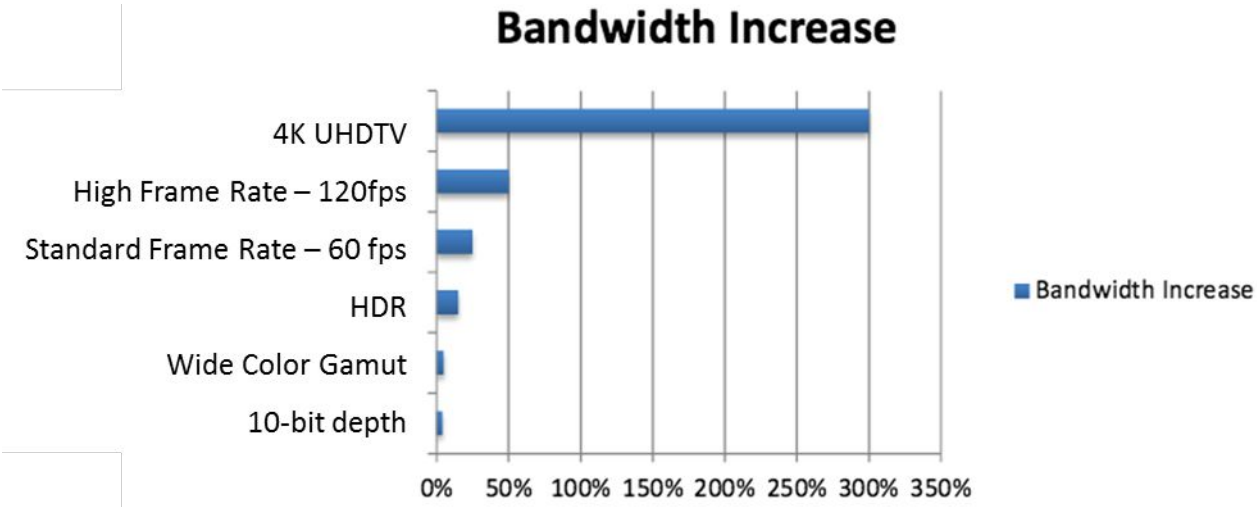


Figure 43 Bandwidth increases for various video format improvements

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 UHD technology that can be layered onto Foundation UHD 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

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

[24] "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

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

100/120 fps decoders render the full HFR experience. See Section 13.3 for more information about backward compatibility.

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

13.2 HFR Video Format Parameters

In the early 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 19.

Table 19 2K high frame rate content parameters

Frame Rate	Spatial Resolution	Scan Type	Dynamic Range	System Colorimetry	Bit Depth	Distribution Codec	HDMI Interface [26]
100, 120[27]	HD	Progressive	SDR, HDR	709, 2020	10	HEVC Main 10 Level 5.1	100fps: at least 1.4 120 fps: at least 2.0

13.3 Backward Compatibility for HFR

Both DVB UHD-1 Phase 2 (ETSI TS 101 154 v. 2.3.1) [63] and ATSC 3.0 (A/341) [54] include framerates up to 120 fps. Both documents further include optional temporal sub-layering for backward compatibility to a frame rate half of the HFR. 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 framerate picture when temporal sub-layering is used. Further framerate 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

[26] Earliest 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.

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

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 44 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.

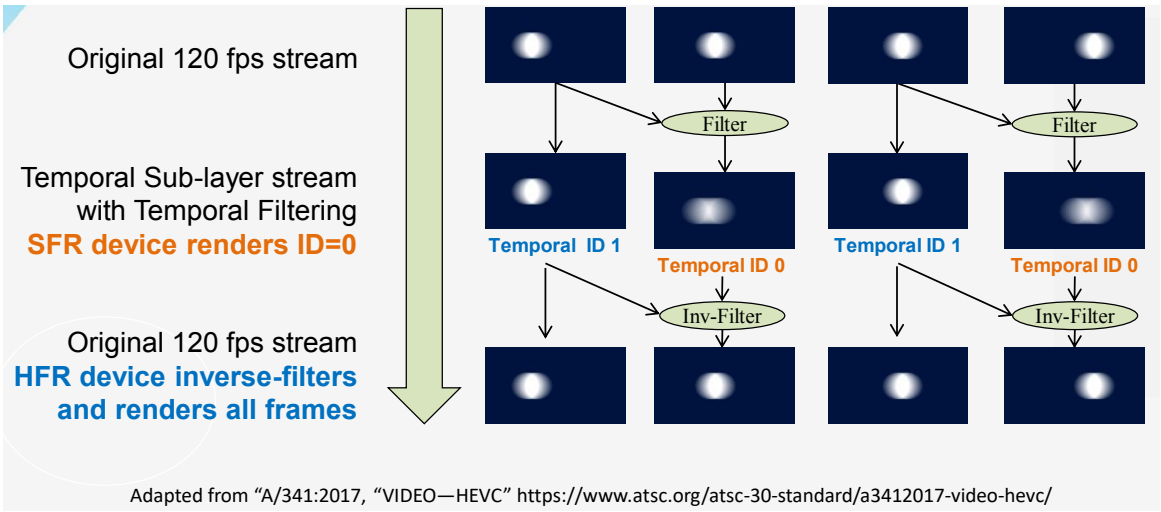
In the case of DVB, separate MPEG-2 TS Packet Identifiers (PIDs) are used to carry the two sub-layers. SFR decoders completely ignore the PID carrying the frames with Temporal ID = 1.

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

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

ATSC 3.0 includes an additional feature for backward compatibility called Temporal Filtering. 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 maybe 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 44.



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

Figure 44 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.

13.4 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 [81] and 2082 [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 [84] 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.

14. Next Generation Audio

14.1 Common Features of 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 alternate audio experiences, alternate languages, dialogue enhancement
- Consistent – Playback experience automatically optimized for each consumer device, e.g. home and mobile
- Object-based Audio – Audio elements are programmed to provide sound from specific locations in space, irrespective of speaker location. By delivering audio as individual elements, or objects, content creators can simplify operations, reduce bandwidth, and provide a premium experience for every audience
- Scene-Based Audio – An arbitrarily large number of directional audio elements composing a 3D sound field are mixed in a fixed number of PCM signals according to the Higher-Order Ambisonics format. Once in the HOA format, the Audio Scene can be efficiently transmitted, manipulated, and rendered on loudspeaker layouts/headphones/soundbars.
- Flexible Delivery - NGA can be delivered to consumers over a number of different distribution platforms including terrestrial, cable, and satellite broadcast, IPTV, OTT, and mobile. It could also be delivered over a hybrid of broadcast and OTT
- Flexible Rendering - NGA can be experienced by consumers through headphones or speakers (e.g., TV speakers, home theater systems including ceiling speakers, sound bars) as shown in Figure 45

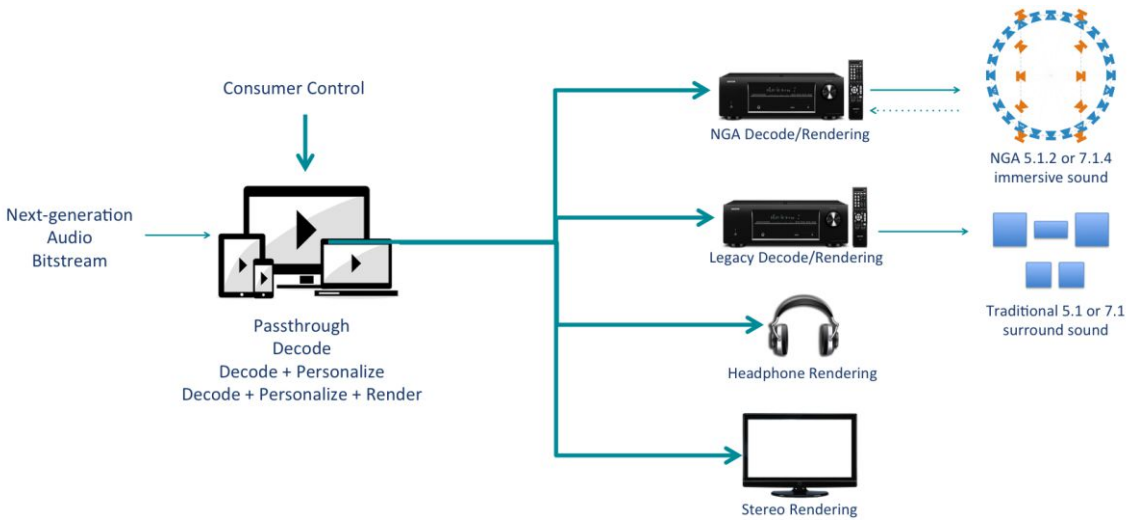


Figure 45 NGA in the consumer domain

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

14.1.1 NGA Use Cases

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

14.1.2 Audio Program Components and Preselections

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

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

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

14.1.3 Carriage of NGA

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

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

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

14.1.4 Metadata

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

14.1.5 Overview of Immersive Program Metadata and Rendering

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

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

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

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 ensures 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.

14.1.6 Audio Element Formats

The information contained in this section 14.1.6 is provided courtesy of the Advanced Television Systems Committee from Standard A/342 Part 1, Audio Common Elements. Readers may find the current version of the document at www.atsc.org.

The NGA systems support three fundamental Audio Element Formats:

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.

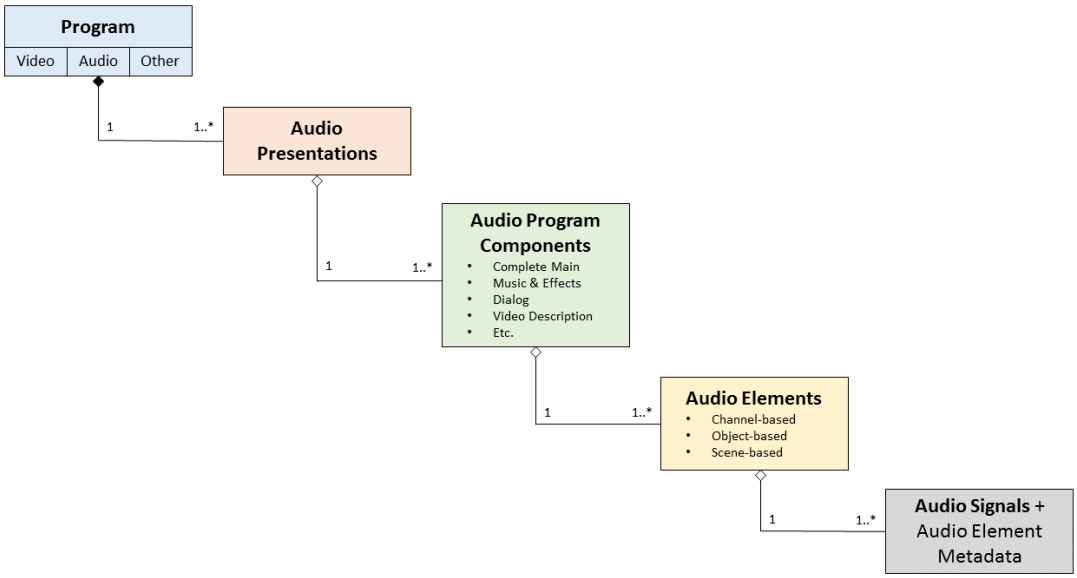


Figure 46 Relationship of key audio terms

Table 20 Mapping of terminology between NGA technologies

Common Term	DASH-IF Term [59]	AC-4 Term [56]	MPEG-H Audio Term [57]	DTS-UHD Term
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

14.1.7 Audio Rendering

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

14.1.7.1 Video Description Service (VDS)

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

14.1.7.2 Multi-Language

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

14.1.7.3 Personalized Audio

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

There are two main concepts of personalized audio:

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

14.2 MPEG-H Audio

14.2.1 Introduction

MPEG-H Audio is a Next Generation Audio (NGA) system offering true immersive sound and advanced user interactivity features. Its object-based concept of delivering separate audio elements with metadata within one audio stream enables personalization and universal delivery. MPEG-H Audio is an open international ISO standard and standardized in ISO/IEC 23008-3 [70]. The MPEG-H 3D Audio Low Complexity Profile Level 3 is adopted by DVB in ETSI TS 101 154 v.2.3.1 [63] and is one of the audio systems standardized for use in ATSC 3.0 Systems as defined in A/342 Part 3 [57]. SCTE has included the MPEG-H Audio System into the suite of NGA standards for cable applications as specified in SCTE 242-3 [77].

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 47, 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.

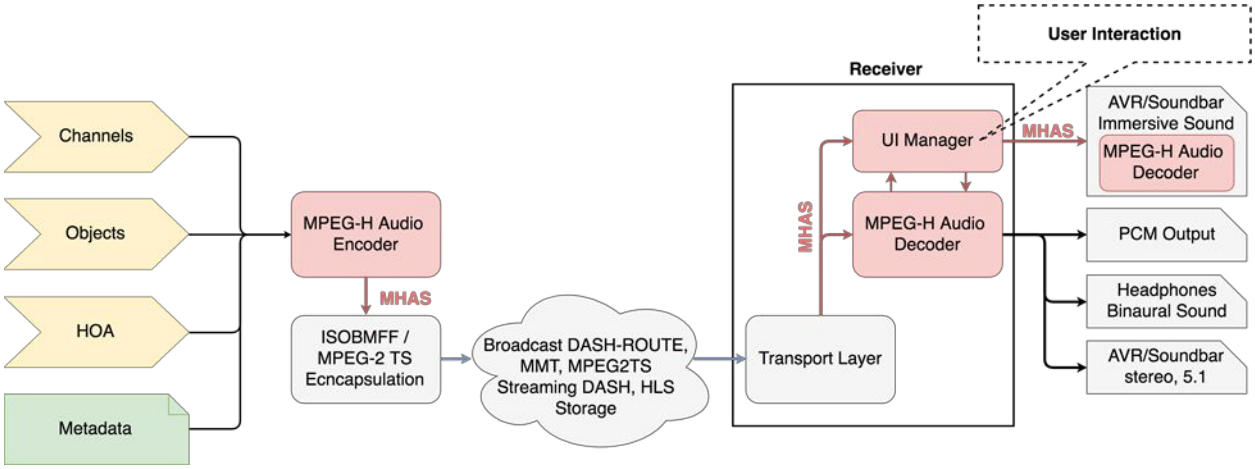


Figure 47 MPEG-H Audio system overview

The MPEG-H Audio Stream (MHAS), described in 14.2.3, contains the audio bitstream and various types of metadata packets and represents common layer for encapsulation into any transport layer format (e.g., MPEG-2 TS, ISOBMFF). The MPEG-H Audio enabled receiver can decode and render the audio to any loudspeaker configuration or a Binaural Audio representation for headphones reproduction. For enabling the advanced user interactivity features in cases where external playback devices are used, the UI Manger can supply the user interactions by inserting new MHAS packets into the MHAS stream and further deliver this over HDMI to the subsequent immersive AVR/Soundbar with MPEG-H Audio decoding capabilities. This is described in more detail in 14.2.1.4.

All MPEG-H Audio features that are described in the following sections are supported by the MPEG-H 3D Audio Low Complexity Profile Level 3 and are thus available in all broadcast systems based on the DVB and ATSC 3.0 specifications. See Table 21 for the characteristics of the Low Complexity Profile and levels.

Table 21 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

14.2.1.1 Personalization and Interactivity

MPEG-H Audio enables viewers to interact with the content and personalize it to their preference. The MPEG-H Audio metadata carries all the information needed for personalization such as attenuating or increasing the level of objects, disabling them, or changing their position. The metadata also contains information to control and restrict the personalization options such as setting the limits in which the user can interact with the content, as illustrated in Figure 48. (See also section 7.4.3 MPEG-H Audio Metadata.)

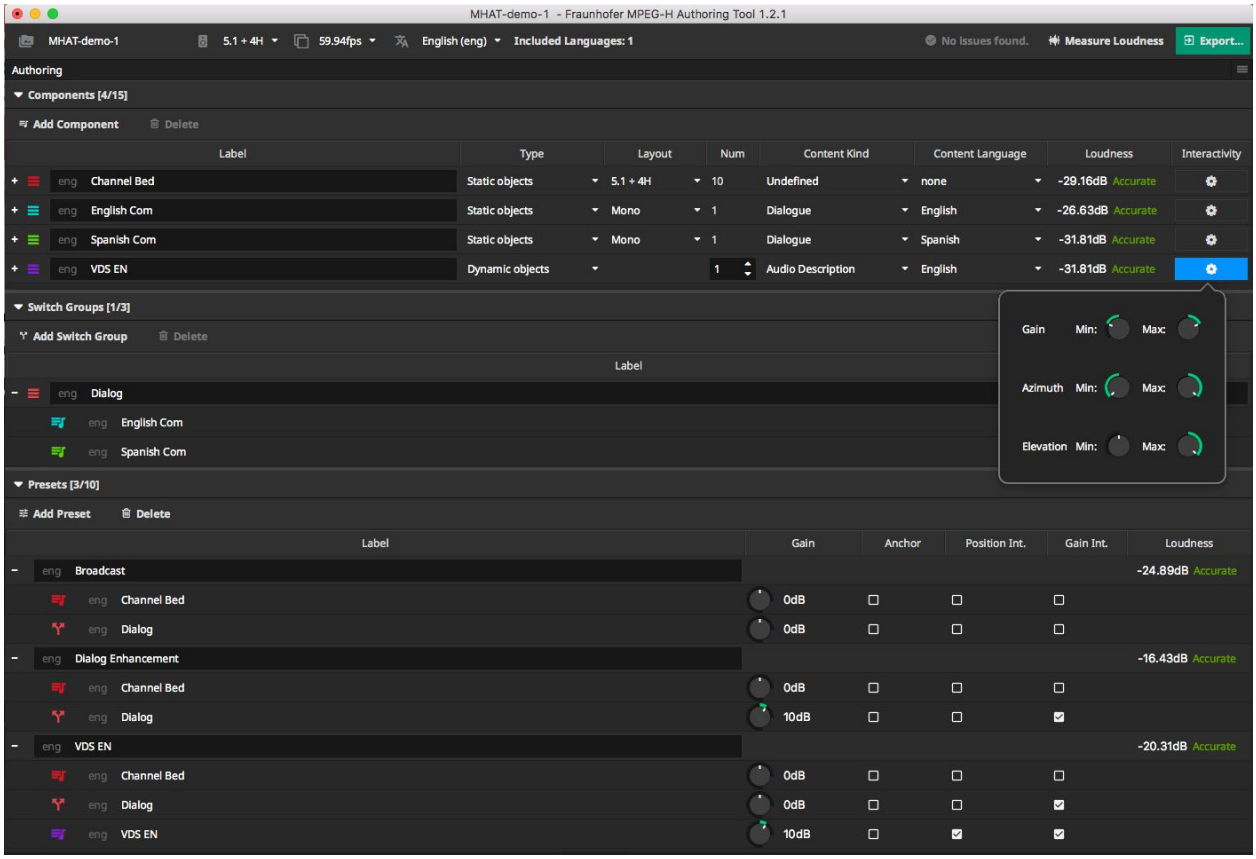


Figure 48 MPEG-H Authoring Tool example session

14.2.1.2 Universal delivery

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

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

14.2.1.3 Immersive Sound

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

The MPEG-H 3D Audio Low Complexity Profile Level 3 allows up to 16 audio elements (channels, objects or HOA signals) to be decoded simultaneously, while up to 32 audio elements can be carried simultaneously in one stream (see Table 21).

14.2.1.4 Distributed User Interface Processing

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

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

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

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

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

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

Figure 49 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.

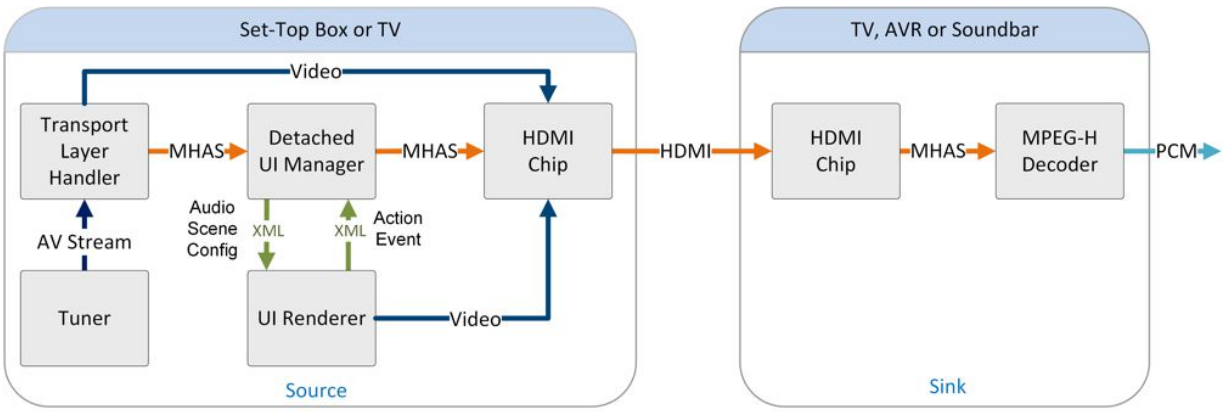


Figure 49 Distributed UI processing with transmission of user commands over HDMI

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.

14.2.2 MPEG-H Audio Metadata

MPEG-H Audio enables NGA features such as personalization and interactivity with a set of static metadata, the “Metadata Audio Elements” (MAE). Audio Objects are associated with

metadata that contain all information necessary for personalization, interactive reproduction, and rendering in flexible reproduction layouts. This metadata is part of the overall set-up and configuration information for each piece of content.

14.2.2.1 Metadata Structure

The metadata (MAE) is structured in several hierarchy levels. The top-level element is the Audio Scene Information or the "AudioSceneInfo" structure as shown in Figure 50. Sub-structures of the AudioSceneInfo contain descriptive information about "Groups", "Switch Groups", and "Presets." An "ID" field uniquely identifies each group, switch group or preset, and is included in each sub-structure.

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

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

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

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

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

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

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

14.2.2.2 Metadata Example

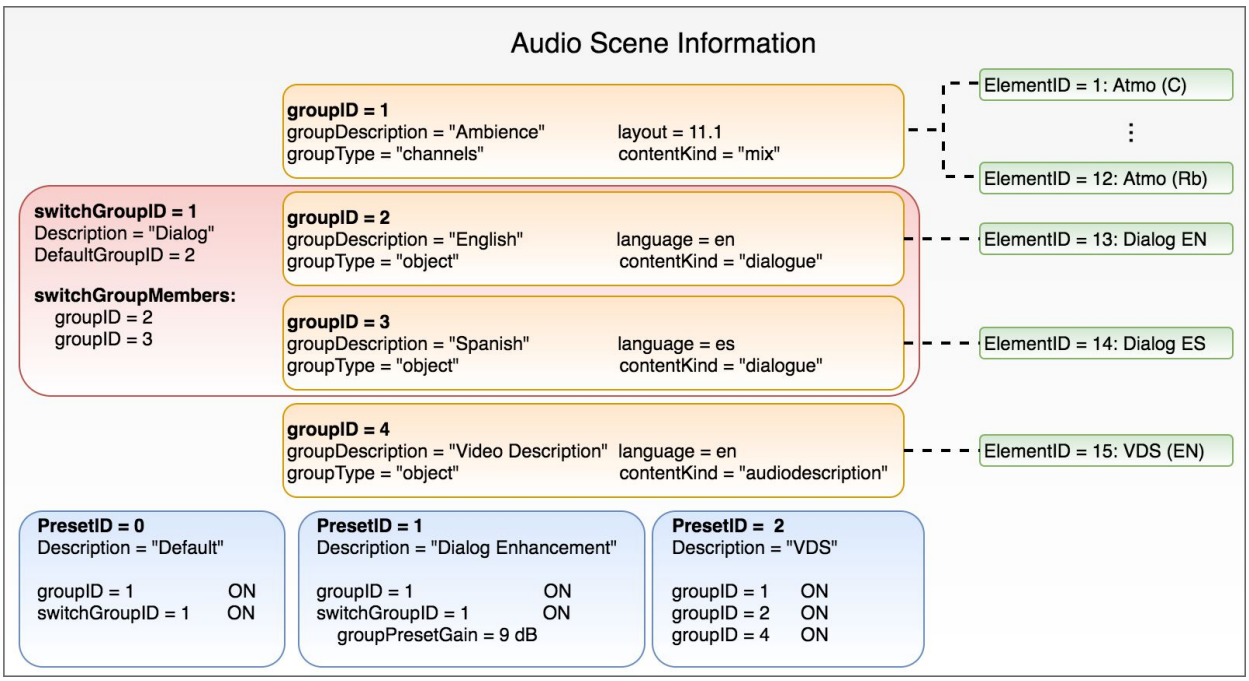


Figure 50 Example of an MPEG-H Audio Scene information

Figure 50 contains an example of MPEG-H Audio Scene Information with four different groups (orange), one switch group (red) and three presets (blue). In this example, the switch group contains two dialogs in different languages that the user can choose from, or the device can automatically select one dialog based on the preference settings.

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

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

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

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

14.2.2.3 Personalization Use Case Examples

Advanced Accessibility

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

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

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

Additionally, MPEG-H Audio allows the user to spatially move the Video Description object to a user selected position (e.g., to the left or right), enabling a spatial separation of main dialog and the Video Description element, as shown in Figure 51. This results in a better intelligibility of the main dialog as well as the Video Description (e.g., in a typical 5.1 set-up the main dialog is assigned to the center speaker while the Video Description object could be assigned to a rear-surround speaker).

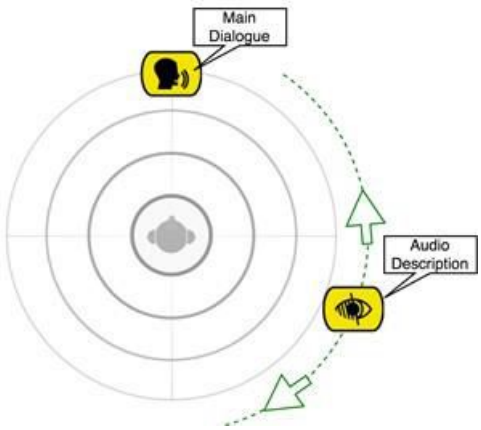


Figure 51 Audio description re-positioning example

Dialog Enhancement (DE)

MPEG-H Audio includes a feature of DE that enables automatic device selection (prioritization) as well as user manipulation. For ease of user selection or for automatic device selection (e.g., enabling TV "Hard of Hearing" TV setting), a Dialog Enhancement preset can be created, as illustrated in Figure 50 using a broadcaster defined enhancement level for the dialog element (e.g., 10dB as shown in Figure 48).

Moreover, if the broadcaster allows personalization of the enhancement level, MPEG-H Audio supports advanced DE which enables direct adjustment of the enhancement level via the user interface. The enhancement limitations (i.e., maximum level) are defined by the broadcaster/content creator as shown in Figure 48 and carried in the metadata. This maximum value for the lower and upper end of the scale can be set differently for different elements as well as for different content.

The advanced loudness management tool of the MPEG-H Audio system automatically compensates loudness changes that result from user interaction (e.g., switching presets or enhancement of dialogue) to keep the overall loudness on the same level, as illustrated in Figure 52. This ensures constant loudness level not only across programs but also after user interactions.

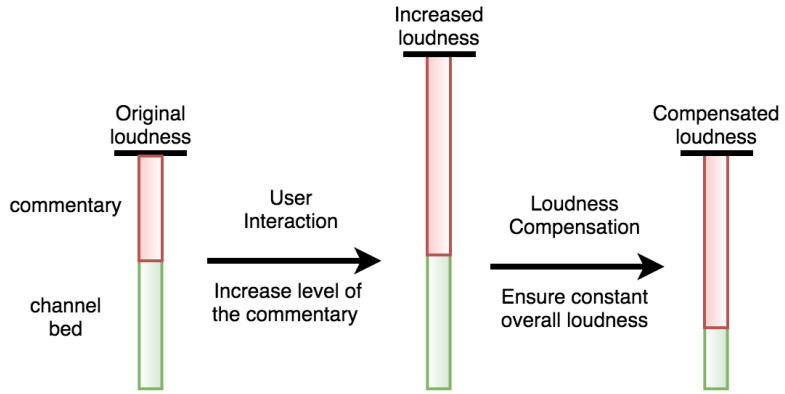


Figure 52 Loudness compensation after user interaction

Multi-language services

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

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

Personalization for Sport Programs

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

14.2.3 MPEG-H Audio Stream

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

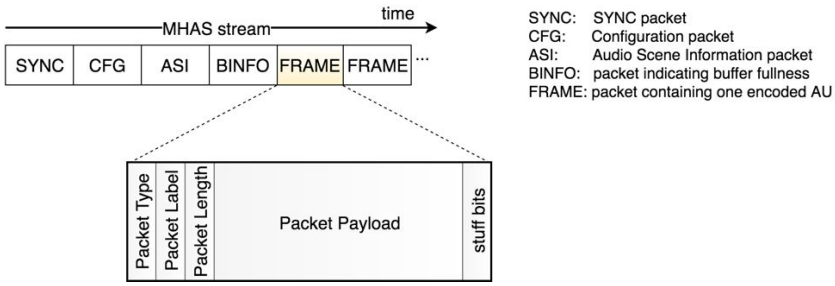


Figure 53 MHAS packet structure

Figure 53 shows the high-level structure of an MHAS packet, which contains the header with the packet type to identify each MHAS packet, a packet label and length information, followed by the payload and potential stuffing bits for byte alignment.

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

14.2.3.1 Random Access Point

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

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

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

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

14.2.3.2 Configuration Changes and A/V Alignment

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

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

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

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

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

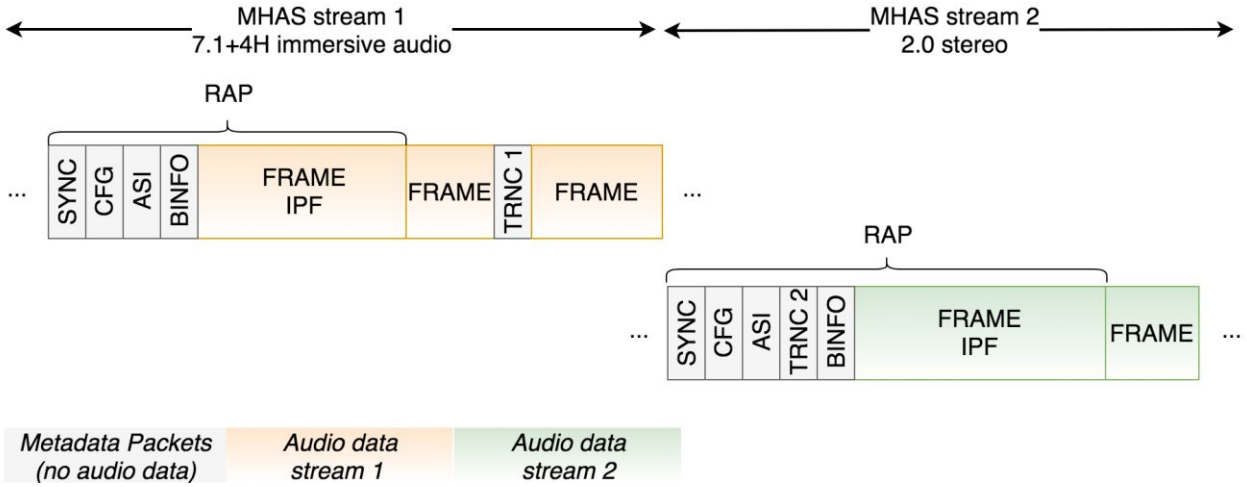


Figure 54 Example of a configuration change from 7.1+4H to 2.0 in the MHAS stream

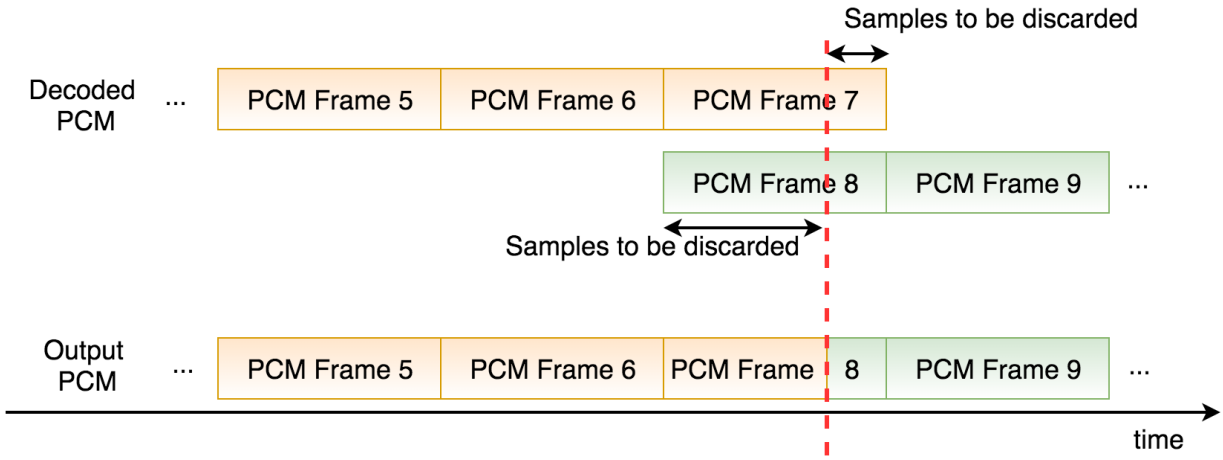


Figure 55 Example of a configuration change from 7.1+4H to 2.0 at the system output

Figure 54 and Figure 55 show an example of a sample-accurate configuration change from an immersive audio set-up to stereo inside one MHAS stream. (I.e., in the ad-insertion use case the inserted ad is stereo, while the rest of the program is in 7.1+4H.)

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

14.3 Dolby AC-4 Audio

AC-4 is a audio system from Dolby Laboratories, which brings a number of features beyond those already delivered by the previous generations of audio technologies, including Dolby Digital® (AC-3) and Dolby Digital Plus (EAC-3). Dolby AC-4 is designed to address the

current and future needs of next-generation video and audio entertainment services, including broadcast and Internet streaming.

The core elements of Dolby AC-4 have been standardized with the European Telecommunications Standards Institute (ETSI) as TS 103 190 [65] and adopted by Digital Video Broadcasting (DVB) in TS 101 154 [63] and are ready for implementation in next generation services and specifications. AC-4 is one of the audio systems standardized for use in ATSC 3.0 Systems [56]. AC-4 is specified in the ATSC 3.0 next-generation broadcast standard (A/342 [55]) and has been selected for use in North America (U.S., Canada and Mexico) as described in A/300 [51].

Furthermore, Dolby AC-4 enables experiences by fully supporting Object-based Audio (OBA), creating significant opportunities to enhance the audio experience, including immersive audio and advanced personalization of the user experience. As shown in Figure 56, 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.

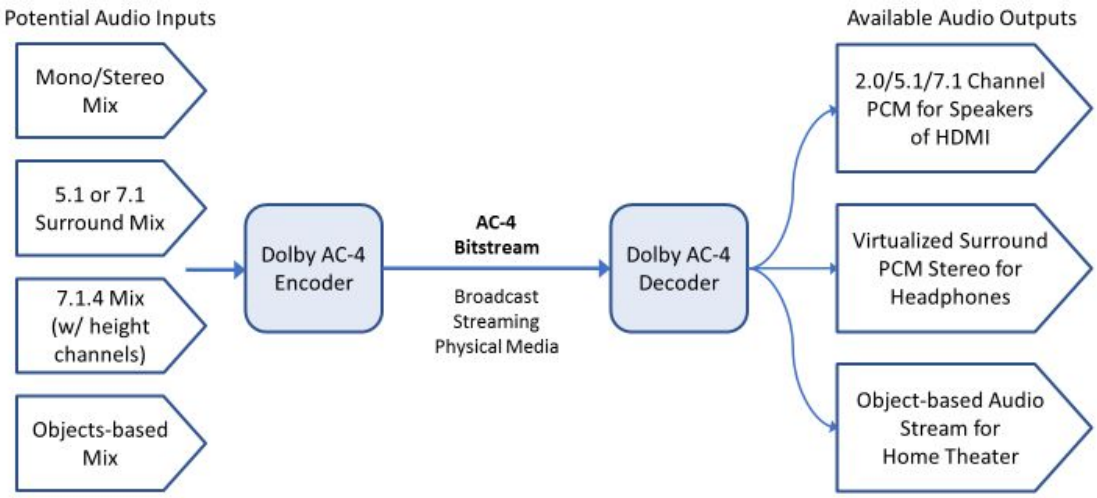


Figure 56 AC-4 Audio system chain

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 22 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 22 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

14.3.1 Dynamic Range Control (DRC) and Loudness

Loudness management in AC-4 includes a novel end-to-end signaling framework along with a real-time adaptive loudness processing mechanism that provide the service provider with an intelligent and automated system that ensures the highest quality audio while remaining compliant with regulations anywhere in the world. Compliant programming delivered to an AC-4 encoder with valid metadata will be encoded, preserving the original intent and compliance (see Figure 57). If the metadata is missing or the source cannot be authenticated, the system switches to an "auto pilot" mode, running a real-time loudness leveler (RTLL) to generate an ITU-R loudness-compliant gain offset value for transmission in the AC-4 bitstream. That gain offset value is automatically applied in the playback system. When compliant programming returns, the RTLL process is inaudibly bypassed. AC-4 is also designed to ensure that loudness compliance is maintained when several substreams are combined into a single presentation (see section III) upon decoding, e.g. M&E+D, or Main+Associated presentations.

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

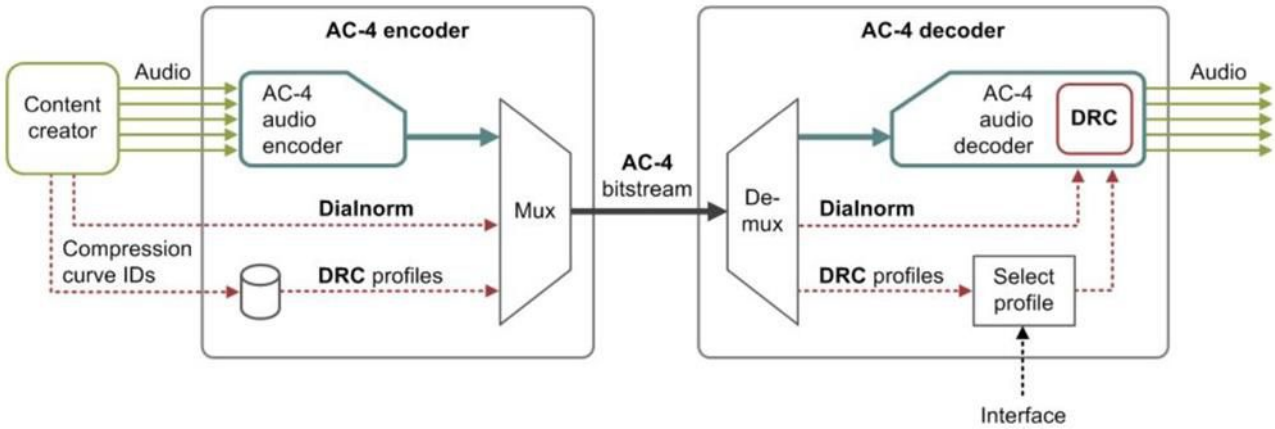


Figure 57 AC-4 DRC generation and application

A flexible DRC solution is essential to serve the wide range of playback devices and playback environments, from high-end Audio Video Receiver (AVR) systems and flat-panel TVs in living rooms to tablets, phones and headphones on-the-go. The AC-4 system defines four independent DRC decoder operating modes that correspond to specific Target Reference Loudness, as shown in Table 23.

Table 23 Common target reference loudness for different devices

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

14.3.2 Hybrid Delivery

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

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

14.3.3 Backward Compatibility

Dolby Atmos audio programs can be encoded using the AC-4 [65] codec or the E-AC-3+JOC [35] codec. When Atmos is used with E-AC-3+JOC streams, backward compatibility is provided for existing non-Atmos E-AC-3 [29] decoders. See Section 11.5 for details. Backward compatibility is achieved in a different way: an AC-4 decoder (e.g., an ATSC 3.0 television or an advanced AVR) can provide a multichannel PCM audio (plus metadata) downmix which is delivered over HDMI and correctly interpreted by current Atmos-enabled devices (e.g., a soundbar) to produce a full Dolby Atmos immersive experience. If the destination renderer only supports stereo or 5.1 channel audio, it will correctly provide a downmix to those legacy formats.

14.3.4 Next Generation Audio Metadata and Rendering

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

- Immersive program metadata – informs Object-based Audio rendering and includes parameters such as position and speaker-dependencies
- Personalized program metadata – specifies audio presentations and defines the relationships between audio elements
- Essential Metadata Required for Next-Generation Broadcast:
    - Intelligent Loudness Metadata – metadata to signal compliance with regional regulations, dialogue loudness, relative-gated loudness, loudness correction type, etc.
    - Program Synchronization – metadata to allow other sources/streams to be synchronized with the primary (emitted) presentation with frame-based accuracy.
    - Legacy Metadata – traditional metadata including dialnorm, DRC, downmixing for Channel-based Audio, etc.

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

14.3.5 Overview of Immersive Program Metadata and rendering

14.3.5.1 Object-based Audio Rendering

Object audio renderers also include control over the perceived object size (see “object width” metadata parameter in Section 14.3.7.3), which provides mixers with the ability to create the

impression of a spatially extended source, which can be controlled within the same frame of reference (see Figure 58).

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

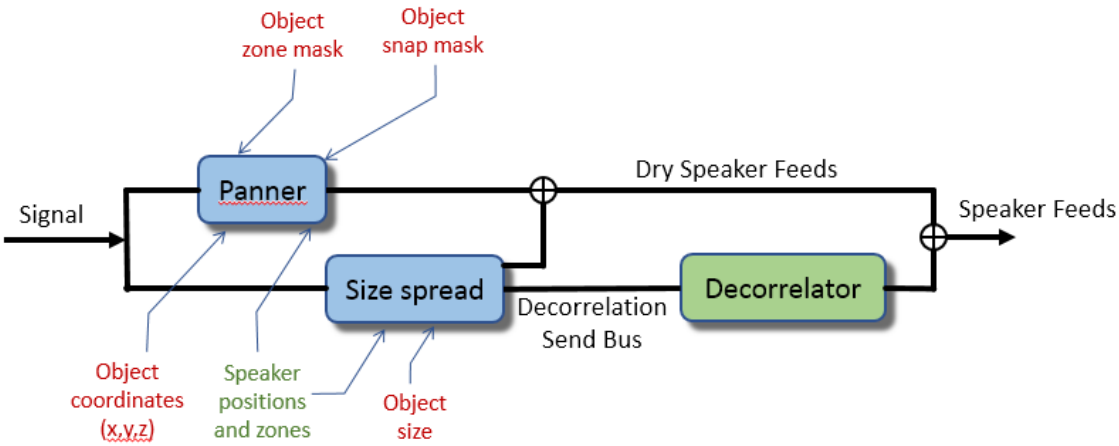


Figure 58 Object-based audio renderer

At the core of rendering are pan and spread operators, each executing a panning algorithm (see Figure 58) responsive to an audio object's coordinates (x,y,z). Most panning algorithms currently used in Object-based Audio production attempt to recreate audio cues during playback via amplitude panning techniques where, a gain vector G[1..n] is computed and assigned to the source signal for each of the *n* loudspeakers. The object audio signal s(t) is therefore reproduced by each loudspeaker *i* as $G_i(x,y,z) \times s(t)$ in order to recreate suitable localization cues as indicated by the object (x,y,z) coordinates and spread information as expressed in the metadata. There are multiple panning algorithms available to implement $G_i(x,y,z)$.

The design of panning algorithms ultimately must balance tradeoffs among timbral fidelity, spatial accuracy, smoothness and sensitivity to listener placement in the listening environment, all of which can affect how an object at a given position in space will be perceived by listeners. For instance, Figure 59 illustrates how different speakers maybe utilized among various rendering (panning) algorithms to place an object's perceived position in the playback environment.

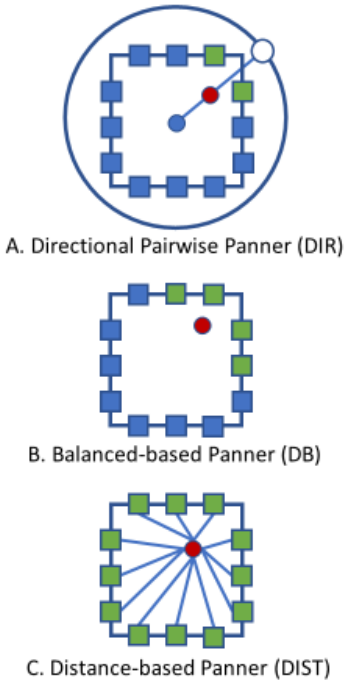


Figure 59 Common panning algorithms

Directional pairwise panning (DIR) (see Figure 59-A) is a commonly used strategy that solely relies on the directional vector from a reference position (generally the sweet spot or center of the room) to the desired object's position. The pair of speakers 'bracketing' the relevant directional vector is used to place (render) that object's position in space during playback. A well-documented extension of directional pairwise panning to support 3D loudspeaker layouts is vector-based amplitude panning which uses triplets of speakers. As this approach only utilizes the direction of the source relative to a reference position, it cannot differentiate between object sources at different positions along the same direction vector. It can also introduce instabilities as objects are panned near the center of the room. Moreover, some 3D implementations may constrain the rendered objects to the surface of a unit sphere and thus would not necessarily allow an object to cross inside the room without going 'up and over'. DIR can cause sharp speaker transitions as objects approach the center of the room with the result that rendering whips around from one side of a room to the opposite, momentarily tagging all the speakers in between.

The balanced-based (DB) panning algorithm, also known as "dual-balanced" is the most common approach used in 5.1/7.1-channel surround productions today (Figure 59-B). This approach utilizes left/right and front/back pan pot controls widely used for surround panning. As a result, dual-balance panning generally operates on the set of four speakers bracketing the desired 2D object position.

Extending to three-dimensions (e.g., when utilizing a vertical layer of speakers above the listener) yields a "triple-balance" panner. It generates three sets of one-dimensional gains corresponding to left/right, front/back and top/bottom balance values. These values can then be multiplied to obtain the final loudspeaker gains:

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

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

In contrast to the directional and balance-based approaches, distance-based panning (DIST) (Figure 59-C) uses the relative distance from the desired 2D or 3D object location to each speaker in use to determine the panning gains. Thus, this approach generally utilizes all the available speakers in use rather than a limited subset, which leads to smoother object pans but with the tradeoff of being prone to timbral artefacts, which can make the sound seem unnatural.

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

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

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

14.3.5.2 Rendering-control metadata

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

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

To achieve this, AC-4 employs additional metadata to dynamically reconfigure the object renderer to "mask out" certain speaker zones during playback of a particular audio object. This is shown as the zone mask metadata in Figure 58. This guarantees that no loudspeaker belonging to the masked zones will be used for rendering the applicable object. Typical zone masks used in production today include: *no sides, no back, screen only, room only and elevation on/off*.

The main application of speaker zone mask metadata is to help the mixer achieve a precise control of which speakers are used to render each object in order to achieve the desired perceptual effect. For instance, the *no sides* mask guarantees that no speaker on the side wall of the room will be used. This creates more stable screen-to-back fly-throughs. If the side speakers are used to render such trajectories, they will become audible for the seats nearest to the side walls and these seats will perceive a distorted trajectory "sliding" along the walls rather than crossing the center of the room.

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

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

Another useful aesthetic control parameter is the *snap-to-speaker mode* represented by snap mask metadata (see Figure 58). The mixer can select this mode for an individual audio object to indicate that consistent reproduction of timbre is more important than consistent reproduction of the object's position. When this mode is enabled, the object renderer does not perform phantom panning to locate the desired sound image. Rather, it renders the object entirely from the single loudspeaker nearest to the intended object location.

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

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

14.3.6 Overview of Personalized Program Metadata

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

14.3.6.1 Presentation Metadata

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

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

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

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

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

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

14.3.7 Essential Metadata Required for Next-Generation Broadcast

This section provides a high-level overview of the most essential metadata parameters required for enabling next-generation audio experiences. Essential metadata is capable of being interchanged for both file-based workflows (as per the ITU-R BWF/ADM formats [72], [73]) AND in serialized form for real-time workflows and interconnects utilizing SMPTE ST 337 [36] formatting/framing.

14.3.7.1 Intelligent Loudness Metadata

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

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

Loudness Practice Type - This parameter indicates which recommended practice was followed when the content was authored or corrected. For example, a value of "0x1" indicates the author (or automated normalization process) was adhering to ATSC A/85 [24]. A value of "0x2" indicates the author was adhering to EBU R 128 [62]. A special value, "0x0" signifies that the loudness recommended practice type is not indicated.

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

Dialogue Gating Practice Type - This parameter indicates what dialogue gating practice was followed when the content was authored or corrected. This parameter is typically 0x02 – "Automated Left, Center and/or Right Channel(s)". However, there are values for signaling manual selection of dialogue, as well as other channel combinations, as detailed in the ETSI TS 103 190 [65].

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

Program Loudness, Relative Gated - This parameter is entered into the encoder to indicate the overall program loudness as per ITU-R BS.1770-4 [37]. In ATSC regions, this parameter would typically be -24.0 LKFS for short-form content as per ATSC A/85 [24]. In EBU regions, this parameter would typically indicate -23.0 LKFS (LUFS).

Program Loudness, Speech Gated - This parameter indicates the speech-gated program loudness. In ATSC regions, this parameter would typically be -24.0 LKFS for long-form content as per ATSC A/85 [24].

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

max_truepk - This parameter indicates the maximum true peak value for the audio program measured per ITU-R BS.1770 [37].

loro_dmx_loud_corr - This parameter is used to calibrate the downmix loudness (if applicable), as per the Lo/Ro coefficients specified in the associated metadata and/or emission bitstream, to match the original (source) program loudness. Note: this parameter is not currently supported in the pending ITU-R BWF/ADM [72] format.

ltrt_dmx_loud_corr - This parameter is used to calibrate the downmix loudness (if applicable), as per the Lt/Rt coefficients specified in the associated metadata and/or emission bitstream to match the original (source) program loudness. Note: This parameter is not currently supported in the pending ITU-R BWF/ADM [72] format.

Note regarding the loudness measurement of objects: The proposed system supports loudness estimation and correction of both Channel-based and Object-based (immersive) programs utilizing the ITU-R BS.1770-4 [37] recommendation.

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

14.3.7.2 Personalized Metadata

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

The primary controls for personalization are:

- Presentation selection
- Dialogue element volume level

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

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

14.3.7.3 Object Audio Metadata

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

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

The primary purpose of the object audio metadata is to:

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

Within the production system, a subset of the object audio metadata fields is essential to provide the best audio experience and to ensure that the original artistic intent is preserved. The remaining non-essential metadata fields described in ETSI TS 103 190-2 [65] are used for either an enhanced playback application or aiding in the transmission and playback of the program content.

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

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

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

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

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

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

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

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

Zone / Elevation mask - The zone and elevation mask metadata fields describe which speakers, either on the listener plane or height plane of the playback environment, shall be enabled or disabled during rendering for a specific object. Each speaker in the playback

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

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

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

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

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

14.3.8 Metadata Carriage

In the production system, different methods are introduced for enabling the carriage of metadata described above within file-based and real-time (HD-SDI) contribution/distribution workflows to address a wide range of industry needs related to interoperability and reliability necessary for day-to-day operations.

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

With the growing interest across the worldwide broadcast industry to enable delivery of both immersive and personalized (interactive) experiences, additional information (i.e., the metadata) must co-exist to describe fully these experiences. The EBU Audio Definition Model [61] has provided the foundation for the development of an international recommendation within ITU-R WP6B, which produced the ITU-R Audio Definition Model (ADM) [72].

The ITU-R ADM [72] specifies how XML data can be generated to provide definitions of tracks and associated metadata within Broadcast Wave (BWF), RF64 files or as a separate file that references associated essence files. In general, the ADM describes the associated audio program as two parts via the XML. The *content* part describes what is contained in the audio (e.g. language, loudness, etc.), while the *format* part describes the technical detail of the underlying audio to drive either decoding and/or rendering properly – including the rendering of Object-based Audio as well as signaling of compressed audio formats in addition to LPCM.

ITU-R BS.2088 [73] incorporates the ADM into the Broadcast Wave (BWF) and RF64 File formats (BW64) as well as well as incorporating metadata within the legacy BWF format as defined in Recommendation ITU-R BR.1352 [74]. ITU-R BS.2088 allows the ubiquitously supported audio file format, B-WAV, to carry numerous audio program representations including Object-based immersive along with audio programming containing elements that are intended to be used for personalization.

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

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

The reliable carriage of audio metadata across real-time interfaces and workflows within HD-SDI has been a long-standing challenge for the industry over the years. Moreover, the existing method(s) could only describe a limited number of Channel-based Audio programs along with limitations in terms of extensibility to support future needs. In the production system, there is a framework and accompanying bitstream format to be carried across any AES3 channel pair within the HD-SDI format. One embodiment of this framework/bitstream is a SMPTE 337 [36] formatted derivative of the Extensible Format for the Delivery of Metadata (EMDF) originally defined in ETSI TS 102 366 Annex H[28] [64]. (See Figure 60)

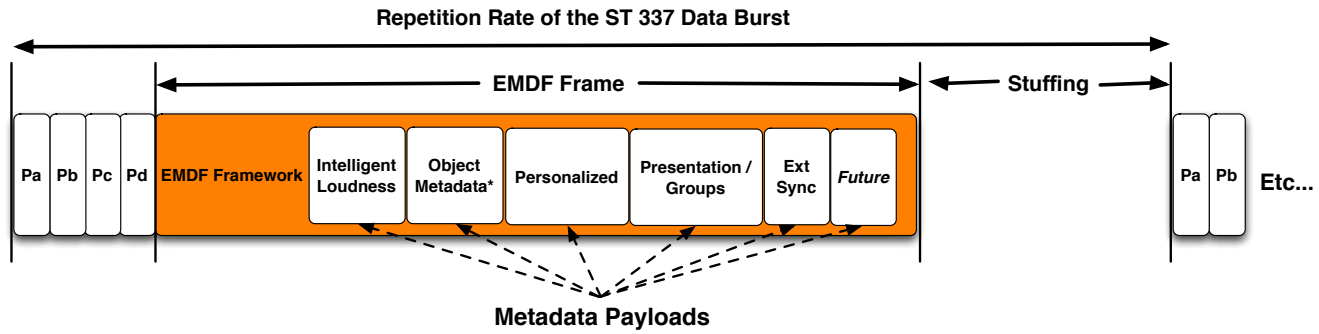


Figure 60 Serialized EMDF Frame formatted as per SMPTE ST 337 [36]

The EMDF specifies the carriage of metadata in a serialized (and efficient) form made up of 'payloads' each with a unique ID. Payload IDs can signal the carriage of several types of metadata (and associated DRM system-specific protection information) necessary for next-generation audio including immersive (object), personalized, intelligent loudness (i.e., as per ETSI TS 102 366 [64] Annex H payload_id 0x1), second-stream synchronization, and so on. Tools to translate to/from the metadata format defined in ITU-R BWF/ADM referenced earlier

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

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

are necessary, including conversion(s) to support frame-based splicing and cloud-based distributed processing required by interchange, distribution and emission systems. The open standardization of the EMDF framework/bitstream and associated payloads allow efficient real-time interchange of immersive (object), personalized, intelligent loudness, second-stream sync metadata, etc. within the SMPTE 337 family of standards for use in today's HD-SDI environments, while also ensuring the design supports efficient transport of metadata payloads for IP-based environments going forward. Operational note: The SMPTE 337 EMDF bitstream is a critical component to enable automated (and seamless) switching of a broadcast emission encoder to accommodate day-to-day operations where legacy programming is interleaved with next-generation programming utilizing program-specific embedded audio channel (or audio object) layouts.

14.4 DTS-UHD Audio

14.4.1 Introduction

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

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

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

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

DTS-UHD has been standardized with the European Telecommunications Standards Institute (ETSI) in TS 103 491 [91], and is included in the DVB Specification TS 101 154 [92] as well as also supported by the Society of Cable Television Engineers in SCTE 242-4 [93] and 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.

14.4.2 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 environment.

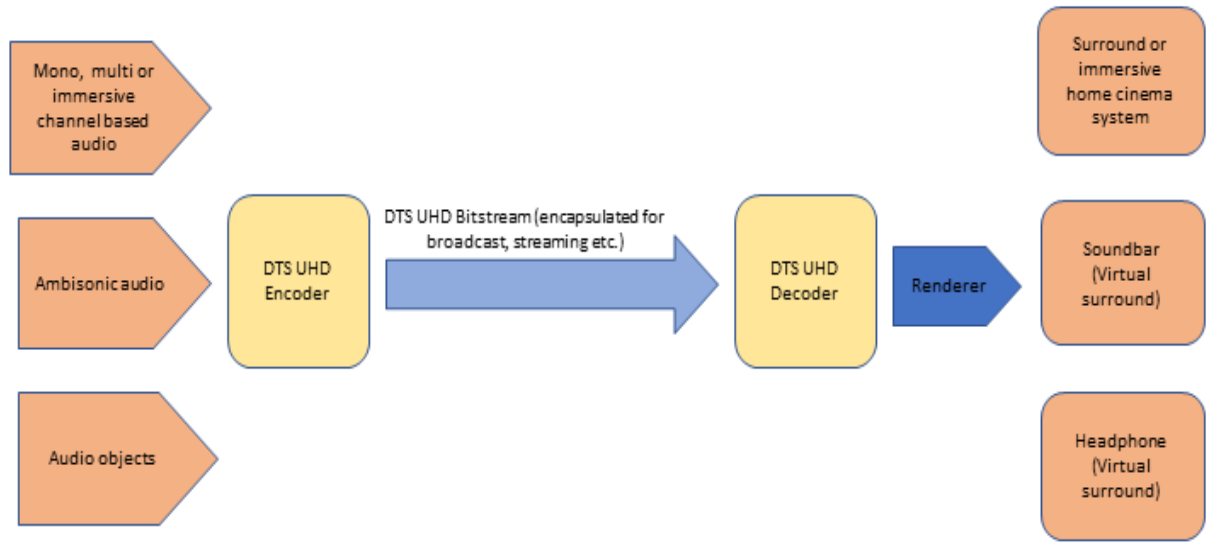


Figure 61 DTS-UHD System Overview

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

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

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

DTS-UHD allows hybrid delivery of different components of the audio content. This would allow the main audio and video service to be delivered via IP Multicast, with additional audio streams, such as language services or audio description delivered via an additional distribution method such as DASH. A DTS-UHD supporting application can receive, decode, sync and render the content for the consumer.

14.4.3 DTS-UHD Bitstream

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

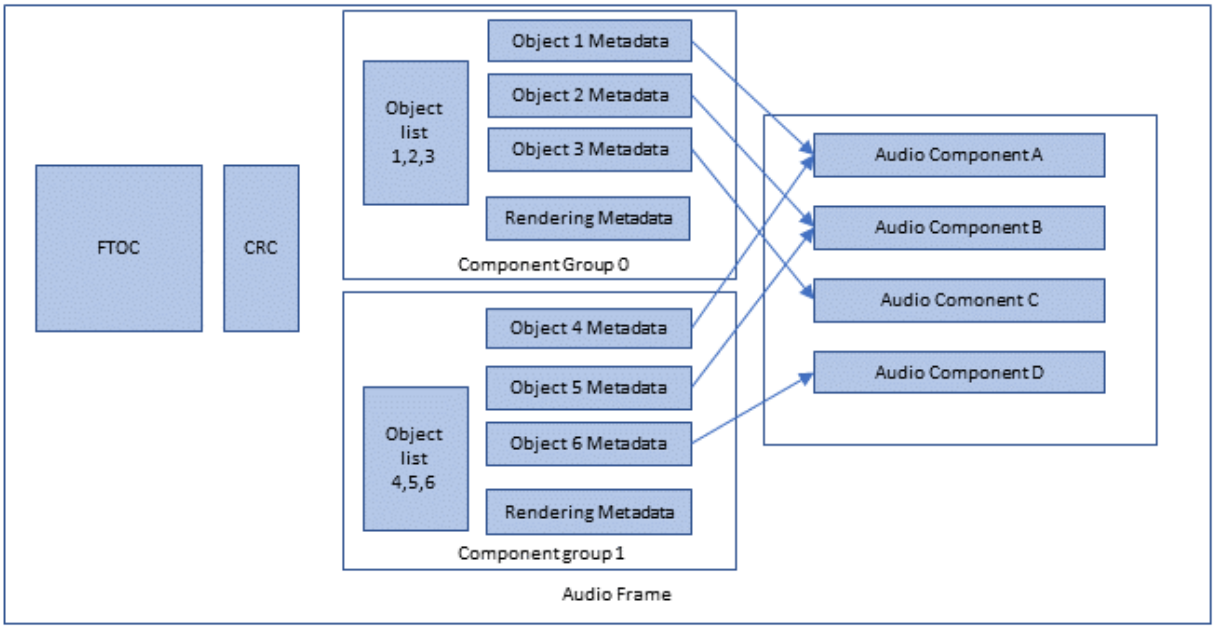


Figure 62 DTS-UHD Audio Frame Structure Example

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

14.4.3.1 Sync and Non-Sync Frames

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

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

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

14.4.4 Metadata

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

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

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

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

- Metadata describing different audio presentations may be present.
- The list of audio presentations/audio objects to be decoded within this stream is passed through a decoder instance API by one of three methods listed below. More detail of the object selection is shown in 14.4.6.4.
    - Play default presentation. In this case, no parameters are presented to the decoder and the audio presentation with the lowest presentation index where bEnblDefaultAuPres is TRUE will be selected, and the default objects within that presentation will be played. This case is indicated by m_ucAudPresInterfaceType = API_PRES_SELECT_DEFAULT_AP.
    - Play by presentation index. In this the desired audio presentation is indicated by a single parameter and the default objects within that presentation will be played. This API is aware only of the selectable audio presentations and non-selectable audio presentations are not counted in ucDesiredAuPresIndex. This case is indicated by m_ucAudPresInterfaceType = API_PRES_SELECT_SPECIFIC_AP.
    - Play an explicit list of objects. In this case, an ordered list of object IDs are presented to the decoder and only the audio presentations containing objects from this list are unpacked and played. If some of the listed object IDs are that of an object group, then that group's object activity mask is respected and the corresponding referenced objects are played. This case is indicated by *m_ucAudPresInterfaceType* = API_PRES_SELECT_OBJECT_ID_LIST.

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

14.4.4.1 Loudness

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

14.4.4.2 Dynamic Range Control and Personalization

Multiple selectable and custom dynamic range compression curves can be associated with an Audio Program to facilitate adaptation to various listening environments. The presence of a selectable DRC curve is indicated by the bitstream metadata parameter **m_bCustomDRCCurveMDPresent** as defined in ETSI TS 103 491 v1.2.1(2019-05) [91]. Different curves can be used to accommodate various playback environments. These curves are based on the DRC compression types and parameters based on the general symmetry between the amount of boost against attenuation. Specific slow/fast attack and release times are associated with each profile.

Table 24 Common DRC curves

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

For the boost vs attenuation parameter:

- A has less aggressive attenuation to loud content.
- B has less aggressive boost to quiet content.
- C has equal amount of attenuation and boost

Other legacy DRC curves are also supported within the system for film, music and speech. Additionally, a fully customized cureve can be included in the metadata, as described in ETSI TS 103 491 [91]

14.4.4.3 Metadata Chunk CRC Word

To ensure error detection, if the CRC flag (transmitted within FTOC: Metadata and Audio Chunk Navigation Parameters) corresponding to particular metadata chunk is TRUE the metadata chunk CRC (16 bit) word is transmitted in order to allow verification of the metadata chunk data. This CRC value is calculated over metadata fields, starting from and including the MD Chunk ID and up to and including the byte alignment field prior to the CRC word.
The decoder will:

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

14.4.5 Audio Chunks

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

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

For each audio chunk:

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

14.4.6 Organization of Streams

14.4.6.1 Objects, Object Groups, Presentations

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

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

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

14.4.6.2 Properties of Objects

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

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

- Describe the object's loudness and dynamics properties.

14.4.6.3 Properties of Object Groups

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

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

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

14.4.6.4 Audio Presentations and Rendering

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

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

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

Figure 63 and Figure 64 show two playback examples; the first one using Default Playback and the second one using specific object and object group selection. In both examples the darker blocks indicate the active elements.

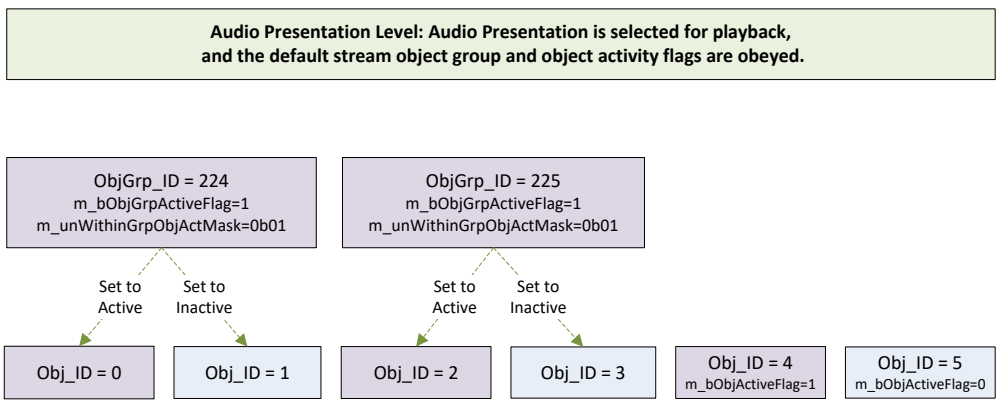


Figure 63 Default Playback

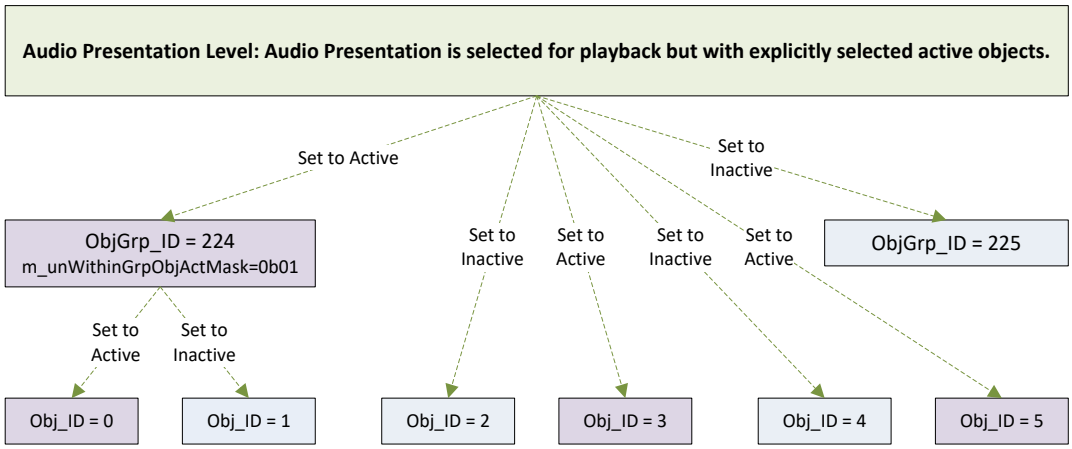


Figure 64 Specific Object and Group Selection

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

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

Figure 65 is an example of playback using the default audio presentation (*m_bEnblDefaultAuPres*=1), i.e. the lowest indexed selectable audio presentation (AP1). The default object group and object activity flags within AP1 are being obeyed. In addition external object groups are activated.

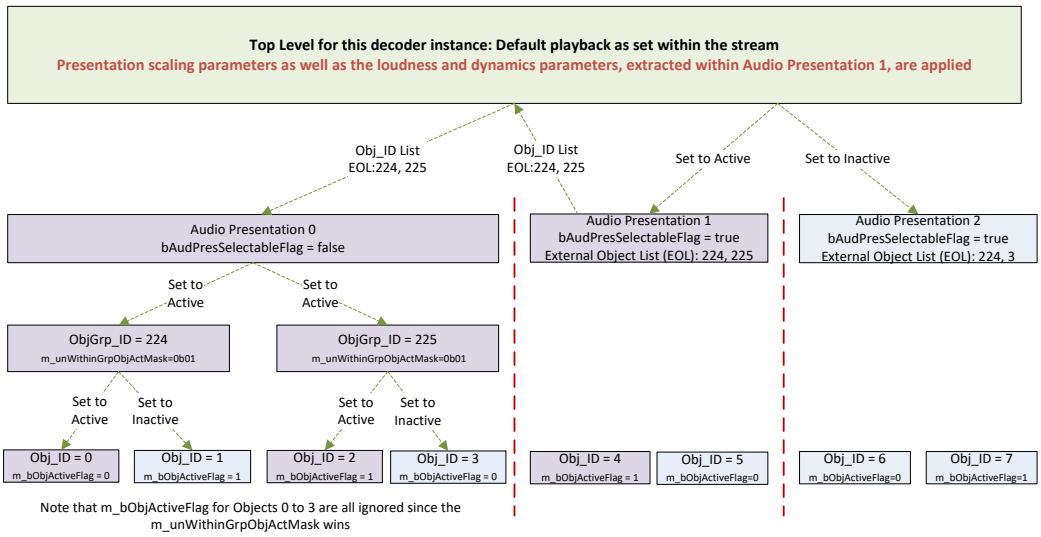


Figure 65 Playback using Default settings

The diagram, Figure 66, shows default object group and object activity flags within AP2 are being obeyed. In addition, external object groups are activated.

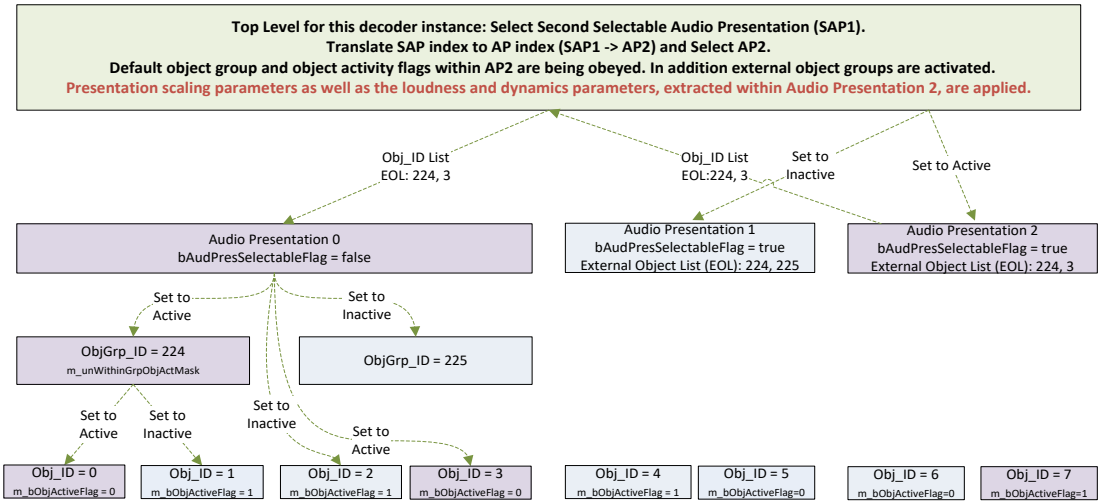


Figure 66 Example of Selecting Playback of Audio Presentation 2

The diagram in Figure 67 shows no default presentations being selected; rather an explicit playlist is selected which can override all defaults.

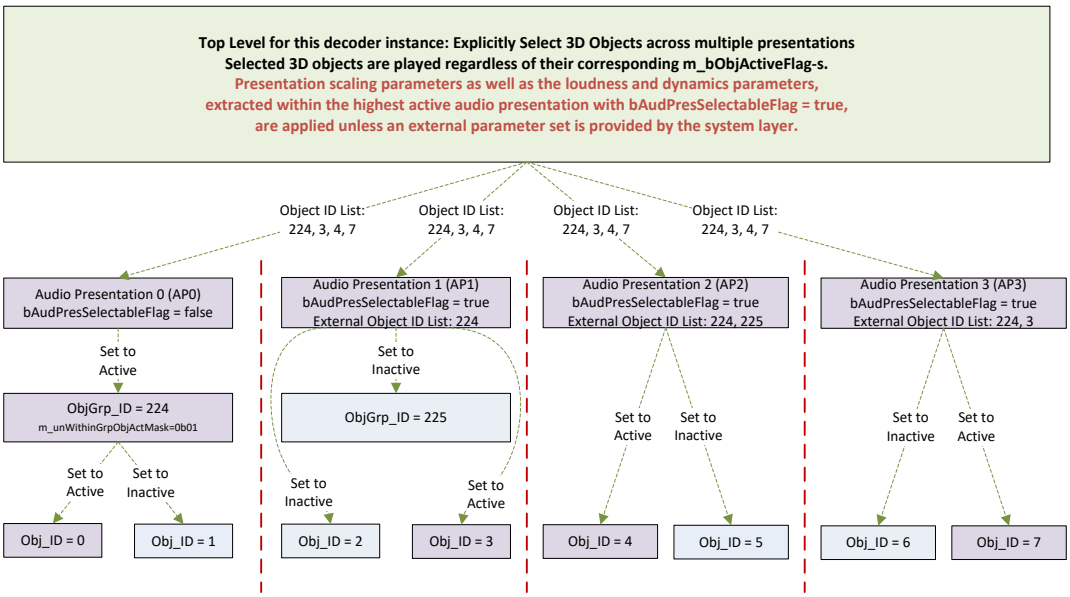


Figure 67 Example of Selecting Desired Objects to Play Within a Single Stream

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

14.4.7 Multi-Stream Playback

All of the above examples of playback have used a single DTS-UHD audio stream and as has been stated each audio stream requires a DTS-UHD decoding instance. However, this does not

preclude the use of additional streams, as in the hybrid example in the previous section. The presentation may be made up of multiple streams, with a main stream and additional auxiliary streams. There are two options available in this instance to decode the streams in this case:

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

In the example of Figure 68 three elementary streams contribute to a particular preselection. Component #2 is from the highest ordered stream in a multi-stream preselection. The renderer will first look for metadata from Component #2 to perform the final scaling of the mix. If some metadata is missing, then the renderer looks at the metadata delivered with Component #1, and finally Component #0, in order, to fill in the missing metadata.

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

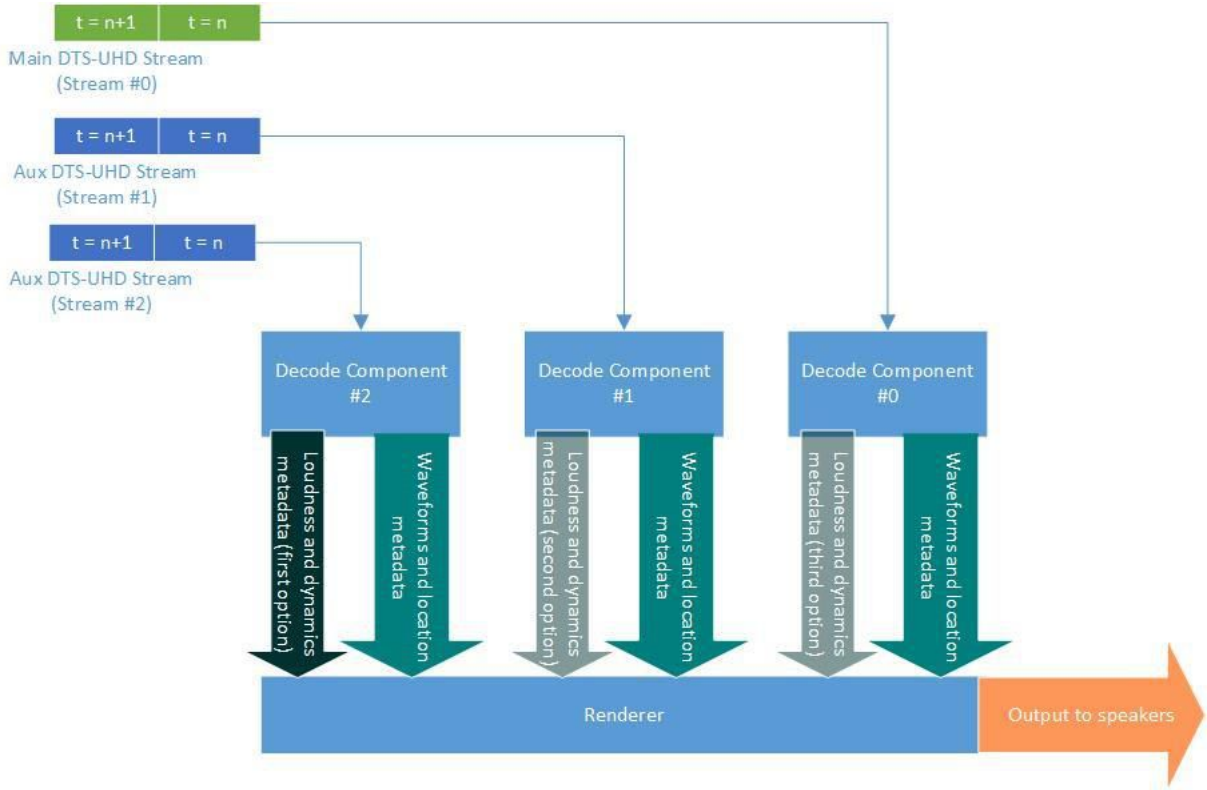


Figure 68 Example of multi-stream decoding

14.4.8 Rendering

For rendering DTS uses a point source renderer based on the ego-centric model. In section 14.1.5 the differences between allocentric and ego-centric rendering has been explained, however in practice the production sound mixer will place objects within a soundfield or space, and the renderer itself uses either the allocentric or ego-centric method for audio object placement at reproduction. The renderer uses metadata as previously described in the DTS-UHD bitstream in order to manage the placement and gain of the objects.

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

The location of a point is specified by polar coordinates - azimuth (θ) elevation (φ) and radius (r). Only the points on the unit sphere are needed, which means the radius shall be equal to 1. The listener is at the origin ($\theta = 0, \varphi = 0, r = 0$), facing the location ($\theta = 0, \varphi = 0, r = 1$) as shown below.

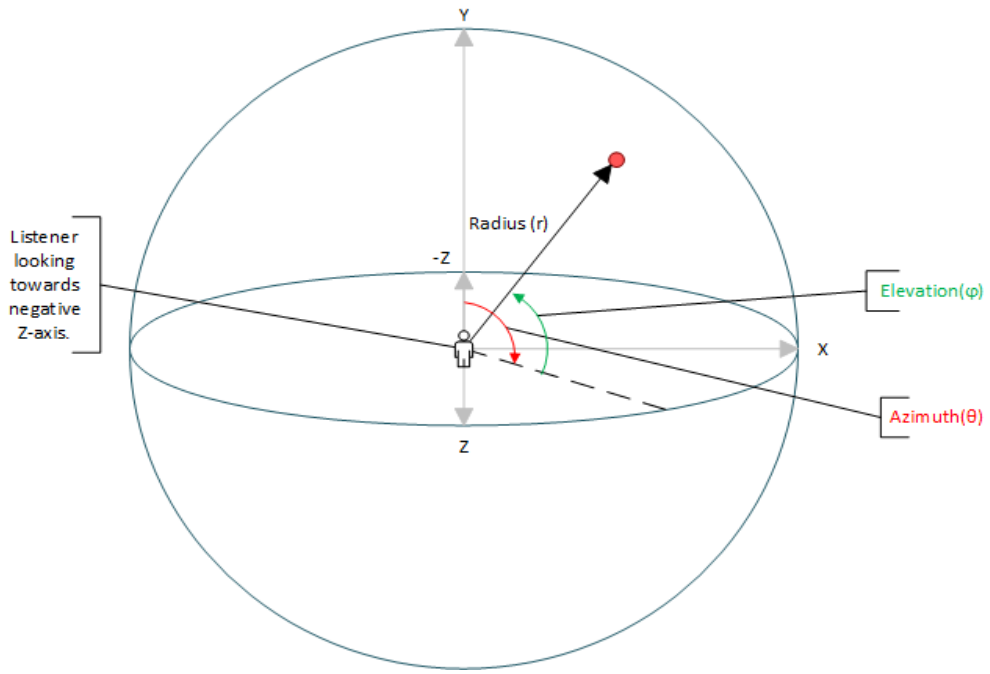


Figure 69: Point Source Object Renderer Coordinate System

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

In order to reproduce audio within the above sphere the speaker layout is used to create a mesh with a convex shape. As many speaker layouts have speakers at large angular spacing, this prevents a convex mesh being created. The DTS renderer uses virtual speakers to create a complete 3-dimensional convex array of speakers within a given sound space. In this case, vector based panning rendering is done over the full set of both physical and virtual speakers, with the fold-down of the virtual speakers to the physical speakers then carried out as a post vector based panning process.

Figure 70 shows the virtual speaker setup for a standard 7.1 configuration. The figure displays virtual speakers in both the upper and lower hemispheres with yellow vertices.

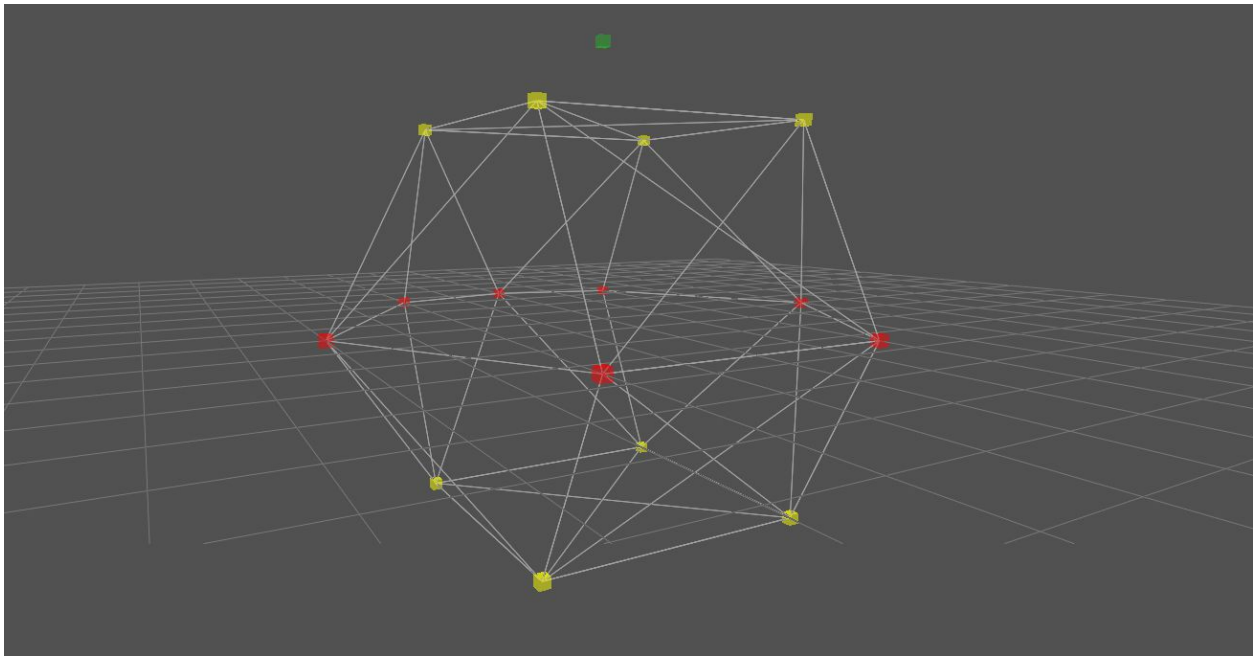


Figure 70: 7.x Output Configuration with Predefined Virtual Speakers

14.4.9 Personalization

As has been previously stated object metadata from the DTS-UHD bitstream may be overridden by the user during playback for the purposes of dialogue enhancement for example. Metadata in the bitstream may also be set to limit or disable a user interaction. The object interactivity manager enforces these rules and applies any user changes to the metadata before calling the renderer. Figure 71 shows that the object interactivity manager sits just before the renderer, where it handles the user input and the limit rules specified by the bitstream creator.

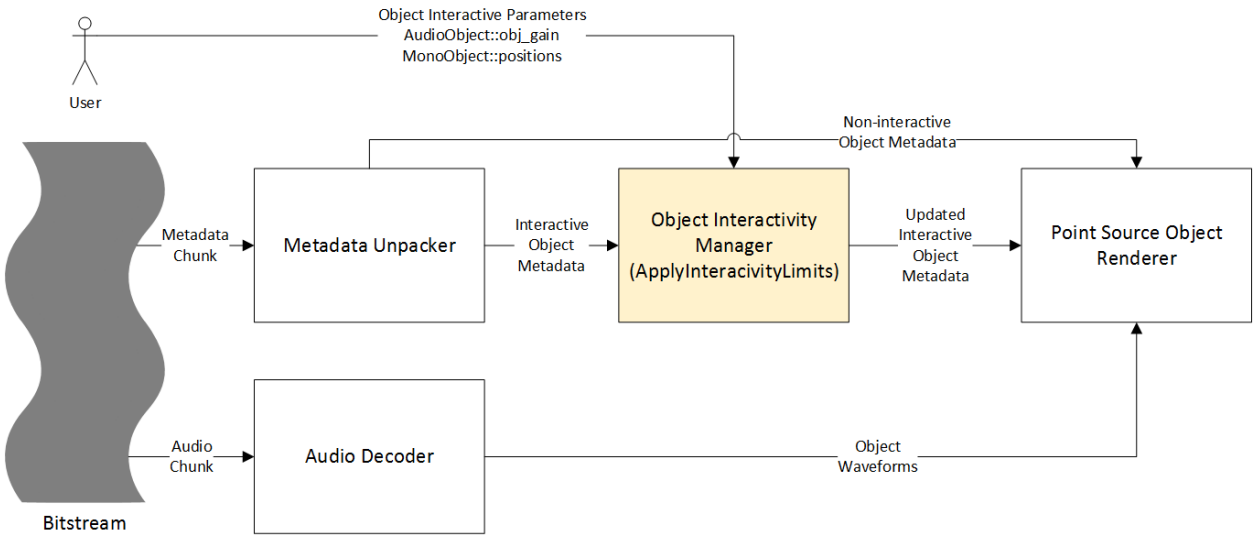


Figure 71: Object Interactivity Manager

15. Content Aware Encoding

15.1 Introduction

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

We will describe in this chapter how CAE works, how it can be applied to Ultra HD and the benefits of using CAE for the transmission of UHD program material. CAE is not a standard, but a technique applied on the encoder side that is expected to be decoded by an HEVC Main 10 decoder. Regarding adaptive streaming, the only existing specification is iOS 11[29]. Ultra HD Forum is seeking DASH IF Guidance to support VBR encoded content on the client.

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

15.1.1 Adaptive Bitrate Usage for UHD

For OTT, ABR is already the most common way to deliver content. CAE is applied on top of ABR in the encoding process. Currently only iOS11[30] has done that, but we expect a wider support such as from Android, DASH, and DVB in the future. For managed IP networks (Cable, Telcos), we also see ABR being used.

Cable operators can broadcast Live over either QAM or ABR or over IP (DOCSIS® 3.0 [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.

15.1.2 Per-title Encoding

Content aware encoding was introduced in production by Netflix®[31] in 2015 using “per-title encoding”[32]. 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.

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

[30] ibid

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

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

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

Some of the more recent implementations of CAE ladder generators reviewed[33] 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.

15.1.3 VBR Encoding

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

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

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

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

15.2 Content Aware Encoding Overview

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

[33] 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

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[34].

15.2.1 Principles

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

CAE techniques can be applied at different levels, described in Table 25.

Table 25 CAE granularity

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

*This might bring unacceptable additional delay (latency).

15.3 Content Aware Encoding applied to UHD

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

Table 26 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

[34] 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

differ in rates. The first ladder (shown in column 4) is a fixed CBR encoding design, assigning bitrates that are chosen independently, regardless of the type of content being encoded. The ladders shown in columns 5 and 6 are examples of CAE ladders generated for two different types of content. The CAE ladder in column 5 is produced for easier-to-encode content resulting in an average savings of more than 50%. The CAE ladder in column 6 is produced for more difficult content, resulting in an average savings of 40-50% vs. CBR encoding, depending on the content complexity.

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

Table 26 Examples of fixed and CAE encoding ladders for live sports

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

We draw in Figure 72 the bitrate vs. resolution of CAE vs. CBR at the same quality level. For simplicity for CAE, we use a more conservative example ladder, resulting in 40% savings.

We are plotting in Figure 72 CAE vs. CBR bitrates, assuming the same visual quality at a given resolution.

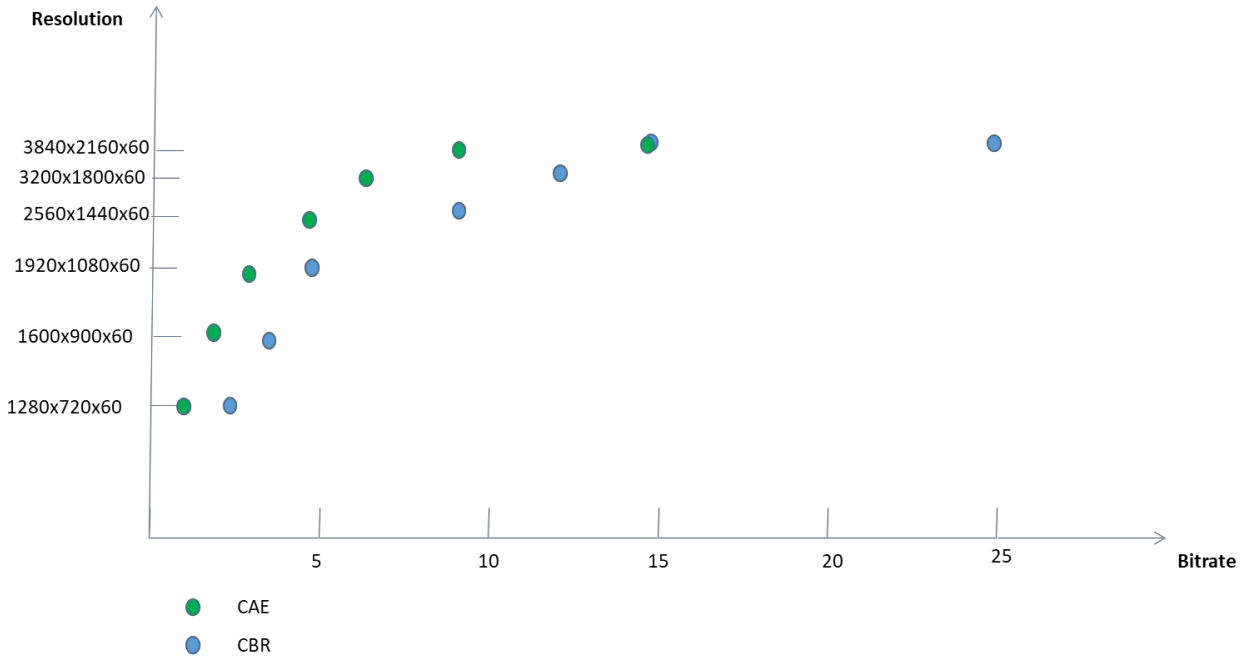


Figure 72 CAE encoding chart

15.4 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].

15.5 Application for Content Aware Encoding

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

15.5.1 Internet bandwidth

From Belson[35], Figure 73 shows the Internet speed distribution over various regions of the world.

[35] 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

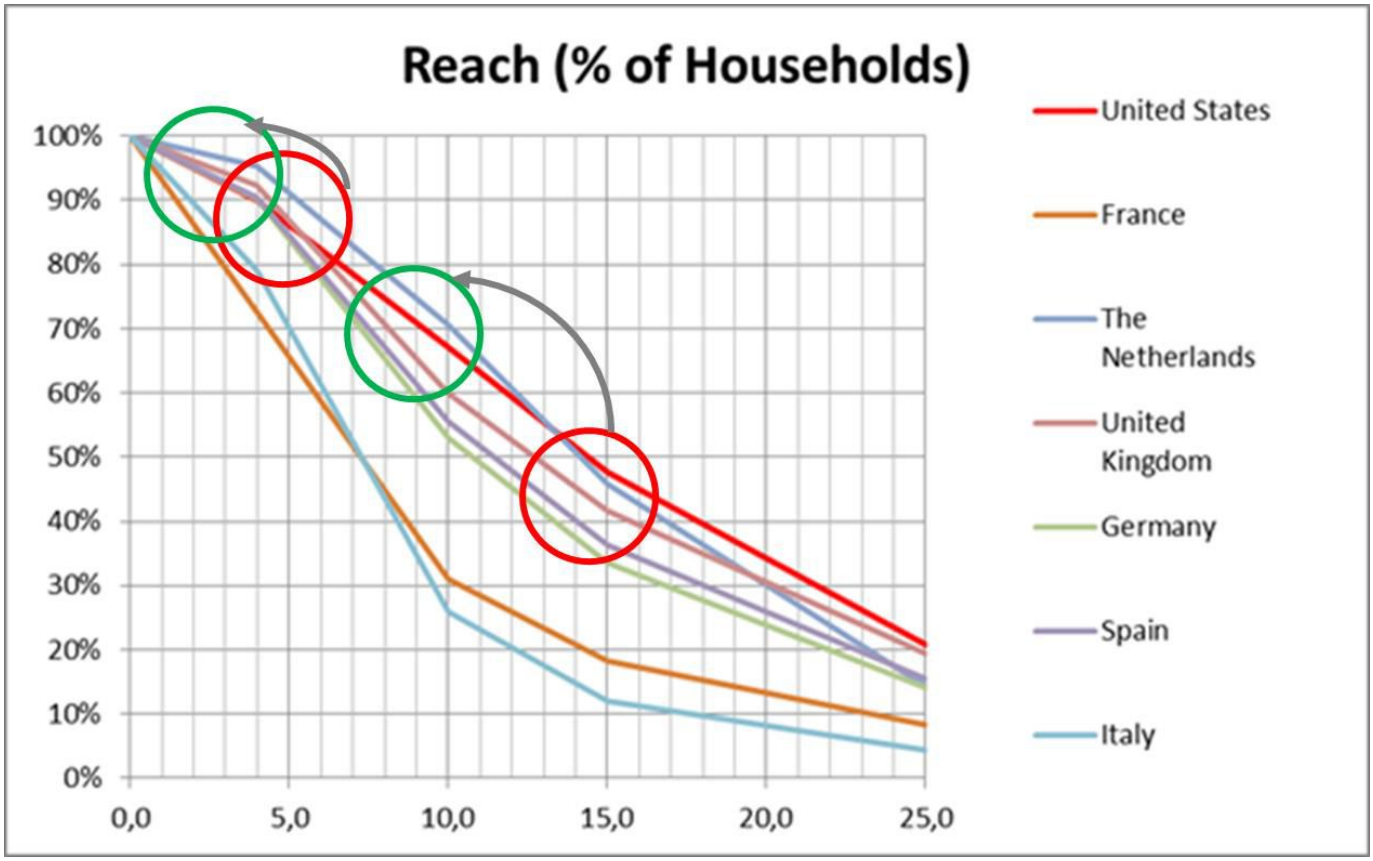


Figure 73 Internet speed distribution per countries (source Akamai)

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.

15.5.2 CAE Sweet Spot for UHD

Based on the previous section finding, the CAE sweet spot is when 2160p60 can be delivered at a lower bitrate than the CBR case. We describe in Figure 74 the CAE sweet spot.

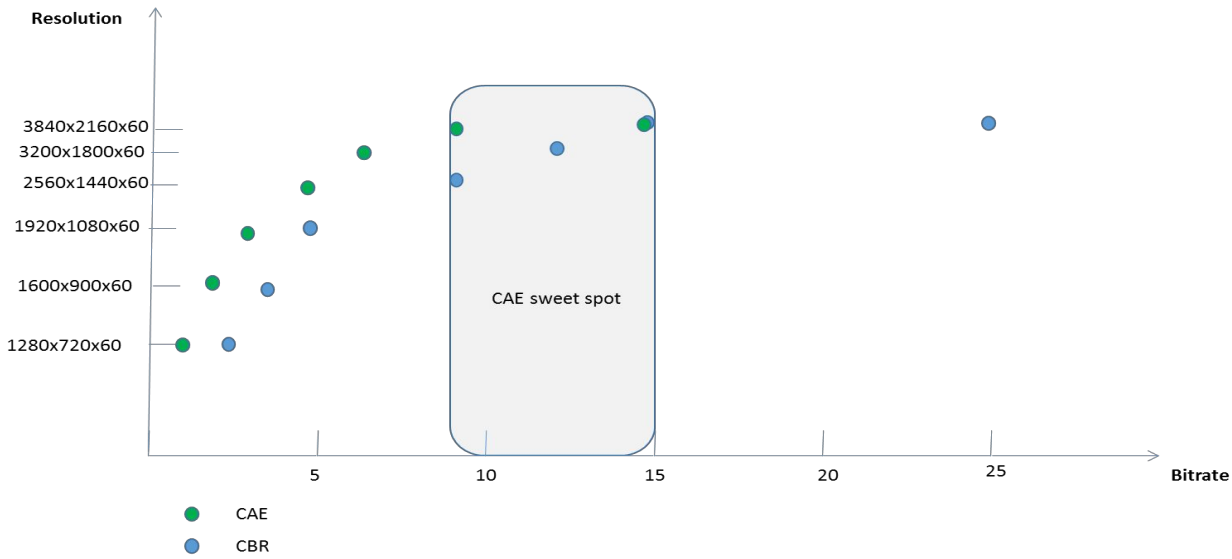


Figure 74 CAE sweet spot vs. CBR

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.

15.6 Content Aware Encoding Benefits

15.6.1 CDN cost

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

15.6.2 Quality of experience

Because the bandwidth required to carry CAE vs. CBR is reduced by 40-50%, the content will be transmitted in a smoother way across the delivery chain. Video services have reported up to a 50% reduction in re-buffering events and a 20% improvement in stream start times for VOD services. As the traffic is the same for Live or VOD, we expect the same network performances to apply for Live[36].

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 75.

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

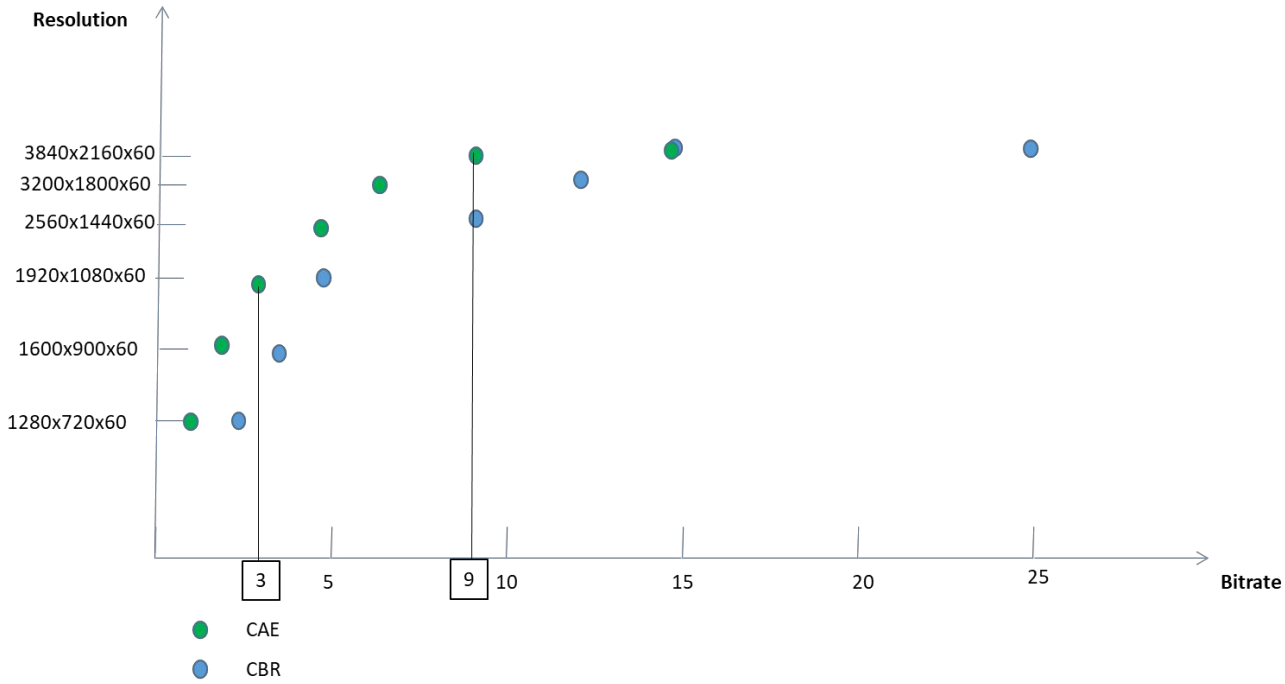


Figure 75 Junction bitrates chart

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

16. Annex A: Real World Foundation UHD Deployments

This annex describes several "real world" use cases of Foundation UHD deployments. The Ultra HD Forum gratefully acknowledges the many organizations that contributed their experiences to this document.

16.1 CBS and DirecTV Major Golf Tournament

Thanks to CBS and DirecTV for providing this information to the Ultra HD Forum about this workflow.

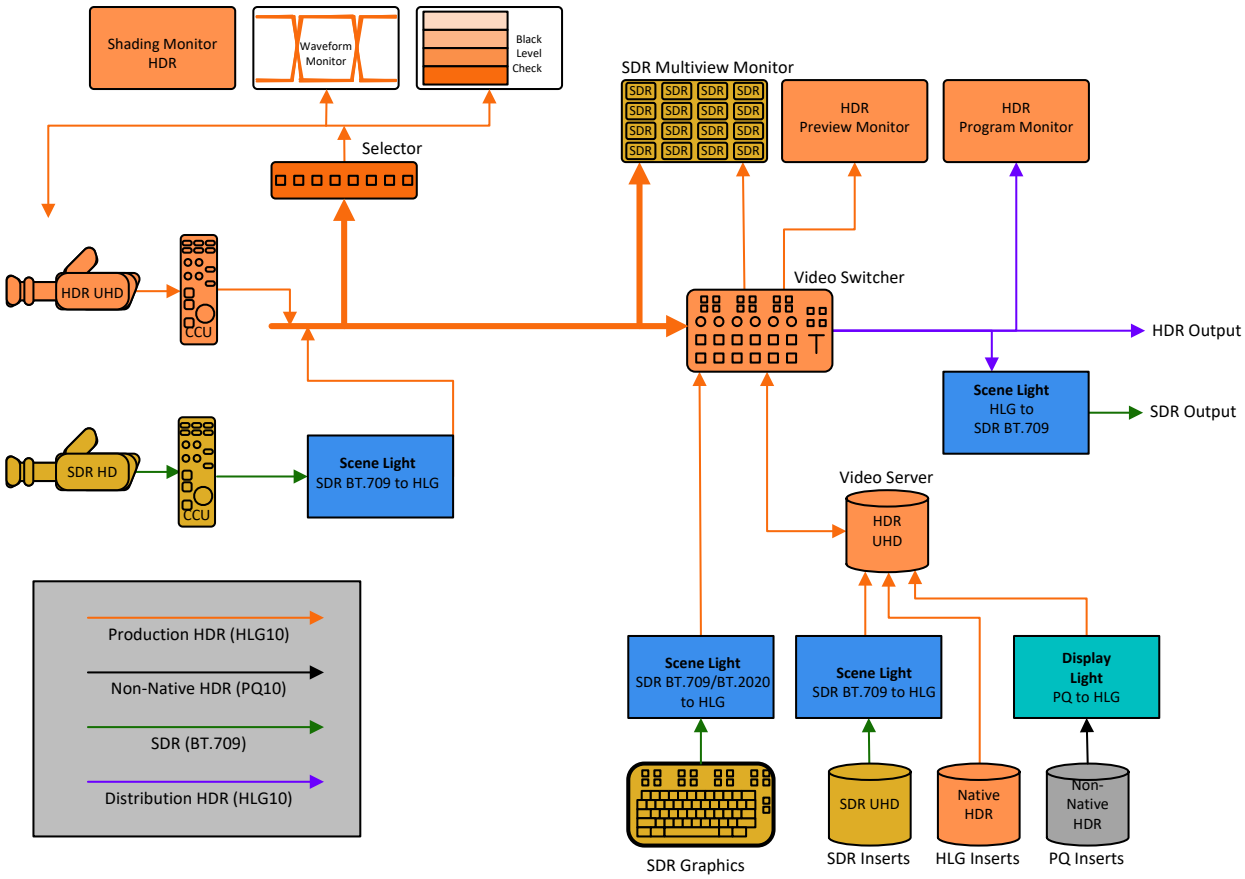


Figure 76 CBS and DirecTV Major Golf Tournament Workflow

For the production illustrated in this diagram, five holes of a golf course were captured in 4K HDR using Sony 4300 cameras operating in BT.2100 HLG mode. HDR content is mastered at 1,000 nits, with BT.2020 color.

The remaining elements of production were captured in SDR.

The video server shown in lower right of the diagram is used for replays and for inserting motion graphic elements. SDR interstitials and SDR graphics are inserted in a "Scene light" manner and switched through the Video Switcher. Graphics, legacy SDR inserts and

interstitials are native SDR and are converted to HDR with 100% SDR mapped to approximately 75% HLG. There is no highlight expansion of SDR sources.

An SDR version of the program is created as is shown in the "SDR Output" source on the right side of the diagram.

16.2 Amazon Major Parade

Thanks to Amazon for providing information to the Ultra HD Forum about this workflow.

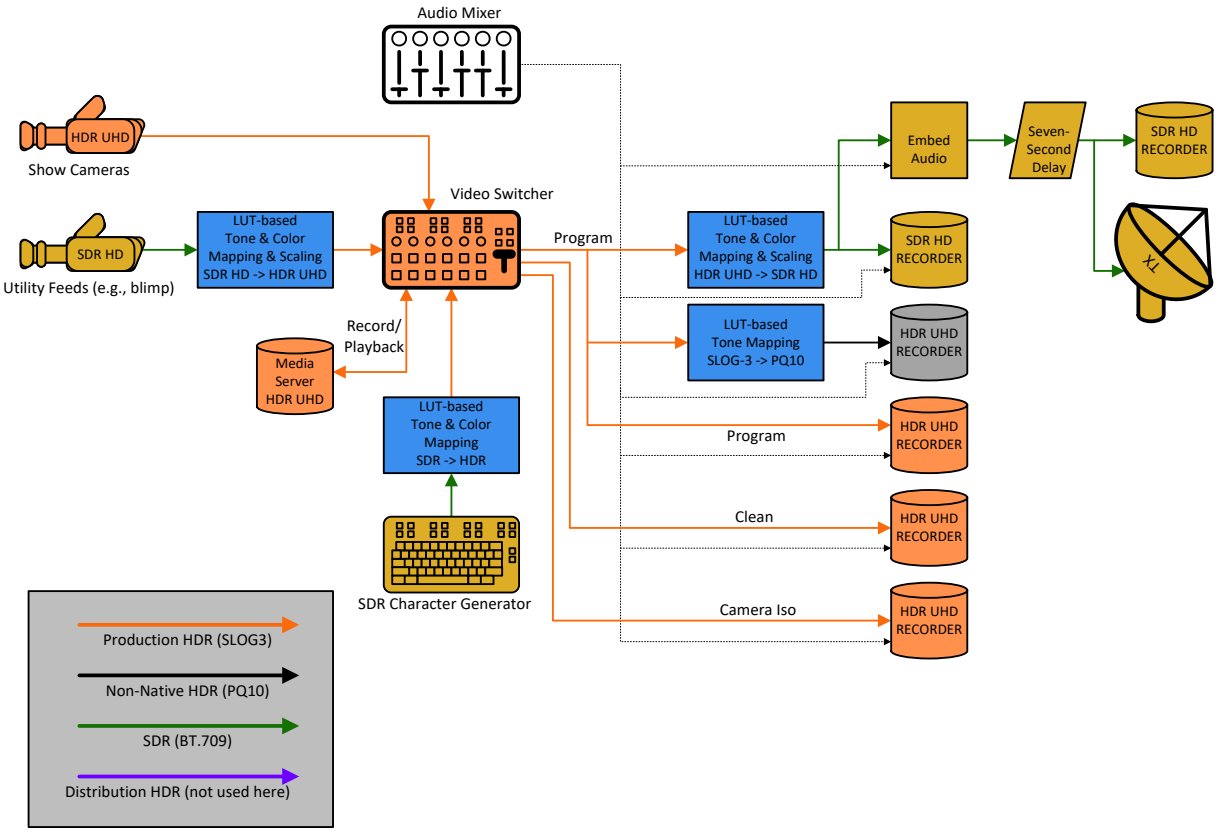


Figure 77 Amazon Major Parade Workflow

16.3 NBC Universal Olympics and 2018 World Cup

Thanks to NBC Universal (NBCU) for providing information to the Ultra HD Forum about their "production HDR" to "PQ Distribution HDR" for linear (live) and VOD workflows.

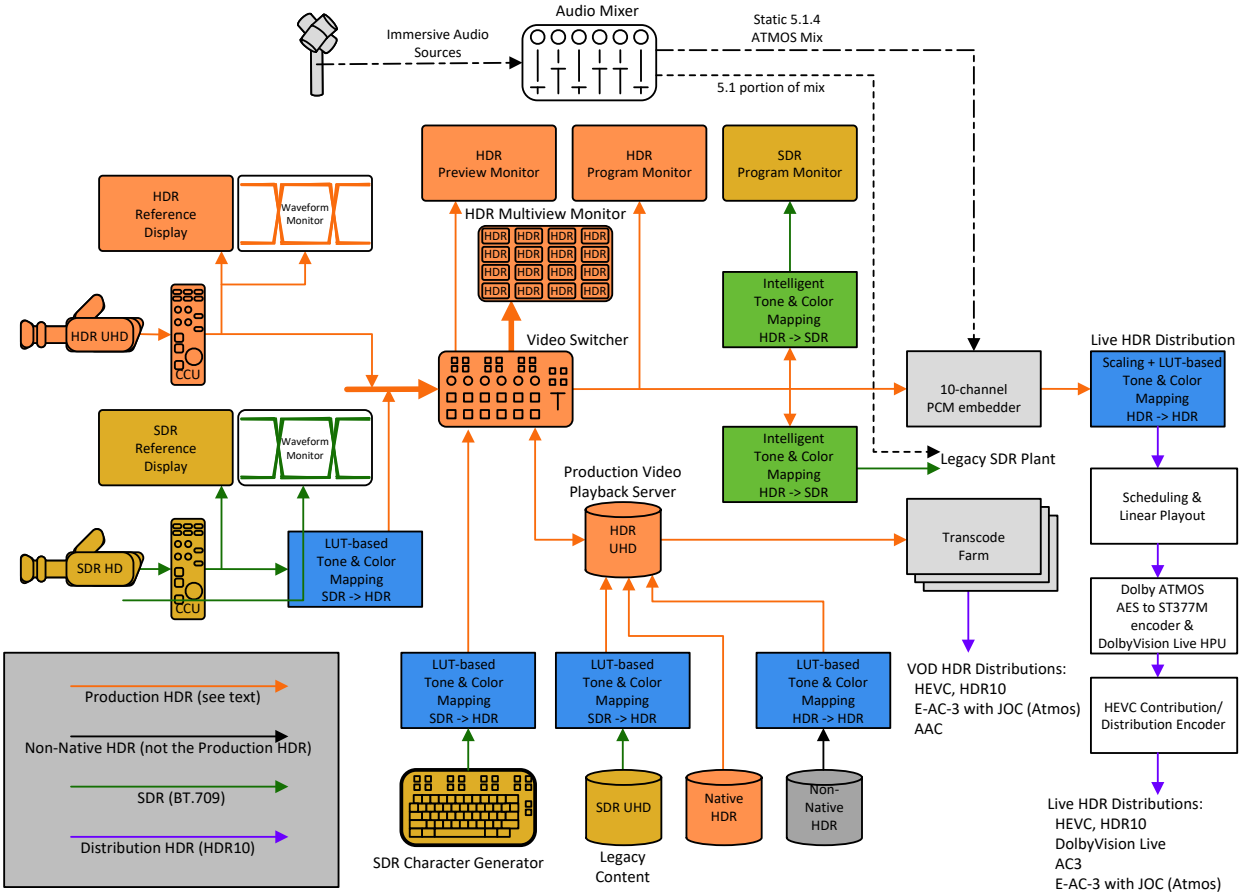


Figure 78 NBCU Olympics and 2018 World Cup UHD Workflow

This workflow includes three different production formats: S-Log3 Live, HLG, and PQ. NBCU chose PQ as a final distribution format, and converted all other formats, including SDR, to this common production format. The Pyeongchang Olympics were captured in HLG and the World Cup was captured in S-Log3 Live.

NBCU has strong preference to for PQ for the distribution format, citing that PQ ensures consistency in terms of final delivery of content quality. NBCU has found S-Log3 provides a good basis for conversion to PQ. BT.2020 color is used.

For shading, NBCU chose a starting point is of diffuse white at 200-300nits and then balances SDR and HDR for the best tonal range.

16.4 Sky/BBC Royal Wedding and Major Tennis Tournament

Thanks to Sky for providing information to the Ultra HD Forum about their distribution of the Royal Wedding in the UK in 4K resolution, a joint UHD production between Sky and the BBC. See https://news.sky.com/story/world-first-for-sky-news-royal-wedding-coverage-11355412 for additional details.

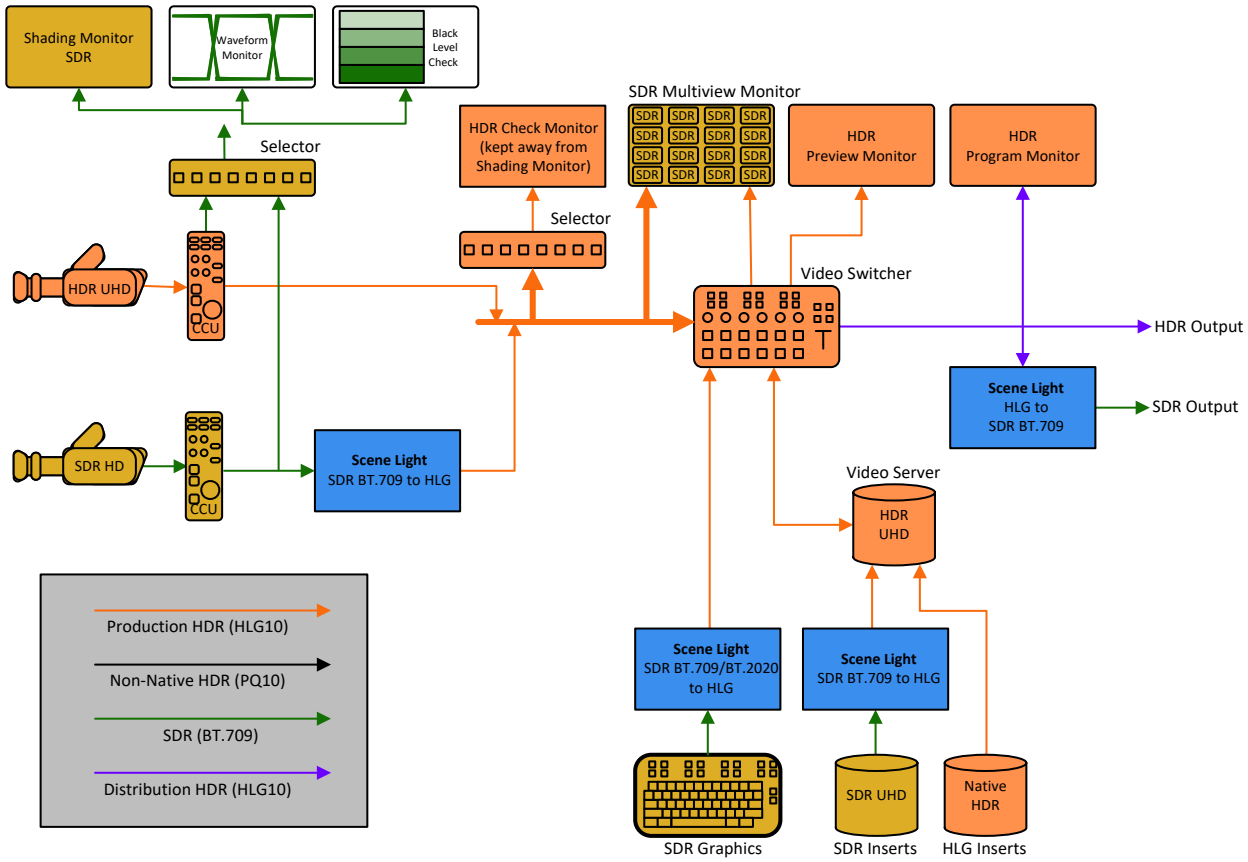


Figure 79 Sky/BBC Royal Wedding and Major Tennis Tournament Workflow

Although the BBC captured their portion of the content in wide color gamut BT.2100 HLG HDR, the main delivery for this event was HD in SDR/709 with Sky distributing UHD SDR/709 in the UK. The use of a single production workflow was necessary and SDR/709 the prime delivery but maintaining the capability to capture the wedding chapel in HDR was a key desire. Sky worked with the BBC to convert BT.2100 HLG to SDR/709 for the main Sky 4K and HD distribution feed. Numerous tests and trials were conducted in advance to ensure the conversion LUT was working satisfactorily across all colors and luminances. The initial "technically" correct LUTs followed the Macbeth chart closely, but when the color volumes were pushed by adding shiny saturated colored surfaces and glare (using a custom-built test chart), the chroma matching to SDR-matrixed cameras strayed. This had to be addressed due to the highly saturated and nature of the fabric of the guards' uniforms and other elements of the event that were expected to challenge the BT.709 capabilities. See also https://www.nepgroup.co.uk/post/the-royal-wedding-in-high-dynamic-range and https://www.bbc.co.uk/rd/blog/2018-05-ultra-high-definition-dynamic-range-royal-wedding-uhd-hdr for additional details.

Sky notes that while there were challenges, the event was a good experience and a big step toward having a single live production using both UHD HLG HDR camera outputs and SDR/709 cut seamlessly together delivering both HDR and SDR, UHD and HD.

A similar workflow was adopted by NEP for their Centre Court coverage of the Wimbledon 2018 tennis championships (https://www.nepgroup.co.uk/post/live-from-wimbledon-host-broadcaster-pacing-like-an-expectant-father-as-coverage-goes-in-house-for-the-first-time). This time, however, a mixture of HLG HDR and SDR/BT.709 specialist cameras were used. In order to ensure a good color match between HDR and SDR cameras, "scene-light"

conversions were utilized and the signal “clippers” on the SDR cameras relaxed to EBU R103 levels (-5%/+105%) to extend their effective color gamut.

16.5 BBC 2019 Football Association Challenge Cup

The BBC’s 2018 live UHD HDR productions used parallel UHD HDR and HD SDR workflows, to minimize the risk of the compromising the picture quality for HD SDR viewers. However, their joint coverage of the 2018 Royal Wedding with Sky (see Section 16.4) proved that format conversion technology had matured sufficiently to allow a high quality SDR signal to be derived from a single HLG HDR production workflow, greatly simplifying the production and reducing costs.

So, for the BBC’s coverage of the FA Cup quarter finals, semi-finals and final, the BBC adopted a single UHD HLG HDR workflow to feed both their domestic UHD HDR and HD SDR distribution, as well as to provide UHD and HD SDR signals to other rights holders. The simplified production architecture is illustrated in the figure below.

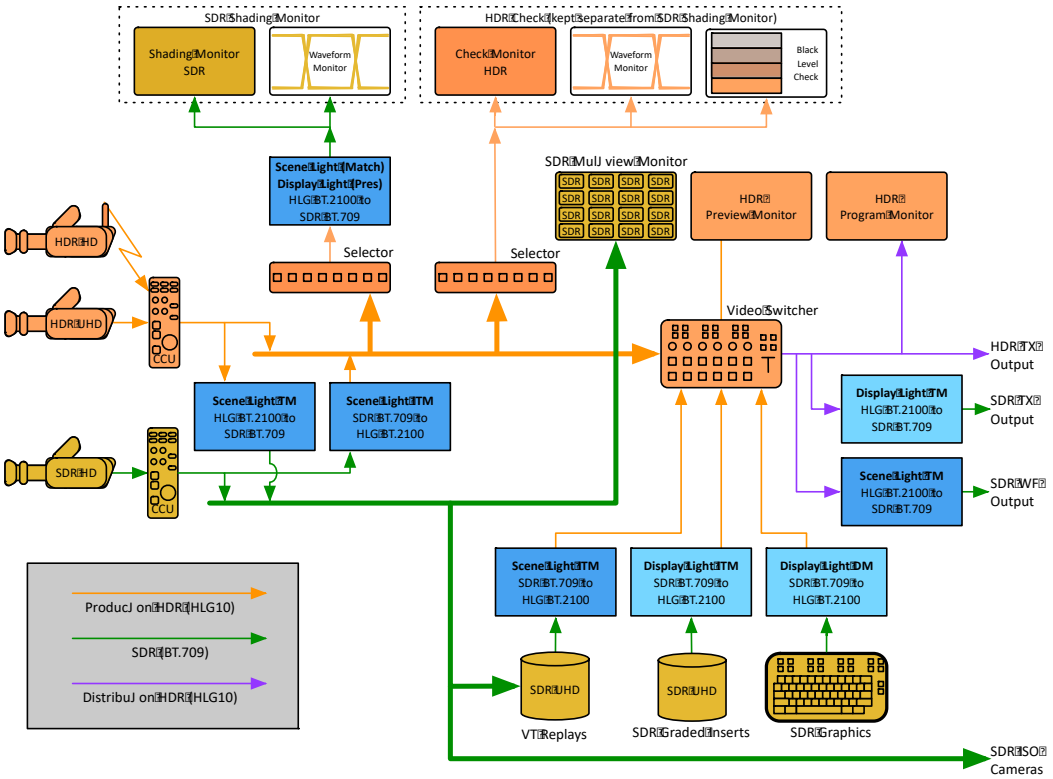


Figure 80 BBC 2019 Football Association Challenge Cup Workflow

Football is one of the most technically challenging sports to produce, making extensive use of specialist cameras and sophisticated graphics for ball-tracking and AR (augmented reality). The complexity of the BBC’s production increased through each round of the tournament. By the time of the final the BBC’s coverage was spread across two OB trucks controlling 41 cameras. It included six super-slomo cameras, four RF cameras, a Spidercam, a helicopter camera (helicam), polecams and robotic cameras (robocams). One truck provided

the match coverage, and the other "presentation" truck provided the BBC's domestic output, which included a local presentation studio.

The native displayed color of objects within a scene is different for each production format as they utilize different end-to-end opto-optical transfer functions (OOTFs) and color primaries. So, with such a complex mix of sources, it is important to use the correct type of HDR/SDR format conversion to achieve a good color match:

- "Scene-light" conversions based on the light falling on a camera sensor, should be used for matching camera sources. They are calculated using cameras OETFs and their inverse.
- "Display-light" conversions based on the light reproduced by a reference display, should be used for graphics and graded content. They are calculated using display EOTFs and their inverse.

More details of the recommended conversions for different signal sources can be found in the ITU report BT.2408-2 "Guidance for operational practices in HDR television production" [95].

Whilst the main cameras (including RF and helicam) were operating in HLG HDR, the super-slomo and specialist cameras could only operate in SDR BT.709. So "scene-light" conversions were used to convert the SDR camera outputs to HLG HDR, thereby providing a good color match with native HLG HDR cameras. A small "boost" was applied to the SDR camera highlights (known as inverse tone mapping (ITM) or "up-mapping") to better match the appearance of the native HDR cameras.

The "VT" area (record/edit/playback) providing action replays and pre-prepared program inserts, was also limited to 8-bit SDR, and every single camera was made available to other broadcasters in SDR (shown as SDR ISO feeds in Figure 80). So a "scene-light" conversion was used to convert the HLG HDR camera outputs to SDR BT.709, thereby providing a signal with a good color match to the native SDR cameras covering the event. Highlights from the HDR cameras were compressed using a "knee" type function (known as tone mapping (TM) or "down-mapping") as part of the conversion to SDR. This not only improved the SDR picture quality but reduced the "round-trip" losses when converting back to HDR for re-insertion into the program on the "VT" output.

In addition to the UHD HDR output, the BBC provided two different SDR program feeds:

- a "clean" (i.e. without graphics) "World-Feed" as a scene-light conversion from HLG to BT.709, to match the camera ISO feeds (SDR) and the SDR cameras of other broadcasters;
- a "dirty" (i.e. with graphics) BBC transmission output as a display-light conversion from HLG to BT.709, thereby ensuring identical colored graphics in the BBC's HDR and SDR programs.

To ensure the highest quality SDR output, the cameras were shaded using an SDR monitor, fed from the HLG signals using the same converters that provided the SDR program outputs. For operational reasons, cameras shaded in the "match" truck used the same scene-light conversion as the SDR "World Feed". Cameras shaded in the "presentation" truck used the same display-light conversion as the BBC's SDR transmission feed. In theory either could have been used. They differed slightly in terms of color saturation and shadow detail, but were both within the usual artistic tolerances for SDR football.

Program graphics were generated in SDR and converted to HLG using a display-light "direct mapping" conversion i.e. without boosting of highlights. The display-light conversion ensured that the graphics colors were maintained across both BBC outputs, and matched those of a conventional SDR production.

17. Annex B: $ICTCP$ Color Representation

The expanding range of display technologies, from various color primaries to increasing dynamic range, is creating a marketplace where color management is becoming increasingly important if artistic intent is to be maintained. In many applications, traditional color transformations may not be possible due to limitations in bandwidth, speed, or processing power. In these cases, image processing such as blending, resizing, and color volume transform must be performed on the incoming signal. With growing color volumes and the increasing need for color processing, distortions already known to be caused by non-uniformity of standard dynamic range (SDR) non-constant-luminance (NCL) $Y'C'_BC'_R$ (hue linearity and constant luminance) will become more prevalent and objectionable.

$ICTCP$ color signal format that is a more perceptually uniform color representation based on the human visual system. The improved decorrelation of saturation, hue, and intensity make $ICTCP$ ideal for the entire imaging chain from scene to screen. $ICTCP$ follows the same operations as NCL $Y'C'_BC'_R$, making it a possible drop-in replacement. These color processing improvements are achieved by utilizing aspects of the human visual system and by optimizing for lines of constant hue, uniformity of just-noticeable-difference (JND) ellipses, and constant luminance. The perceptually uniform design of $ICTCP$ allows for complex tasks such as color volume transform to be easily performed on HDR and WCG imagery with minimal error.

$ICTCP$ is included in BT.2100 [5] and is being deployed by OTT service providers as well as implemented by numerous consumer TV manufacturers.

18. Annex C: ACES Workflow for Color and Dynamic Range

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

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

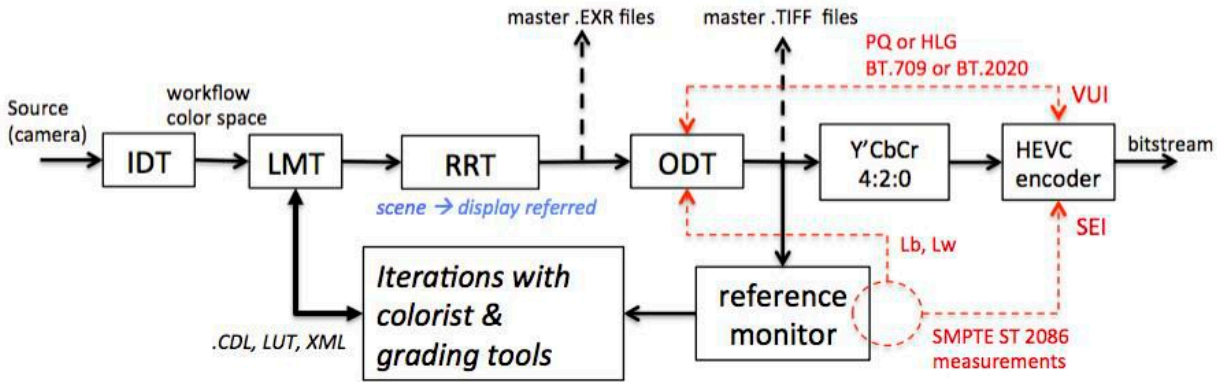


Figure 81 ACES Workflow Model

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

Table 27 ACES Workflow Model

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

In the diagram, the blue annotates the Reference Rendering Transform (RRT, e.g., “scene linear -> display referred”). The red detail indicates metadata input to the encoder for the system colorimetry and transfer function (VUI = video usability information). This could include the Master Display Color Volume (MCDV) metadata (e.g., for HDR10). VUI can optionally populate an Alternative transfer characteristics SEI message to support backwards compatibility. (See also Section 6.1.9.)

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

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

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

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

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

19. Annex D: ISO 23001-12, Sample Variants

For Forensic Watermarking as described in Section 7.2, transport of Variants can be done by different mechanisms. One alternative is transport at the container layer. Variants metadata can be transmitted in an alternate transmission channel next to the video content at the container layer. E.g., Variants metadata could be placed within a MPEG2-TS bitstream using a dedicated Packet Identifier (PID) and the Program Clock Reference (PCR) to synchronize the video and metadata channels. Alternately, Variants metadata could be incorporated as an extra track in an ISOBMFF [28] file. In that case, synchronization can be achieved by aligning samples across different tracks. When Variants metadata is handled as a component separate from the video, proper care shall be taken to guarantee its protection if needed with relevant content protection techniques.

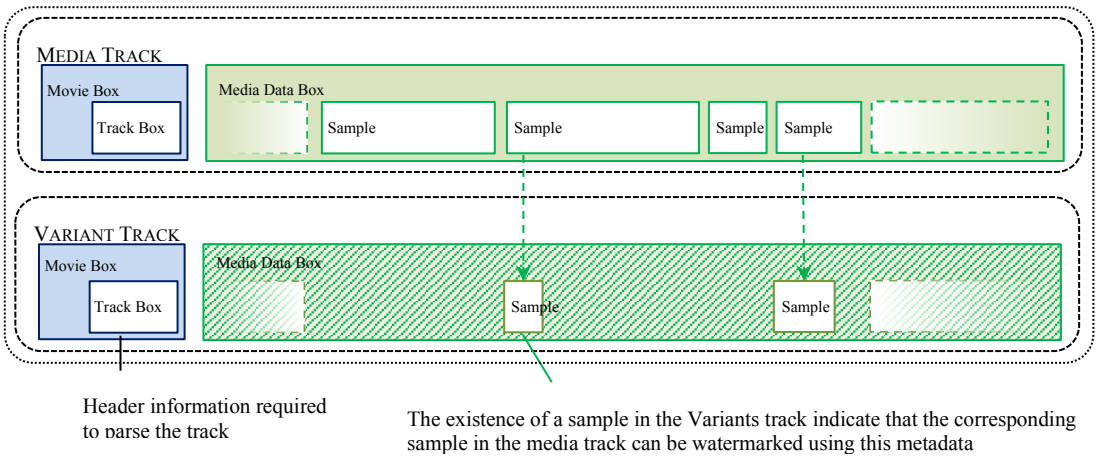


Figure 82 Transport at the Container Layer Using a Track in an ISOBMFF File

An example of how to transmit Variants metadata as an extra track in an ISOBMFF [28] file is described in [40][37]. This standard applies to file-based delivery, e.g., on physical media with embedding on the client side. The baseline principle is to define a dedicated track in an ISOBMFF file that describes how to construct watermarked video samples. For instance, a constructor for a sample indicates which portion of the video track shall be kept and which portions shall be replaced by a Variant available in the variant track. Access to the MPEG variant constructors is subject to cryptographic keys. Different users/devices will have a different set of keys and thereby would be able to only produce different watermarked video samples using different constructors. Moreover, the Variants are double encrypted to serve as an anti-bypass mechanism. A player that would not perform the watermark embedding operation would not be able to display a good quality video since some segments of the video would still be encrypted. The strong link between encryption and watermarking requires

[37] Noted that the terminology “variants” is slightly different in the MPEG standard and these guidelines. In the MPEG standard, a variant is a full MPEG sample composed of parts of the original bitstream and parts of the Variants, as defined in this document i.e. segments of bitstream that can be used interchangeably at a given location in the bitstream.

collaboration between CAS/DRM and watermarking systems, e.g., for key management and provisioning. The virtue of this design is that it enables a secure integration of the watermark embedding module on open devices outside of the secure video path or trusted execution environment.

20. Annex E: AVS2

The Digital Audio and Video Coding Standard Working Group (AVS Workgroup) of China delivered their latest generation Advanced Video Coding Standard (AVS2) to target UHD and HDR content for both broadcast and broadband communications and for storage. AVS2 was adopted by the State Administration of Radio, Film, and Television (SARFT) as the UHD video standard for industry [38] in May, 2016 and by the General Administration of Quality Supervision, Inspection and Quarantine (GAQSIQ) as the Chinese national standard[39] for UHD video, in December, 2016. Both were published in Chinese, and the English language version was standardized by the IEEE as 1857.4[40] in July 2018.

20.1 Why AVS2

AVS2 is the successor to the earlier video coding standard AVS+[41], which was successor-in-turn to AVS1[42,43]. 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[44], 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.

20.2 Deployment

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

---

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

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

[40] IEEE 1857.4-2018 "IEEE Approved Draft Standard for 2nd Generation IEEE 1857 Video Coding"

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

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

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

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

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

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.

20.3 Technology

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

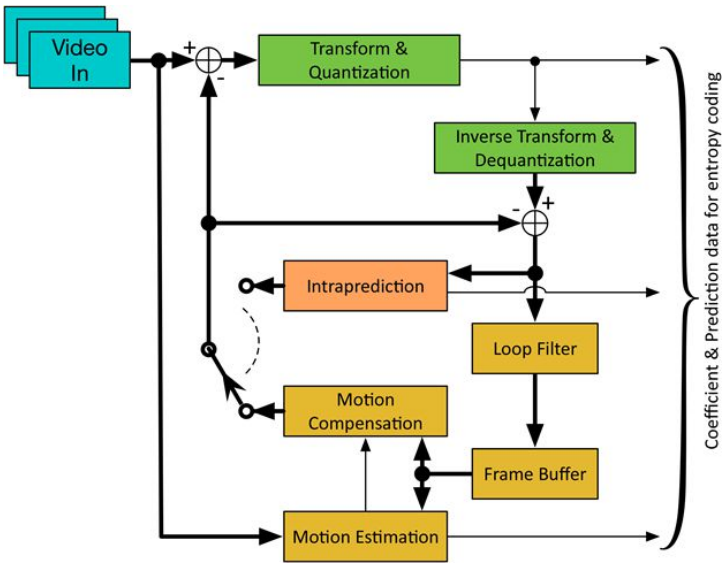


Figure 83 AVS2 coding framework

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

Where there are similarities between HEVC and AVS2, for example the overall processing flow, quad-tree partitioning, certain prediction modes, and motion vectors, transcoding from HEVC to AVS2 can be especially efficient[45].

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

8377 Fremont Blvd., Suite 117, Fremont, CA 94538 UNITED STATES