SCTE · ISBE

# JOURNAL OF DIGITAL VIDEO SUBCOMMITTEE

Volume 2 Number 1
July 2017

SCTE · ISBE

Society of Cable Telecommunications Engineers
International Society of Broadband Experts

# JOURNAL OF DIGITAL VIDEO

**VOLUME 2, NUMBER 1**
**July 2017**

Society of Cable Telecommunications Engineers, Inc.
International Society of Broadband Experts™
140 Philips Road, Exton, PA 19341-1318

# Table of Contents





**Editorial Correspondence:** If there are errors or omissions to the information provided in this journal, corrections may be sent to our editorial department. Address to: SCTE Journals, SCTE/ISBE, 140 Philips Road, Exton, PA 19341-1318 or email journals@scte.org.

**Submissions:** If you have ideas or topics for future journal articles, please let us know. Topics must be relevant to our membership and fall under the technologies covered by each respective journal. All submissions will be peer reviewed and published at the discretion of SCTE/ISBE. Electronic submissions are preferred, and should be submitted to SCTE Journals, SCTE/ISBE, 140 Philips Road, Exton, PA 19341-1318 or email journals@scte.org.

**Subscriptions:** Access to technical journals is a benefit of SCTE/ISBE Membership. Nonmembers can join at www.scte.org/join.

## From the Editors

Welcome to the third issue of the *Journal of Digital Video*, a publication of collected papers by the Society of Cable Telecommunications Engineers (SCTE) and its global arm, the International Society of Broadband Experts (ISBE). This issue, Volume 2, Number 1/July 2017, focuses on topics such as advanced audio standards, over-the-top (OTT) streaming and delivery practices.

One memorable *Monty Python* scene is a debate with a prospective corpse who replies: "I'm not dead yet!" These three papers take the debate from "video is dead" to the rapid evolution of video program services. To be sure, the days of a traditional linear program stream meeting the needs of the digital video consumer are gone; instead, "digital video" now means enhanced features across multiple platforms, devices, locations, and delivery modes, with the potential of global content flexibility.

The three papers in this issue each address separate, but important, aspects of taking a traditional video service to multiple platforms and devices.

First, expansion of what used to be an unsung utility channel for signaling and housekeeping is addressed in *Metadata for OTT*. Metadata has become an essential component of a program stream, now controlling rights and revenue, as well as the consumer experience. A blank screen resulting from a rights-related blackout is no longer acceptable to the viewer/consumer.

Second, preparation of content for multiple platforms and multiple device families can quickly lead to an exponential increase in cost with quantity of content formats. *Adoption of Unified Distribution Architectures* provides a look into the workflows and options to keep throughput and efficiencies as top-of-mind topics when undertaking new products and services.

Finally, remember the audio! New audio features, such as multi-language support and object-oriented sound (delivering sound above, around, and behind) are described in *MPEG-H*. MPEG-H is one of the next-generation audio systems undergoing standardization work in SCTE/ISBE's Digital Video Subcommittee.

We thank all who contributed to this issue of the *Journal of Digital Video*, including the authors, peer reviewers and SCTE•ISBE publications and marketing staff. We hope you enjoy this issue and that the papers spark the innovative ideas and essential knowledge required to advance video systems and services.

In closing, if you have editorial concerns or topics you would like us to consider for the fourth issue of *SCTE•ISBE Journal of Digital Video*, please refer to the "editorial correspondence" and "submissions" sections at the bottom of the table of contents for instructions.

*SCTE•ISBE Journal of Digital Video* Senior Editor,

Paul Hearty
Vice President, Technology Standards Office, Sony Electronics, Inc.

# Metadata for OTT Streaming of Broadcast Television

A Technical Paper prepared for SCTE/ISBE by

Alan Young, Chief Operating Officer, Crystal
4550 River Green Parkway, Suite 220
Duluth, GA 30096
alan.young@crystalcc.com
678-987-2925

# Table of Contents



## List of Figures



# 1. Introduction

This paper discusses how to extract and format the metadata linked to precise frames in video signals created during the origination of broadcast television so it can be used to drive the move of broadcast quality television onto the Internet.

Such metadata enables the frame-accurate (seamless) replacement of content delivered via the Internet e.g. as an Over-The-Top (OTT) or a TV Everywhere (TVE) service. The same metadata enables dynamic ad insertion at any point downstream of the origination and it also enables the automation of C3 and D4 Video On Demand (VOD) asset creation which could reduce the time it currently takes to produce those assets from hours to near instantaneous.

Evolving viewer habits continue to drive a shift to watching content that is delivered via the Internet. This evolution started with user-generated videos, e.g. YouTube, and quickly became a revolution as consumers were able to watch movies and episodic content online and on-demand, e.g. Netflix. The next phase is to move broadcast TV to the Internet and create packages of channels (just like a cable operator) but delivered via the Internet. This is, of course, already well underway (DirecTV and Dish have both launched such services – DirecTV Now and Sling TV) and more recently YouTube [1] announced a service to compete with the established multichannel video programming distributors and there are several others planning launches in the next year.

Last year at an event in Mexico City, Reed Hastings, Netflix's CEO, predicted that broadcast television will die by 2030, arguing, "It's kind of like the horse, you know, the horse was good until we had the car."[2] Mr. Hasting's projection that broadcast television will become obsolete now that the Internet can carry such high-quality video may or may not happen – time will tell. Though, at the current time it is hard to contemplate the Internet being able to handle events like the Superbowl without broadcast television – 100 million simultaneous streams of about 2 Mbps each equates to 200 Terabits per second of additional capacity for a few hours, once a year. Verizon, for example, currently advertises a network capacity of 28 Terabits per second [3] on its EdgeCast network.

It is easy, however, to contemplate a world where anything you can watch as a live broadcast via traditional means will also be available online and on any device at more or less the same time (broadcast delays are in the range of a few seconds whereas delivery over the Internet is typically delayed between 30 seconds to a minute). In fact, making traditional television broadcasts available for Internet consumption on any device (including a TV) has become a necessity for broadcasters as they seek to serve cord-cutters and cord-nevers in the face of increasing competition from online-only providers like Netflix. Unfortunately, delivering broadcast television via the Internet is more complex than inserting a multi-bit rate (MBR) transcoder and packager at the output of the broadcast origination system.

There are significant business challenges to overcome such as content rights not being uniformly available across legacy distribution and the Internet. You can't just send content to any device, anywhere, any time because the distributor may not have the rights to do so – and the penalties can be severe (i.e. financial penalties and/or loss of all rights). In almost all cases, some portion of the content of a broadcast linear channel needs to be replaced (because of rights issues) before it can be delivered over the Internet.

There is also a requirement to turn a profit. The vast majority of the revenues for broadcasters still come from the traditional linear delivery and this justifies the very significant operating costs of putting a linear channel on air. The cost of duplicating the existing workflow to create an OTT/TVE version of a linear channel cannot be supported today by the revenues generated purely through Internet delivery.

Fortunately, with the right metadata, it is possible to re-use most of the existing linear workflow to create OTT/TVE versions of linear feeds for far less than the cost of duplicating everything, by either selectively replacing content in the linear channel (for rights purposes, dynamic ad insertion or personalization) or totally automating the extraction of parts of the linear signal to automatically create VOD files.

## 2. Frame Accuracy

There are three key requirements for replacing part of a linear television signal with new content – whether it be because of rights management restrictions or for dynamic advertising inserted into the stream being viewed on an iPhone – in a manner that results in the end viewer being unable to perceive that a switch or substitution has occurred at all (seamless):

1. The technical and editorial quality of the substitute content must match the content it is replacing. This means the same compression parameters, the same audio formats, the same graphics, etc. If this is not the case, the user will see the difference. Fortunately, in practice this is straightforward to execute for the technical parameters, as they are known, which do not vary often. With respect to the editorial quality, again, this is largely under the control of the programmer and so is not very difficult to control, especially if the same people are making the decisions for the broadcast feed and Internet delivery – which is typically the case.
2. The duration of the replacement content (or advertisement) must be as close as possible to the same duration as the content it is replacing. Ideally the durations should be exactly the same – to the frame – but if the replacement content is a few frames short it can be filled out with black frames without being too noticeable. If the replacement content is long – even by a few frames – there will be a discontinuity (i.e. the first few frames of the next program will be missed or the last few frames of the replacement content will be lost) if the switch is made in the linear video signal (i.e. SDI or MPEG-2 TS). See Figure 1 below.

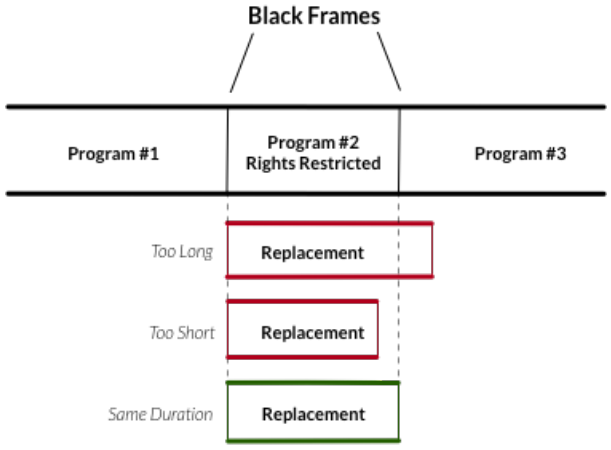



**Figure 1 – Length of Replacement Content is Critical for Program Continutity**

If the switch is made after the linear video has been fragmented for MBR delivery (e.g. HLS, MPEG-DASH or other HTTP streaming format), there is a little leeway; the streaming protocol will just follow the manifest (m3u8, MPD, etc.) and will delay the live stream until the replacement content has finished. Obviously if the difference is too big, there will still be a discontinuity somewhere that will have to be corrected – there are only 86,400 seconds in a day.

The ‘switch’ between the existing feed and the replacement content must occur during black and silence. This may not require absolute frame accuracy if there are black frames between the two pieces of content – which is typically the case – but the more black frames there are, the more noticeable it will be to the user. Regardless it is vital that the precise timing of the start of each segment of a linear program align with the end of the previous segment, because if the switch takes place too early it will cut off frames of video from the end of the program currently airing. If the switch occurs too late there will be a few frames of embargoed content airing that should not be. Both will result in a discontinuity which will also be objectionable to the viewer - see Figure 2 below.

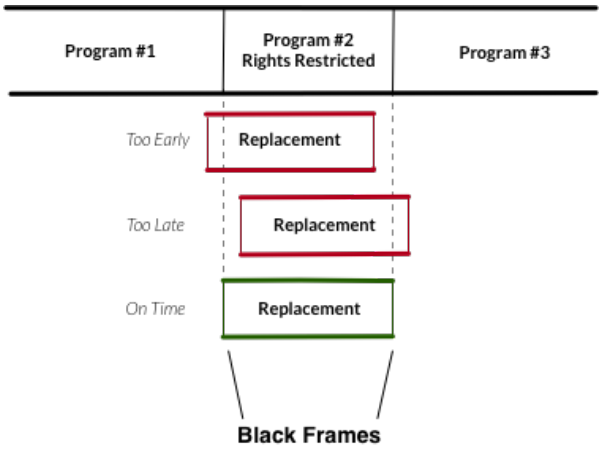



**Figure 2 – Instance the Switch Occurs is Critical for Program Continuity**

The requirement to be seamless is as critical for linear television broadcast via satellite, cable or over the air as it is when delivered over the Internet. So having the metadata regarding the frame boundaries of the content is essential. Without this metadata the only option is to completely re-originate the channel and this is highly duplicative (in most cases 90%+ of the daily schedule for the OTT/TVE channel would be the same as for the main broadcast) and expensive, and it doesn’t help at all with more complex rights management situations, i.e. certain NFL games cannot be delivered over wireless networks that don’t belong to Verizon.

The information regarding the precise times at which the last frame of a program and the first frame of the next piece of content (e.g. an advertisement) are broadcast is only available from the broadcast automation system for pre-programmed content or the Master Control Room switcher for live content.

## 3. SCTE 35

The SCTE 35 standard [4] has been used for many years to signal local commercial breaks so that cable operators can insert their own local advertising into national program feeds. The SCTE 35 standard has evolved over the years and now defines signaling for the start and end of all programming, promotional and advertising content as well as the ID of the content, using segmentation descriptors. This would be ideal for the applications described in the previous section but for the fact that the implementation of the SCTE 35 standard by broadcasters to date is not uniform, complete and/or frame-accurate in most cases despite the Recommended Practice defined in SCTE 67 [5].

In many cases, SCTE 35 signaling is implemented using General Purpose Interface (GPI) contact closures driven by secondary events in the automation system playlist. Different GPIs are assigned to different message types – start of program, end of program, start of ad break, etc. Apart from the fact that the more GPIs are used the more the broadcast center looks like a 1970's telephone exchange, the only SCTE 35 messages that can be inserted using this method are 'canned messages'. So even if the GPIs are frame-accurate, the content ID is usually missing and the number of different message types that can be employed is limited. The lack of a content ID in the in-band SCTE 35 message makes it harder to identify content that is being signaled which may be required to implement a SCTE 224 [6] policy, for example. When using GPI, the content ID must be determined using an out-of-band method such as matching events from the Electronic Programming Guide (EPG). This is a highly error prone mechanism because if a SCTE 35 message is missed or there are late-breaking changes to the schedule, the system gets out of sync quickly and will 'think' it is dealing with a different piece of content than it actually is. There is also the problem with cluttering up the broadcast origination playlist with secondary events, which in turn, in the extreme, can cause operational issues as human operators 'can't see the woods for the trees', especially during live events.

There is a lack of uniformity in the parts of the linear stream that are signaled and precisely what is signaled. This is a problem for both programmers and their distributors. For example, if programmer 1 is signaling the start and end of each ad break with a message using provider advertisement type (provider advertisement is a segmentation_type_id in SCTE 35) and programmer 2 signals the same ad break as a placement opportunity (another segmentation_type_id in SCTE 35) things get very complicated for the distributors quickly.

It is often also the case that the distributors themselves ask for different SCTE 35 message formats from the same programmer and some look to use proprietary APIs that have nothing to do with SCTE 35. SCTE 35 is a cable TV standard; OTT operators tend to be more comfortable with APIs. Verizon Digital Media Services, for example, has a published API for its Live Slicer [7] which can be used to signal timing and content ID, amongst other things, without SCTE 35 (or SCTE 104).

## 4. Extracting and Formatting the Necessary Metadata

There is a better method than the hardwired GPI approach for extracting the timing data associated with the individual programs and advertisements in a linear television signal: direct integration with the automation system. This is contemplated in SCTE 104 [8] - but the standard doesn't describe how to actually execute (and nor could it, given that there is no standard automation system). SCTE 104 messages are designed to be converted into SCTE 35 by a compression system, so the Standard itself doesn't address the programmers' and OTT operators' issues with SCTE 35 described in the previous section.

Fundamentally, all that is required from the automation system (or Master Control Room (MCR)) is a content ID and the time that event will go on air with enough accuracy to be able to identify the first frame of video. Although there are many automation systems, they all fundamentally work in the same way. A few automation system vendors even provide their own API which allows direct querying of the system to extract the necessary timing information. In all cases, it is possible to gather the necessary timing data without impacting the automation system at all – which is a critical requirement.

The content ID is usually a house ID relating to the asset to be played and is therefore local to the origination operation. Downstream, this will likely not be of much use to distributors who are looking for

program titles, descriptions, actors, directors, etc. Linkage to the title of the program or advertisement (as well as other metadata) needs to come from other sources. So, to get this additional metadata the playlist content IDs are used to query traffic system databases, asset management systems (MAM), etc. to extract the necessary information. The wide adoption of EIDR [9] and Ad-ID [10], for instance, makes these very useful metadata identifiers for distributors and it is a relatively straightforward matter to take a content ID and query a database to acquire the Ad-ID or EIDR code (or several others).

This, together with the timing associated with the asset, can be passed as a SCTE 104/35 message or through a proprietary API and, importantly, multiple SCTE 104/35 message formats and APIs can be derived from the same content ID and timing data. See Figure 3 below. This essentially means that Linear TV channel providers can easily satisfy practically any and all metadata formats required by distributors – both traditional and OTT – without having to change their existing automation system in any way or run twisted pair wiring for GPIs.

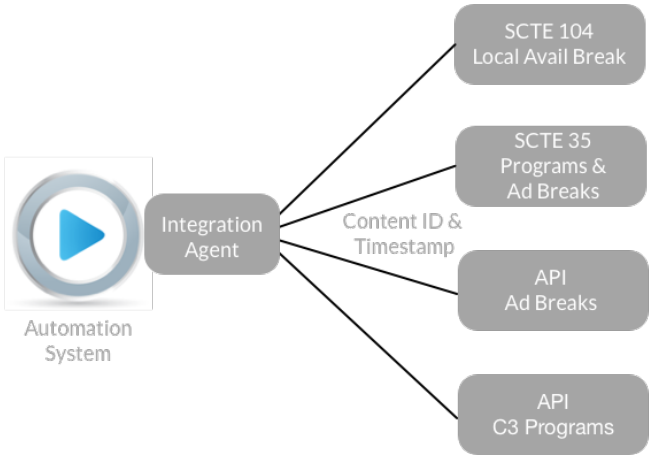



**Figure 3 – One Automation Interface Can Provide All Necessary Metadata Formats**

# 5. Applications and Conclusions

Once the rich metadata describing a linear television channel in terms of EIDR (with chapters) and Ad-ID with the start and end frame of each component have been extracted and formatted (i.e. as SCTE 35 messages or proprietary API calls), a number of applications can be automated easily. Some of these applications are critical in terms of resolving the business challenges associated with providing an OTT version of a linear channel.

## 5.1. Content Replacement

As discussed already, content distribution rights are not uniformly available across legacy distribution and the Internet. You can't just send content to any device anywhere any time because the distributor may not have the rights to do so. This leaves many broadcasters in the position of having to remove and/or replace programs on their existing channels for which they do not have the rights to deliver over the Internet.

Today's methods of executing this content replacement are undesirable because they either significantly increase cost or potentially damage the brand. Re-originating a linear channel is expensive, but if the offending content is just blacked out (to save cost), it turns viewers off. The slate may say "we'll be right back" but viewers won't wait. Consumers are a lot savvier today and have multiple alternatives for their

video consumption demands. It's still all about the eyeballs and engagement to ensure maximum ad value ratings.

With the linear television feed from an origination system appropriately marked with metadata describing each event's start and end times, it is possible to replace content seamlessly with something else. This replacement can take place in the broadcast center under the control of the broadcaster for critical replacements or it can be executed downstream. SCTE 224 is the standard which describes the Event Scheduling and Notification Interface. This essentially allows the rights associated with a piece of content to be sent out-of-band in the form of policies that can be executed when the in-band metadata described in this paper is detected. As an example, a policy could be written using SCTE 224 which, in English, indicates that a particular program (which could be identified by an EIDR code) is to be substituted by another piece of content when the user is on a particular network or in a particular location. As soon as that EIDR is detected in the SCTE 35 message, the exact frame at which to make the switch is known and so the content can be replaced seamlessly. Using this mechanism it is possible to generate thousands of different streams from the same original TV broadcast.

### 5.2. Dynamic Advertising Insertion

Because delivery of content over the Internet makes it possible (in fact, necessary) to send each user an individual stream, tailoring the advertising to each user greatly increases the relevance (and value). Given that the advertiser can have knowledge of what is currently being aired (through SCTE 35), where the user is, what device they are using, and likely also their browsing history, contextual advertising based on all of these factors can be delivered to greatly enhance the relevance and effective cost per thousand (CPM) of advertisements.[11] It is even possible to describe policies using SCTE 224 that place advertising for users based on the programming directly before the ad break. Although Dynamic Advertising Insertion (DAI) technology is more or less ubiquitous today, the timing is often not frame accurate resulting in ads cutting off the end of programming and/or rejoining the programming late. This is because DAI providers have re-used the legacy signaling technology used by cable operators to insert local commercials because that is all that was available. With direct automation integration, the frame accurate timing, identification of each ad within a break and contextual information is all signaled in-band enabling operators to truly take advantage of the opportunity DAI provides.

### 5.3. Live-to-VOD

To boost revenue, advertising supported broadcast TV programs which 'air' initially on a channel are made available for later viewing on-demand and on any device. [12] For the first three days after a program airs, the same advertising that was originally broadcast is reused. Under Nielsen's C3 ratings system, any viewings occurring during the 3-day window count as if it were watched live. Afterwards (known as D4), the advertising needs to be changed to maximize content monetization. [13] If the show has low ratings it may also be advantageous to replace some of the ads before the C3 window ends. Unfortunately, today generating these on-demand files in the first place is a manual process which is inefficient, and expensive. The current process also limits the scalability, especially when it comes to re-purposing content with new advertisements at the end of the C3 window.

Today the manual processing takes most broadcasters several hours to execute, which eats into the C3 window, thus reducing the value. However, with the right metadata (i.e. that obtained through direct integration with the broadcast automation system) the entire end-to-end process can be completely

automated. This not only increases revenue (because it provides more opportunity for the program to be viewed online), it also reduces costs because manual processing is no longer required.

### 5.4. Start-over-TV

Many cable operators and MVPDs offer start-over[1] capability to users – enabling those that miss the start of a program to restart it from the beginning. It is a very popular feature; however, the flaw is the use of the EPG start time rather than the actual start time – the difference can be several minutes or longer depending on the circumstances. In order to ensure that nothing is missed, the user is transported back to just before the beginning of the show usually midway through the segment directly before it, which results in a poor experience. Obviously with the metadata described herein signaled in SCTE 35 makes it possible to implement a totally frame accurate start-over function. Additionally SCTE 224 policies can describe the rights for each individual piece of content – whether or not it can be started again, DVR'ed etc. – and there is much less risk of errors in implementing these policies because the asset IDs in the SCTE 35 in-band messages can automatically be matched with the policies associated with those assets in SCTE 224.

### 5.5. Interactive TV

Perhaps some of the most interesting capabilities enabled by signaling frame accurate metadata are interactive applications. The ability to signal the playlist in advance enables users to be notified in some way that certain content is about to start – this could be a program or an advertisement. This enables a much richer experience to be provided by the broadcaster to their viewers. The possibilities are literally limitless.

It is clear that the train has already left the station with regard to broadcast TV being streamed over the Internet but efficiently extracting the right metadata could significantly speed up that train as well as give the carriages a lot more functionality. Rather than killing broadcast TV, OTT can actually save it!

[1] Startover™ is a Trademark of Time Warner Cable.

# 6. Abbreviations and Definitions

## 6.1. Abbreviations

| | |
|---|---|
| DAI | Dynamic Ad Insertion |
| EPG | Electronic Programming Guide |
| GPI | General Purpose Interface |
| HLS | HTTP Live Streaming |
| HTTP | Hypertext Transfer Protocol |
| ISBE | International Society of Broadband Experts |
| MAM | Media Asset Management System |
| MBR | Multi-Bit Rate |
| MCR | Master Control Room |
| OTT | Over-The-Top |
| SCTE | Society of Cable Telecommunications Engineers |
| SDI | Serial Data Interface |
| TVE | TV Everywhere |

## 6.2. Definitions

| | |
|---|---|
| Downstream | Information flowing from the hub to the user |
| M3u8 | M3U file encoded in utf-8 |
| MPD | Media Presentation Description |
| MPEG-2 TS | Transport stream used in digital video broadcast and DVDs |
| MPEG-DASH | Transport stream for internet delivery |

# 7. Bibliography and References


1. http://www.cnbc.com/2017/02/28/youtube-announces-skinny-tv-bundle-.html
2. http://www.dailymail.co.uk/news/article-2853481/It-s-kind-like-horse-good-car-Netflix-CEO-says-broadcast-TV-dead-2030.html
3. https://www.verizondigitalmedia.com/blog/2017/03/get-a-platform-with-smarter-delivery-to-solve-your-video-quality-challenges/
4. ANSI/SCTE 35 2016, Digital Program Insertion Cueing Message for Cable
5. ANSI/SCTE 67 2014, Recommended Practice for SCTE 35 Digital Program Insertion Cueing Message for Cable
6. ANSI/SCTE 224 2015, Event Scheduling and Notification Interface (ESNI)
7. https://support.uplynk.com/doc_live_slicer_api.html
8. ANSI/SCTE 104 2015 Automation System to Compression System Communications Applications Program Interface (API)
9. http://eidr.org
10. http://www.ad-id.org
11. https://www.martechadvisor.com/articles/display-native-ads/2017-predictions-the-intersection-of-native-advertising-and-contextual-commerce/
12. http://www.huffingtonpost.com/entry/capitalising-on-mobile-video-demand_us_58b9c4c8e4b02eac8876ce27
13. http://telestreamblog.telestream.net/2017/02/daic4c3/

# Adoption of Unified Distribution Architecture for Overcoming Cost and Monetization Barriers in Multiscreen Service Environment

A Technical Paper prepared for SCTE/ISBE by

Imagine Communications Research & Development
3001 Dallas Parkway, Suite 300
Frisco, TX 75034

# Table of Contents



# 1. Introduction

Disruption continues to roil the premium video marketplace. Content producers and distributors are under mounting pressure to find a uniformed approach to reaching all screens with minimal costs, strategic flexibility and a clear pathway to revenue growth.

Providers of every stripe have been contending with audience fragmentation by investing heavily in IP-optimized transcoding, streaming and other infrastructure elements to ensure their content is accessible on connected devices even as they continue to rely on the modes of processing and distribution that undergird legacy pay TV services. But while it made sense at the outset of the multiscreen era to tailor infrastructure support for IP video streaming as an adjunct to the core pay TV infrastructure, today the dual-silo approach stands in the way of realizing the full ROI potential in both domains.

Spending on components performing duplicative processes, along with continued reliance on proprietary hardware to support traditional pay TV functionalities, drives costs ever higher while limiting responsiveness to consumer demand and new revenue opportunities. Consequently, there's nothing more important to the success of all players than implementing a Unified Distribution architecture that eliminates unnecessary costs and opens the door to seamless execution of advanced advertising and content strategies across all end user access points.

Wide-scale moves in this direction have been blocked to this point by business realities. Content distributors are prevented from replacing legacy infrastructure elements, especially set-top boxes and the associated local pay TV transport mechanisms, without assurance of revenue that will justify the expense. Simply collapsing everything end to end onto the IP infrastructure is not a viable option as long as non-IP set-tops dominate the subscriber landscape.

Yet, it's essential that broadcasters and distributors be empowered to bring the advantages of IP technology into the legacy domain, including monetization through dynamic advertising; personalization and localization of content; virtualization of processes through use of software running on commodity servers; efficient aggregation and analysis of data, and much else. The challenge, then, is to create a Unified Distribution architecture that can be cost effectively deployed to eradicate as much of the legacy infrastructure as possible while extending the benefits of IP technology over QAM and IPTV links to set-top boxes.

To meet this challenge, content distributors and broadcasters require a distribution architecture that takes an innovative approach to utilizing the mechanisms that have made HTTP (Hypertext Transfer Protocol)-based adaptive bitrate (ABR) streaming a viable means of delivering TV-quality video to connected devices.

By making fragmentation for ABR distribution central to the encoding process at core distribution facilities, broadcasters and MVPDs (multichannel video programming distributors) can rely on one mode of distribution for every piece of content, eliminating the need at various secondary and tertiary staging points for the transcoding and packaging processes currently used to format premium video for HTTP delivery to connected devices. This, of course, requires a departure from the way providers utilize terrestrial backbones, where premium content is encoded and encapsulated in IP packets at core facilities for multicasting in continuous UDP (User Datagram Protocol) streams to regional and local distribution facilities.

To make this possible while ensuring support for TV distribution in the legacy domain, an edge platform designed to convert ABR fragments into the continuous UDP-encapsulated MPEG-2 Transport Streams that are compatible with legacy set-tops is required. Ideally, this HTTP-to-UDP gateway should be software based and run on the same type of COTS (commodity off-the-shelf) hardware that's commonly used to support today's content delivery networks (CDNs).

With an HTTP-to-UDP gateway at the network edge, content is converted for local distribution over legacy pay TV links. In this way, set-top-connected TV sets are able to deliver a viewing and advertising experience that is more closely aligned with what viewers experience when accessing content on IP-connected devices. Consolidation onto the ABR infrastructure also adds efficiencies to time-shifting in the legacy domain through use of content fragmentation in cloud DVR, catch-up and trick-play applications.

From a monetization standpoint, the incorporation of the TV set into the dynamic advertising paradigm creates opportunities for both broadcasters and service providers to enable converged multiscreen ad campaigns. Such convergence, enabling the localized and personalized ad targeting across all screens that has long been sought by ad agencies and their clients, has the potential to drive higher CPMs and, with them, higher total revenues than can be attained under current conditions, where linear advertising over legacy channels is treated as a business apart from dynamic advertising in the digital realm.

These benefits can be achieved by broadcasters and network service providers working independently of each other in the sale, respectively, of national and local avails. But there's also an opportunity to grow revenue even farther in both segments through collaboration in the use of ABR technology with HTTP-to-UDP gateways. Service providers, by making CDN capabilities available to broadcasters on a wholesale basis, could generate new revenue for themselves while expanding the higher-priced targeted ad reach for broadcasters' national ad campaigns.

## 2. The Growing Reliance of the Pay TV Business on OTT Strategies

As things stand today, broadcasters' and distributors' efforts to ensure their offerings reach the broadest possible audience by bringing IP-connected devices into the viewing matrix have grown increasingly costly without a compensating increase in revenue. Judging from trends in consumer behavior and the emergence of new video formats, these costs will continue to mount at unacceptable levels so long as providers continue to rely on separate infrastructures to deliver content in the legacy and IP-connected domains.

The pay TV market is undergoing major disruption as providers move beyond the original TV Everywhere (TVE) paradigm where IP-streamed offerings of live and on-demand premium video were targeted to pay TV subscribers as a way to enhance the appeal of the legacy services. Now, growing numbers of broadcasters and distributors see an urgent need to use Internet streaming technology to engage consumers who are not drawn to traditional premium packages – the so-called cord nevers and a growing legion of cord cutters.

### 2.1. Trends Lines in Consumer Viewing Behavior

In the U.S., according to Parks Associates, as of Q3 2016, 63 percent of broadband households were subscribing to at least one OTT video service, and 31 percent were subscribing to two or more. [1] Parks reported that all of the top ten OTT services increased their subscriber bases in 2016. Some of these providers, including the top two, Netflix and Amazon, focus on delivering on-demand content, including

TV episodes, while others feature linear broadcasts as well as stored content. In all cases, they are providing viewers alternative means of reaching content that was once only available through legacy pay TV subscriptions or over-the-air broadcasts.

Network traffic tracker Sandvine said the average North American household in 2016 had at least seven active devices in use daily, with video streaming accounting for 65 percent of usage. [2] By 2019, the average number of connected media-enabled devices per North American household will climb to ten, according to IHS projections. [3]

The TV, of course, is now among the devices playing a big role in the growth of online video consumption, thanks to the proliferation of smart TVs and a wide range of IP streaming media players. In the U.S., 34 percent of all broadband households own a smart TV and 26 percent own streaming media players supplied by Roku, Apple, Amazon, Google and others, according to Parks Associates. [4]

By the end of 2017, streaming media player penetration of U.S. broadband households will reach 40 percent, according to NPD Group, marking a 150 percent increase since the beginning of 2014. [5] Overall, counting media players, smart TVs, game consoles and Blu-ray players, 211 million devices will be in place to enable OTT viewing on TV sets in U.S. households by YE 2017, NPD predicts.

These trends are replicated worldwide. U.K. analyst Ovum in a recent report predicted online video subscriptions, which topped 100 million worldwide at the end of 2015, will increase to 177 million by 2019. [6] In another report, Digital TV Research projected that by 2020 online subscription revenues will reach $21.6 billion, three times the 2014 total, with penetration exceeding 33 percent in ten countries. [7]

Variations often occur among findings from different research firms and all end-user surveys can potentially yield subjective results based on the audience polled or the phrasing of a question. In addition, market projections are often updated frequently to reflect shifts in purchasing or usage trends.

## 2.2. Broadcaster Responses

In response to these trends, broadcasters are licensing content to a greater number of OTT subscription services, including not only competitors to traditional MVPDs but also a growing number of MVPDs that are moving beyond the TV Everywhere model to offer standalone OTT services. At the same time, many, if not most, broadcasters have implemented direct-to-consumer (DTC) strategies that leverage the versatility of IP-based production, post-production and distribution processes to deliver unique blends of their branded content to consumers, often without requiring them to be pay TV subscribers.

DTC strategies have become especially important to building an international presence for many U.S.-based TV and cable networks. Discovery Communications, for example, reaches some three billion subscribers in more than 220 countries and territories through Discovery Channel and multiple other network brands. With its launch of DPlay in 2015, Discovery entered the DTC subscription market, starting with an $8 monthly service combining live and on-demand content in Norway, Denmark, Italy and Sweden. And the company has become a major sports network for Europe and Asia-Pacific regions through its recent acquisition of Eurosport.

Another ambitious undertaking is NBCUniversal's GlobalNetworks, which manages a variety of acquired properties in Europe and Asia as well as internationally formatted iterations of branded NBC cable networks. Fox, too, has an aggressive global OTT initiative in place through Fox International, which partners with leading regional platforms to make programming available online, sometimes licensing

content to third-party aggregators, sometimes offering a DTC package on its own. Disney/ABC, Viacom, AMC Networks, Scripps Networks and Time Warner's HBO and Turner Broadcasting System are other examples of U.S. media companies with notable international expansion strategies in play.

### 2.3. MVPD Responses

Meanwhile, the ranks of MVPDs offering standalone OTT services now include all the Tier 1 providers in the U.S. and many more abroad. Dish Network's Sling TV, the first MVPD offering to launch in the U.S. is now ranked sixth in Parks' top ten behind Netflix, Amazon Prime, Hulu, MLB.TV and WWE Network and ahead of HBO Now, Crunchyroll, Showtime and CBS All Access.

Like Sling TV, Verizon's more recently launched Go90 OTT subscription service and AT&T's DirecTV Now are available nationwide. DirecTV Now, which launched in late 2016 with packages that include a 100-channel lineup, currently priced at $60 monthly following an introductory offering at $35 per month, attracted over 200,000 subscribers in its first month of operation.

Taking a somewhat less aggressive approach, cable leaders Comcast and Charter Communications are limiting access to their offerings, Xfinity Stream and Spectrum TV Stream, respectively, to their own broadband subscribers. And the number three telco, CenturyLink, which has abandoned its legacy Prism IPTV service, says it may launch its own OTT service in some of its operating territories during 2017.

As in the U.S., MVPDs elsewhere are reacting by going beyond TVE enhancements to legacy pay TV services with their own standalone OTT services. Some are in the SVOD mode, such as Netherlands-based Altice Group, which in late 2015 launched its €9.99 Zive SVOD service in France followed by expansion in 2016 to six other countries. The MVPD says the service, now offering 15,000 HD and 400 UHD titles, will be rolled out by Altice USA, the third-ranked stateside MSO, at an unannounced date. [8]

Other European operators are offering both live and VOD programming with their new OTT services. Major players in this space include Sky with NOW TV in the U.K. and other versions offered through Sky Deutschland and Sky Italia, Belgium cable operator VOO with Be tv Go, Orange in France and Vodafone in Germany and Spain.

Increasingly, MVPDs are embracing unified modes of video processing across legacy and OTT outlets as they upgrade facilities to support more advanced services. For example, Vodafone Deutschland is offering its new UHD-enabled GigaTV service both as a traditional pay TV service and as an OTT option, which at €9.99 per month includes 120 TV channels, 55 catch-up portals and pay-per-view access to over 3,000 VOD movie titles.

## 3. The Emergence of a TV-Caliber ABR Infrastructure

All these strategies underscore the growing confidence of MVPDs and broadcasters in the viability of IP-based streaming as a means of delivering services that measure up to the performance standards set for legacy pay TV. ABR streaming technology, originally used primarily for delivering on-demand content at relatively low resolutions to personal devices, has emerged as a viable means of broadcasting live sports, news and other TV programming in 1080p HD or even 4K resolution at quality levels required for viewing on the largest displays.

It's especially noteworthy that the biggest challenge for ABR streaming of TV-caliber live programming, namely coverage of fast-action sports, has been met with resounding success. OTT delivery of

professional and college-level football, baseball, basketball, soccer and other sports to mass audiences in HD is now a routine feature of broadcast operations worldwide with 4K transmissions entering the picture as well.

Perhaps nothing better illustrates what can be accomplished with IP technology utilizing ABR streaming than NBC's live OTT delivery of all events from the 2016 Summer Olympics. One hundred million unique users spent an aggregate time of over 45 million hours watching live coverage on connected devices during the 17-day schedule, according to NBC. [9] IP-based production and post-production processes facilitated storage and preparation of live-captured content for time-shifted viewing with enhanced features offering statistics, athletic profiles and much else relevant to each event and to the games as a whole.

### 3.1. The Technical Foundation for Robust Streaming

Many technological advances have come into play to enable such capabilities over the open Internet, starting with enhancements in the functionalities intrinsic to the ABR streaming platform. ABR was designed for use with HTTP servers as a way to augment the packet-loss recovery mechanisms of TCP (Transmission Control Protocol) in support of maintaining continuity of the A/V stream by adjusting bitrates to accommodate fluctuations in available bandwidth during a user session.

Fundamental to today's ABR streaming capabilities are software-based platforms that utilizes high-density COTS processors to transcode content from original sources at bitrates suited to reaching multiple types of devices at various levels of resolution and frame rates. These platforms are designed to perform all this processing, which might include use of more than one codec for each bitrate profile, in real time for live feeds and at accelerated speeds for content stored for on-demand distribution.

State-of-the-art premium video transcoders execute many other tasks as well, such as de-interlacing of NTSC files to progressive mode; adding IDR (instantaneous decoder refresh) frames to enable SCTE 35-based ad insertion; performing GOP (group of pictures) alignment to facilitate smooth output in the ABR streaming process; making automatic loudness adjustments, and processing and synchronizing ancillary feeds such as closed captioning, picture-in-picture displays and foreign language subtitles.

ABR streaming packagers, software components that can be co-located with transcoders or positioned remotely, utilize a communications framework that allows HTTP servers to send to each client device a "manifest" file of information pertinent to the content the user is accessing, starting most fundamentally with a list of the various bitrates at which the content has been transcoded for distribution. The packagers direct processors to fragment each transcoded version of the content into "chunks" of a few seconds duration, the length of which depends on which of several ABR formats is used with a given streaming session.

While the multiplicity of ABR formats, most prominently including HTTP Live Streaming (HLS), Adobe HTTP Dynamic Streaming (HDS) and Microsoft Smooth Streaming, complicated use of the technology through its formative years, HLS has emerged as the dominant mode with close to universal support on recent vintage devices while HDS and Smooth have been made interoperable through the standardized format known as MPEG-DASH. With Apple remaining a non-participant in the DASH initiative, it appears likely DASH and HLS will coexist as anchors to a significantly less-complicated ABR ecosystem for some time to come.

Throughout the streaming session the client device continually signals to the server which bitrate profile should be sent with each succeeding fragment based on how much bandwidth the device determines is available over its connection at that point in time in the context of how much processing power the device CPU has for handling various frame rates and levels of resolution. Thus, for example, even if enough bandwidth is available to send a chunk at 60 frames per second with 1080p resolution, a device equipped to render resolution no greater than 480p at 30 fps will ask for fragments transcoded at the appropriate bitrate, thereby avoiding over use of bandwidth and buffering delays that result from over saturation of the CPU.

## 3.2. Advances Enabling Transformations in Monetization, Services and Operations

Beyond the robust performance capabilities of multi-bitrate ABR streaming, several other factors are contributing to the utility of ABR in pay TV operations. At the top of the list are the functionalities introduced through what is known as “manifest manipulation.”

Over time the information communicated in manifest files has been expanded to include data telling clients where to find and exactly when to pull content such as advertising, special features or alternative programming that the distributor wants any given user to receive at any point during the session. These adjustments through manifest manipulation are implemented in tandem with use of HTTP applications servers, ad decision systems and other elements, all of which are managed by orchestration platforms that employ data analysis and policy servers to call for assets relevant to a particular transaction based on the location and profile of the end user.

Another function crucial to using ABR in premium video distributions is automated support for time shifting. Applications servers linked to the packaging platforms enable implementation of multiple time-shifting modes under the control of end users, including trick-play functions, catch-up viewing in limited time windows and cloud-based DVR options utilizing long-term storage facilities.

In addition, it’s now much easier than it once was to apply whatever mode of content protection is suited to the type of content and device associated with a given session. In today’s premium service operations the execution of encryption, device authentication, user certification and DRM management is performed under the guidance of multi-platform systems that support on-the-fly association of policies and DRMs specific to user and device profiles for each live and on-demand session. Some of these systems also incorporate support for applying traditional conditional access (CA) protection to content destined for viewing over legacy Pay TV streams.

Rounding out the advances that have made ABR a reliable mode of delivering TV-caliber performance to connected devices, including large-screen displays, are quality assurance (QA) platforms that orchestrate both the QoS (quality of service) functions that enable fast, often proactive measures against network-based threats to expected service performance and the QoE (quality of experience) functions that track and analyze data from every viewing session in support of various business models. Utilizing advanced analytics engines that tap into data flows from network elements and user devices, these QA platforms provide content producers and distributors the means to ensure the caliber of performance they’re looking for is adhered to in the production, post-production and distribution processes as well as in the application of dynamic ad placement and other functionalities tied to manifest manipulation.

### 3.3. The Expanding Role of CDNs

A major factor underlying the viability and flexibility of today's IP video infrastructure is the fact that all the foregoing functions as well as all the functions performed by IP-optimized production and post-production platforms are now executed by software systems running on COTS facilities. As a result, all the advances supporting shared use of resources through standards-based virtualization technologies can be leveraged by every element in the end-to-end IP video infrastructure. This means that broadcasters and distributors will be able to exploit whatever cost benefits emerge with ongoing advances in the virtualization domain to enable adjustments to ever-changing market needs in the most cost-efficient ways possible. Selective application of virtualization, however, is sometimes still required. Depending on the mix of applications, Virtualization does not always yield statistical multiplexing gains. It also can introduce tuning conflicts between guests or between guests and a host. In addition, the implementation of a hardware abstraction layer can at times reduce performance.

Anchoring the implementation of all these capabilities is the fact that all providers in the global premium video ecosystem rely on the processing capabilities of public and private CDN facilities positioned at the edges of metropolitan areas and often at points closer to end users. Originally designed to provide caching support for the most frequently accessed on-demand video content, COTS-based CDN facilities in growing numbers have been expanded and enhanced to support the more advanced capabilities of ABR technology as discussed above for both on-demand and live feeds.

Now these facilities can be farther enhanced to make it possible to integrate legacy pay TV distribution with the advanced IP streaming infrastructure in support of the Unified Distribution architecture described in the following section. Critically, these enhancements can be readily implemented in software utilizing COTS appliances, thereby avoiding the need for investments in purpose-built hardware.

Universal reliance on CDN resources in the IP streaming arena creates a location at the network edge where those ABR fragments can be converted to MPEG-2 transport streams that can be conveyed in IP packets over UDP transport to end users. This requires new software that can be easily installed to run on the COTS servers that populate CDN facilities.

## 4. The Mounting Case Against Operating in a Two-Silo Environment

Notwithstanding the emergence of ABR as a TV-caliber mode of distribution, the premium video industry continues to rely on the legacy mode of end-to-end content distribution as the core architecture while investing increasing amounts in the overlaid ABR infrastructure, resulting in duplicative spending on transcoding platforms and other elements to keep up with evolving requirements in the IP and legacy pay TV domains. This dual-silo approach to infrastructure investment is becoming ever more untenable as the industry moves into a new era in TV technology.

### 4.1. The Need to Support UHD in the Legacy and IP Domains

Ultra HD (UHD) has been slow to get off the ground, much as was the case in the early going with HD. But whatever the time frame for introducing UHD services on a mass scale turns out to be, it makes no sense to have to invest in two infrastructures to support implementation of those services.

The prospects for market adaptation to UHD have been greatly improved with the inclusion of HDR (High Dynamic Range) technology as a true differentiator in user experience. As a result, supporting UHD has come to mean investment in not only the encoding and CPE infrastructure required for

transmitting programming at 4K resolution but also the processing mechanisms that are needed to add HDR enhancements to 4K content and, likely, HD content as well. And, of course, the same types of dual-silo cost barriers loom when it comes to planning for future investments in other formats, including virtual reality (VR) and, eventually, 8K UHD.

Moreover, the onset of UHD brings another major infrastructure adjustment into play in conjunction with meeting new security requirements for licensing the highest-value content, including HD as well as UHD programming. With the need to thwart video piracy emerging as a major priority across the motion picture and TV programming sectors, the Enhanced Content Protection (ECP) specifications issued in 2015 by the motion picture studios' tech consortium MovieLabs are now widely viewed as a template for protecting all types of premium content.

This is going to require many adjustments in content protection infrastructure. New stipulations include not only the well-publicized need for mechanisms supporting per-session applications of forensic watermarking codes but also stringent requirements for end-to-end link security bearing on the design of core hardware processors, where no CPU will be allowed to perform security-related functions, and use of robust DRM technology that is superior to native DRMs employed with many types of devices.

### 4.2. The Need to Execute New Monetization and UX Capabilities in Legacy Pay TV

Beyond duplicative investments in new technology, there's another, potentially much greater cost incurred with continued reliance on the legacy architecture, which is the difficulty of refurbishing traditional pay TV service with the user experience (UX) and monetization advances intrinsic to the IP streaming architecture. As the UX offered with OTT services delivered to IP-connected TVs becomes ever more compelling, the competitive challenge to legacy services intensifies. To even begin to replicate the IP-platform capabilities in the legacy domain requires ever more investment in infrastructure elements specifically designed to work within the limitations of legacy mechanisms.

Consequently, as advertisers become more enthusiastic about the dynamic targeted advertising opportunity offered in the OTT space, there's no readily available way for broadcasters and distributors to capture higher CPMs for locally and demographically targeted ad placements in legacy pay TV programming. As a result, even though the industry has made some strides with regard to dynamic ad placement in VOD content, it has been stalled in building out the business and supporting infrastructure framework essential to realizing the full potential of addressable advertising in linear programming.

## 5. Realizing the Potential of a Unified Distribution Architecture

### 5.1. Technical Requirements

As the imperatives to eliminate the unnecessary costs of duplicate distribution architecture and to bring the benefits of IP technology into the pay TV arena intensify, the search for a way to address these issues has become mission critical. In other words, it's becoming mandatory that broadcasters and distributors define and implement a Unified Distribution architecture that eliminates duplication, enables use of cloud technology to consolidate, simplify and add DevOps flexibility to operations, preserves the legacy pay TV access infrastructure and extends the benefits of IP-based UX and monetization into that domain.

Clearly, as evidenced in the foregoing discussion, the streaming infrastructure has been proven sufficiently robust to play the role as the core architecture. But even as ever more distributors come to this

realization, the drawback to wide-scale adjustment to this new reality is a perception that it can't be done without incurring the extraordinary costs of replacing legacy STBs and stranding investment in the supporting access infrastructure.

Consequently, three major changes in how things are currently done must be accomplished to achieve the goals of a new Unified Distribution architecture. Notably:

- The current mode of distributing premium video over terrestrial fiber backbones from core broadcaster and MVPD facilities, which entails encapsulation of traditionally encoded programming in IP packets for multicasting in continuous UDP streams to secondary points of distribution, is eliminated in favor of multi-resolution transcoding and fragmenting of content at the points of origin in support of ABR streaming over those backbones with farther packaging for local streaming performed at the secondary distribution points;
- All processing of the fragmented IP content at those secondary distribution points in support of live and on-demand multiscreen service models, including dynamic advertising, personalization of user experience, adherence to local blackout policies, support for time shifting and UX feature enhancements, is performed in software running on COTS appliances;
- To accommodate delivery of the locally processed IP content to legacy set-tops an HTTP-to-UDP gateway is employed to convert the ABR fragments to continuous MPEG-2 Transport Streams for delivery over UDP, an IP protocol now supported by the vast majority of STBs.

All mechanisms essential to making this Unified Distribution architecture possible are now available in the marketplace, including the HTTP-to-UDP gateway described above.

## 5.2. The Cost-Saving Benefits

Fragmentation and transcoding of mezzanine-encoded content mapped to all the resolutions suited to everything from handsets to 4K TV sets at core points of distribution not only unifies downstream processing onto a single platform; it eliminates many of the transcoding steps that the current architecture requires in both the legacy and IP domains. For example, video delivered by an MVPD to connected devices goes through four transcoding processes, starting with the point of contribution, next at the point of MVPD ingestion and then in two different processing environments, one for distribution to STBs and the other for ABR streaming to connected devices.

This configuration multiplies costs as each point in the chain is adjusted to accommodate new requirements, and it introduces delays and quality degradations at each step. With implementation of the Unified Distribution architecture, once the transcoding and fragmentation are performed on the mezzanine files, the need for farther processing is greatly reduced.

Rather than requiring multiple transcoding stations and separate infrastructures for processing video, the new architecture consolidates processing for IP and legacy pay TV operations with use of just-in-time packaging at the edge distribution points to execute manifest manipulation for DAI (dynamic ad insertion), blacked-out content replacement, personalization of features and other locally oriented functions. With the addition of the software-based HTTP-to-UDP gateway the locally processed programming can be delivered to legacy STBs.

These advantages apply whether the content is delivered for traditional broadcast linear viewing or for time-shifted applications. When it comes to supporting live content distribution over the backbones, the multiple fragmented versions of the content can be delivered in traditional IP multicast mode in keeping with the low latency requirements of linear programming, including sports. Along with conserving

bandwidth this offers the added benefit of eliminating the variations in latency that occur in the two-silo approach, where content delivered to consumers on their personal devices is typically out of synch with content delivered over the legacy feeds. A rare but potentially negative byproduct of normalizing delay across the legacy and IP delivery paths is the increased opportunity for other information delivery vehicles, namely social networks, to provide "breaking news" in greater advance of fragmented content.

Another cost-saving infrastructure benefit has to do with the content recovery mechanisms intrinsic to ABR. Currently, if a UDP packet is dropped or damaged in transit, the content is rendered useless, which means providers must support alternative backup paths for each route on the backbones. HTTP, with its reliance on TCP, provides packet replacement redundancy in the stream, obviating the need for backup paths. It should be noted that a small UDP span, between the gateway and QAM equipment, still exists on the legacy network. Incidences of failure in this span can be reduced by collocating the gateway and QAM equipment.

Consolidating video processing at MVPD facilities lowers costs of operations in many other ways as well. For example, with this consolidation comes the ability to implement content protection mechanisms, including the watermarking and advanced DRM processes required under new content protection specifications, through the new consolidated multi-platform security systems on offer from industry suppliers.

With the Unified Distribution architecture in place MVPDs can also greatly lower costs while adding flexibility in their migrations to all-IP pay TV services. At any point they can implement a cap-and-transition strategy that preserves the investment in legacy STBs while delivering the pay TV service directly from the streaming packagers to new subscribers and subscribers who upgrade to new tiers tied to the IP service paradigm. Unified distribution, more than just protecting legacy investments, puts distributors on the path to a complete migration to all-IP spectrum, which will eventually be needed to meet the video consumption demands of future audiences.

It's also important to note that, with consolidation of processing onto a single IP platform at local distribution points, MVPDs can use ABR mechanisms to break free of encumbrances and costs imposed by the need to use traditional statistical multiplexing technology to maximize bandwidth efficiency on QAM channels. Currently, while many operators want to capitalize on the bandwidth-conserving capabilities of H.264 compression, which is now widely supported by set-tops deployed over the past few years, the greater complexities associated with H.264 compression pose problems for legacy stat muxes, which means they must be replaced with newer equipment or the transition to H.264 has to be delayed.

With reliance on ABR feeds into the HTTP-to-UDP gateway operators can avoid this problem by capitalizing on the ABR transcoding and fragmenting system to replicate the benefits of stat muxing. In this model, the client in the gateway asks the streamer to send fragments for each stream associated with a given QAM channel at the minimum bitrates required to achieve the required quality level, thereby achieving maximum bandwidth efficiency across the entire QAM channel.

Another cost component eliminated with implementation of the new architecture is splicing technology used in blacked-out content replacement and advertising. In the case of blackouts, the need to precisely splice replacement programming at local distribution points through manual operation of splicing equipment is a costly headache that goes away with the use of just-in-time packaging with manifest manipulation to perform the task.

Similarly, using a single HTTP-based ad insertion system tied to the software-managed manifest manipulation process eliminates the need for traditional ad splicers at points of content origin and secondary distribution points. For MVPDs, the shift to HTTP-based insertion has the added benefit of overcoming barriers imposed by legacy ad insertion technology on the use of H.264 compression, which requires upgrades to H.264-compatible splicers.

### 5.3. Enabling DevOps Agility

Reliance on software-defined uses of commodity datacenter hardware to manage and process video at all locations allows all players to benefit from the flexibility of a software-based DevOps environment. In this environment, providers can introduce new user interfaces, on-board new devices, launch new applications, execute software upgrades and deliver other user benefits almost instantaneously whenever they wish across all screens.

Moreover, operating in the IP domain from a software and cloud-based infrastructure, development teams can directly tune into operational results and make whatever adjustments are necessary with minimum disruption to services. Overseers of market trials can access operations-generated data to learn in real time how innovations are performing and, in many cases, make adjustments to the technology parameters without having to end the trial and start all over again with lab testing and a new trial. They can judge the effectiveness of an offer and adjust terms or replace it with another from dashboards that allow them to monitor results and implement changes with the click of a mouse.

### 5.4. The Monetization and UX Benefits of Leveraging Just-in-Time Packaging

As previously discussed, the advances tied to use of manifest manipulation in the ABR stream packaging process have enabled a wide range of capabilities, including dynamic advertising suited to localization and personalization of ads, blacked-out program replacement and personalization of UIs and other aspects of the user experience.

The important point is to underscore the fact that, with deployment of the HTTP-to-UDP gateway, providers can deliver into the legacy pay TV streams many of the applications enabled by the ABR platform for consumption on connected devices. Obviously, the processing performed on content destined for translation through the gateway must take into account the shared viewing environment of STB-connected TV sets. Zone-based local advertising will be the same across all devices, but distributors will need to avail themselves of settings on advanced platforms that adjust the range of just-in-time applications to suit the viewing situation.

## 6. Conclusion

The pay TV industry has reached a crossroads where cost and revenue impediments attending reliance on the current distribution architecture to deliver premium content to legacy STBs and connected devices are severely limiting the ROI potential in both arenas. The need to support separate video processing infrastructures within this architectural framework makes it very hard to meet spending requirements imposed by new developments, especially when it comes to enabling the transition to next-generation TV formats, including UHD in the near term and virtual reality and 8K UHD farther into the future.

This dual-silo architecture also prevents providers from bringing into the legacy pay TV service domain the advances in dynamic advertising, service personalization and localized content management that are intrinsic to IP-based operations. This is impeding pursuit of new monetization opportunities and creating

a widening gap between the pay TV UX and what consumers can expect utilizing connected TVs and other devices to receive programming over the ABR infrastructure.

Given the proven viability of ABR streaming for premium video distribution, the industry has every reason to eliminate the two-silo approach in favor of a Unified Distribution architecture that relies on the transcoding, fragmentation and other components of the ABR infrastructure as the primary video processing framework. But in so doing providers must be able to protect legacy infrastructure investments with technology that supports delivery of pay TV services to legacy STBs.

Moreover, the Unified Distribution architecture must be able to enhance those legacy services with the capabilities that can be derived from an ABR infrastructure equipped with the advanced functionalities enabled by manifest manipulation and various supporting components. These enhancements include addressable advertising, personalization of services, local programming adjustments related to blackout rules and much else.

All the advances required to support such a Unified Distribution architecture are now commercially available, including an HTTP-to-UDP gateway,. This innovation makes it possible to convert the fragmented output of ABR packagers to the continuous transport streams that deliver pay TV content over UDP to STBs, thereby ensuring the monetization and UX enhancements enabled by the IP infrastructure can be brought into the legacy service domain.

All components of the Unified Distribution architecture are designed to run as software-based platforms on COTS appliances at the cores and edges of the broadband ecosystem. As a result, providers can operate with DevOps flexibility in pursuit of new revenue opportunities across all screens while setting the stage for migration strategies that will lead eventually to retirement of the legacy STB-based infrastructure.

# 7. Abbreviations and Definitions

## 7.1. Abbreviations

| | |
|---|---|
| ABR | adaptive bitrate |
| AP | access point |
| bps | bits per second |
| CDN | content delivery network |
| COTS | commercial off the shelf |
| FEC | forward error correction |
| HDR | high dynamic range |
| HFC | hybrid fiber-coax |
| HD | high definition |
| HTTP | hypertext transfer protocol |
| UHD | ultra high definition |
| Hz | Hertz |
| ISBE | International Society of Broadband Experts |
| OTT | Over the top |
| TPC | Transmission control protocol |
| SCTE | Society of Cable Telecommunications Engineers |
| UDP | User Datagram Protocol |

### 7.2. Definitions

| | |
|---|---|
| HTTP-to-UDP gateway | A gateway at the network edge, where content is converted from ABR format to UDP for local distribution over legacy pay TV links |
| Addressable advertising | The ability to target commercials at a narrow audience or individuals |

## 8. Bibliography and References


1. Parks Associates, OTT Video Market Tracker, December 2016
2. Sandvine, Global Internet Phenomena Report, August 2016
3. Broadband News, Five Connected Media Devices per Home by 2019, September 2015
4. Parks Associates, press release, January 2015
5. NPD Group, Connected Home Forecast, January 2015
6. Rapid TV News, "OTT Streaming to Hit 100 MN Subs," April 2015
7. Digital TV Research, Global OTT TV & Video Forecast, June 2015
8. Fierce Cable, Altice Remains Silent on U.S. Launch of Zive, June 2016
9. NBC Sports, NBC's Rio Olympics Is the Most Successful Media Event in History, August 2016

# MPEG-H TV Audio System for Cable Applications

A Technical Paper prepared for SCTE/ISBE by

Adrian Murtaza, Research Engineer, Fraunhofer IIS, SCTE/ISBE Member
Am Wolfsmantel 33
Erlangen, 91058
adrian.murtaza@iis.fraunhofer.de
+49 (0) 9131 776 6224

Harald Fuchs, Group Manager, Fraunhofer IIS, SCTE/ISBE Member
Am Wolfsmantel 33
Erlangen, 91058
harald.fuchs@iis.fraunhofer.de
+49 (0) 9131 776 6008

Stefan Meltzer, Technology Consultant, Fraunhofer IIS, SCTE/ISBE Member
Am Wolfsmantel 33
Erlangen, 91058
stefan.meltzer@iis-extern.fraunhofer.de
+49 (0) 9131 776 6115

# Table of Contents



## List of Figures



## List of Tables

**Title** **Page Number**



# 1. Introduction

The MPEG-H TV Audio system has been developed and implemented for broadcasting applications, based on the MPEG-H 3D Audio Standard. The MPEG-H TV Audio system is also often referred to as Next Generation Audio (NGA) as it introduces new features like immersive audio and interactivity using the concepts of object and scene based audio. Dating back to early 2011, a consortium of a large number of companies have begun the standardization effort in MPEG for specifying a new audio codec. Designed to complement the video coding advances for Ultra-HD (UHD) displays with 4K or 8K horizontal resolution.

In 2015, shortly after the finalization of the ISO/IEC MPEG-H 3D Audio standard [1], the MPEG-H 3D Audio Low Complexity Profile (LC) was specified in Amendment 3 [2]. The MPEG-H 3D Audio LC Profile is a powerful subset of coding tools from the MPEG-H 3D Audio standard with the goal of creating a system which allows decoding immersive content within the computational limits of today's consumer devices, while still enabling all new NGA features. Thus MPEG-H 3D Audio Low Complexity Profile is the natural choice for broadcast applications, especially for delivery of UHD services, and the base for the MPEG-H TV Audio system.

Depending upon the geographical location, MPEG-H immersive programs may be delivered according to ATSC or DVB standards for signal reception specification. The MPEG-H TV Audio System has been successfully evaluated during features and listening tests conducted by the ATSC audio committee, the MPEG-H TV Audio system has been adopted by ATSC and is now part of the ATSC 3.0 suite of standards, as A342 Part 3 [3]. DVB has also selected and included the MPEG-H TV Audio system in the DVB specification ETSI TS 101 154 v2.3.1 that defines the usage of audio and video codecs for DVB systems [4].

Recently, the MPEG-H TV 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 [5] that is based on ATSC 3.0. On May 31, 2017 South Korea launched its 4K UHD TV service and the MPEG-H TV Audio system became the first NGA deployment on a 24/7 basis anywhere in the world.

SCTE/ISBE has technically finalized the work for specifying a suite of standards documenting coding and carriage constraints of NGA systems for cable television, including the MPEG-H TV Audio system. The main scope of these specifications is to enable delivery of NGA services over cable networks.

This paper describes the new features of the MPEG-H TV Audio system and its use in cable applications based on MPEG-2 TS or MPEG-DASH, as specified by SCTE/ISBE. More details about the development of the MPEG-H TV Audio System for ATSC 3.0 and considerations on the production workflow and the end-to-end signal flow are provided in [6] and [7], while [8] summarizes the MPEG standardization project and offers an overview of the system architecture, capabilities and performance of MPEG-H 3D Audio.

# 2. MPEG-H TV Audio System Features

The evolution of previous MPEG Audio multi-channel coding technologies has been mainly based on introduction of new coding tools, and improvements of existing ones, in order to achieve better coding efficiency. In contrast, the development of NGA systems was primarily focused on enabling new features

as well as improving the audio coding efficiency. In this way, the NGA systems enhance the prior audio technologies with support for immersive sound, personalization, interactivity and universal delivery. These features are described in detail, within the context of the MPEG-H TV Audio system framework, in the remainder of this paper.

## 2.1. Immersive Sound

Delivery and reproduction of immersive sound is a key feature of NGA systems. Distinguished from surround sound by expanding the sound image in the vertical dimension (i.e., the sound can come from all directions, including above or below the listener's head), immersive sound offers the user a more enveloping and realistic experience. This increases the spatial impression and makes the viewer feel more "part of the scene", diminishing the awareness of being a remote viewer.

Immersive sound is usually delivered using one of the three well-established formats:

- channel-based, where traditionally, each transmission channel is associated with a precisely defined and fixed target location of the loudspeakers relative to the listener
- object-based, where each individual audio object may be positioned in three dimensions independently of loudspeaker positions based on the associated side information
- scene-based (or Ambisonics), where a sound scene is represented by a set of Spherical Harmonic coefficients that have no direct relationship to channels or objects, but instead describe the sound field

For channel-based transmission and playback, an immersive sound experience is usually created by enhancing the traditional 5.1 and 7.1 surround loudspeaker configurations with overhead loudspeakers. Typically, four height loudspeakers are added on an upper layer above the middle layer (i.e., the horizontal listener plane) in a home environment.

Other immersive setups can be created by using a different number of loudspeakers. Several sound systems for broadcasting applications beyond the 5.1 sound system are described in [9]. In MPEG, the most common loudspeaker configurations are specified in the MPEG Coding-Independent Code Points (CICP) standard [10]. The CICP loudspeaker configurations supported by the MPEG-H TV Audio system are specified in [1], and include, among others, configurations such as: stereo, 5.1 surround, 7.1 surround, 5.1+4H, 7.1+4H, 10.2(b), and 22.2.

Object-based representations of immersive sound scenes are becoming more popular among sound producers. For example, audio objects can be used for conveying sound effects such as the fly-over of a helicopter. The main difference between objects and channels is that the spatial position of an audio object can vary over time, and the positioning information is carried as side information amongst other metadata used to fully describe the object. This metadata enables the decoder to render the object to the final loudspeaker setup at the receiver side. Usage of object based audio requires a change in the way how content is produced, and the metadata created during production plays a crucial role. For these reasons, the ITU has led the effort for specifying an open common metadata model, called the Audio Definition Model (ADM) [11]. This model allows a representation of object based audio independent from the method used to carry and reproduce the content for final reproduction.

The third option of conveying immersive sound is using scene based audio, which in the case of the MPEG-H TV Audio system is accomplished with Higher Order Ambisonics (HOA). The HOA approach is designed for capturing and representing a sound field. The HOA coefficient signals can be easily

manipulated using only matrix operations. This makes the sound field representation adaptable to different reproduction methods, such as, for immersive sound playback over headphones with enabled head tracking. The rotation of the audio scene could be done in this case by only a matrix multiplication between a rotation matrix and the HOA coefficients. This makes HOA very attractive for virtual and augmented reality applications. Similar to object based immersive sound, the scene-based approach allows for adaptation to any output format, since the HOA signals are not associated with any specific channel loudspeaker layout. More details about this approach are provided in [6].

While any one of these three formats can be used individually for the delivery of complete audio scenes, the MPEG-H TV Audio system also supports any combination of them. The most common example for broadcast applications, including delivery over cable networks, is a mixture of a fixed immersive channel "bed" (e.g., 7.1+4H) and several additional audio objects (e.g., several languages for broadcast or overhead spatial effects for cinematic applications).

The MPEG-H 3D Audio standard supports delivery of a high number of channels, objects and HOA signals, for instance, up to 128 channels and 128 objects can be transmitted simultaneously. For playback, the transported signals can be mapped to a maximum of 64 loudspeakers.

The MPEG-H 3D Audio LC Profile has been created for broadcast applications limiting the number of signals to a reasonable number that can be handled in broadcast production workflows and still enable all NGA features. At the same time, these constraints reduce the computational complexity for decoding and rendering in CE devices to practical values. Table 1 provides an overview of the five levels of the LC Profile with increasing complexity ("Level 5" being the most complex), as specified in [1]. From those levels, the "Level 3", with a maximum of 16 simultaneously decoded core codec channels (i.e., channels, objects or HOA signals) and a maximum of 12 loudspeaker outputs, has been selected for ATSC 3.0, DVB and SCTE. Although a "Level 3" compliant bit stream can contain up to 32 different encoded signals, only 16 can be simultaneously decoded, according to the constraints signaled as side information For example, the stream can contain dialog objects in several languages, but only the dialog objects for one language will be decoded and played back).

**Table 1 – MPEG-H 3D Audio Low Complexity Profile**

| Profile Level | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Maximum Sampling Rate, kHz | 48 | 48 | 48 | 48 | 96 |
| Maximum core codec channels in bit stream | 10 | 18 | 32 | 56 | 56 |
| Maximum simultaneously decoded core codec channels | 5 | 9 | 16 | 28 | 28 |
| Maximum Loudspeaker outputs | 2 | 8 | 12 | 24 | 24 |
| Example loudspeaker configuration | 2.0 | 7.1 | 7.1+4H | 22.2 | 22.2 |
| Maximum Decoded Objects | 5 | 9 | 16 | 28 | 28 |

## 2.2. Personalization and Interactivity

Next Generation Audio systems enable viewers to interact more with the content and personalize it to their preference. This implies shifting the production paradigm towards an object-based audio approach. The MPEG-H TV Audio system metadata carries all the information needed to enable viewers to manipulate the audio objects by attenuating or increasing their level, disabling them, or changing their position in the three-dimensional space. Although it would be possible to create complex audio scenes only with audio objects and allow full control over these objects, this is not practical in a broadcast

application. The simplest and most effective way to enable interactivity in broadcast is a combination of channel or HOA based audio with a small number of audio objects that viewers can manipulate.

Usually, content providers and broadcasters desire control over the viewers' degrees of freedom to alter the way in which content is consumed. This is why the MPEG-H TV Audio system offers advanced metadata structures that empower broadcasters to enable or disable interactivity options and to strictly set the limits in which the user can interact with the content. Moreover, the MPEG-H metadata enables broadcasters to provide several versions of the content, as so-called "presets", which describe how all channels, objects and HOA signals are mixed together and presented to the viewer. Choosing between different presets is the simplest way to interact with the content, and will probably be used by most of the TV viewers. Additionally, advanced interactivity settings can be offered to more experienced users to manipulate objects individually. Figure 1 illustrates a typical user interface for providing a simple menu for selection of different presets (left side), and a more advanced interactivity menu after entering the advanced mode (right side).


**Figure 1 – Typical MPEG-H user interface for preset/advanced interactivity.**

### *2.2.1. Dialog Enhancement*

The most common complaint received by broadcasters today is related to dialog intelligibility. Sound engineers always try to create the broadcast mix as a "compromise" between the level of the dialog and the background (e.g., music, stadium crowd, etc.). Still, viewers may prefer a different mix, as shown by existing studies and experiments [16]. For example, hearing impaired people would benefit from a higher volume of the dialog. Similarly, a higher level for dialog during playback can help viewers consuming the content in a secondary language, or one in which the dialog is competing with a very noisy environment. In these situations, the Dialog Enhancement (DE) feature allows the audience to adjust the content to their own personal preferences, enhancing their listening experience.

The MPEG-H TV Audio system enables Dialog Enhancement by transmitting the dialog as an independent audio object. The viewer can therefore increase the dialog level for better intelligibility, or simply adjust it to create their own customized mix. Moreover, the stream can contain one DE preset (or several, such as light DE and heavy DE) and the viewer can either select it manually or the playback device can perform an automatic selection based on the preference settings of the device.

### *2.2.2. Accessibility and Multi-Language Services*

With existing audio codecs, multi-language programs are usually broadcast as separate complete mixes in each language. Using one stream for each mix leads to significant bitrate increase, directly proportional to

the number of additional languages offered. Moreover, if Video Descriptive Services (VDS) have to be provided as additional complete mixes, the required bandwidth would become even more impactful.

The MPEG-H TV Audio system enables a much more efficient way of offering accessibility or multi-language services, using object-based audio, similar to the Dialog Enhancement feature, as described above. With a common channel bed and individual audio objects for dialog in different languages and for audio description, the MPEG-H TV Audio system requires a significantly lower bitrate. For example, assuming a 5.1 program delivered in 3 different languages and VDS service for each language, a legacy system would require transport of six complete 5.1 mixes in six different streams. The MPEG-H TV Audio system uses in this case a single stream carrying one 5.1 mix comprising the ambiance and 6 individual objects for the different languages and VDS services.

Additionally, all features can be enabled in a single stream, simplifying the required signaling and selection process on the receiver side.

For various program types, such as sport programs, additional interactivity options can be provided, such as choosing between biased commentaries, or listening to the team radio communication between the driver and his team during a car race.

## 2.3. Universal Delivery

In the past years, the way media is consumed has changed dramatically, and while the content is delivered over many different channels (e.g., linear broadcast, internet, mobile platforms), the playback device options grew even more diverse (e.g., TV, AVR, PC, portable devices, etc.). Additionally, the listening environment is no longer well-known, the audience watches their favorite shows not only at home using either TV speakers or home theater 5.1 systems, but also on trains and airplanes, with audio options ranging from high-quality headphones to lower quality portable screen speakers.

In this context, the MPEG-H TV Audio system provides not just an audio codec, but a complete integrated audio solution for delivering the best possible audio experience, independently of the reproduction system. It includes rendering and downmixing functionality, and also 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 TV 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 artistic downmixes.

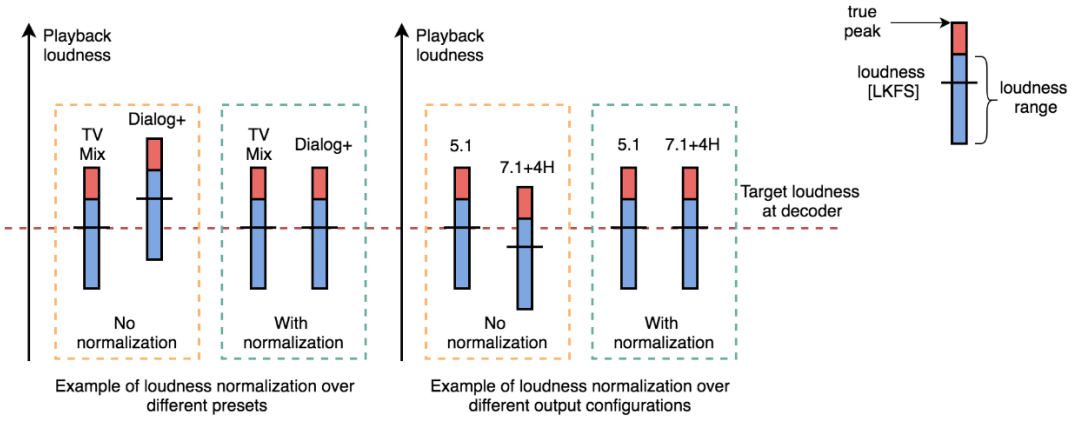



**Figure 2 – MPEG-H Audio Loudness Normalization over different presets and playback configurations.**

Figure 2 illustrates an example of how loudness normalization works for providing the same loudness level for different presets and downmix configurations, based on the loudness information included in the metadata.

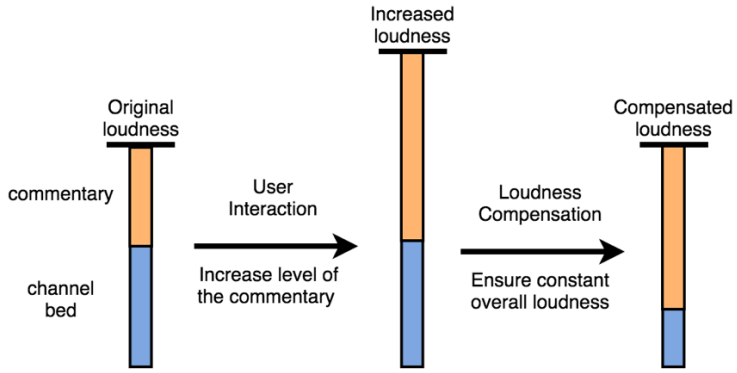



**Figure 3 – MPEG-H Audio Loudness Compensation after user interaction.**

In addition to the loudness normalization module, the MPEG-H TV Audio system includes a loudness compensation component, responsible for adjusting the loudness level after user interaction. For example, if the user increases the dialog level, the overall loudness level would also increase. In this case, the MPEG-H TV Audio system decoder will automatically decrease the level of the complete mix after the user interaction such that the overall loudness level remains constant, as shown in Figure 3.

## 3. MPEG-H TV Audio System

The MPEG-H TV Audio system is an integrated audio solution and consists of several building blocks. Some of those are briefly described in this section with a focus on metadata and stream packaging. For a complete overview, see [6].

## 3.1. MPEG-H TV Audio System Core Codec

MPEG-H 3D Audio is based upon existing MPEG technologies, such as MPEG-4 High Efficiency Advanced Audio Coding (HE-AAC) and MPEG-D Unified Speech and Audio Coding (USAC), to address the coding aspects.

In order to fulfill the new requirements for delivery of immersive audio sound, the MPEG-H 3D Audio core has extended the USAC toolbox, with tools that further improve the coding efficiency based on the perceptual properties of immersive sound reproduction. A set of these tools was selected in the MPEG-H Audio Low Complexity Profile including: Intelligent Gap Filling (IGF), improved Linear Predictive Domain (LPD) mode coding, predictors for Frequency Domain (FD) and Transform Coding Excitation (TCX) modes, and a Multichannel Coding Tool (MCT). A detailed overview of these tools is provided in [6].

## 3.2. MPEG-H Metadata Audio Elements

The MPEG-H TV Audio system enables personalization and interactivity as described above with a set of static metadata, called “Metadata Audio Elements” (MAE). This metadata is static in the sense that it exists once for each piece of content and does not change over time, thus it is part of the overall setup and configuration information of the MPEG-H TV Audio system.

The top-level element is the “AudioSceneInfo”. Sub-structures of the AudioSceneInfo contain the definition of groups, switch groups and presets as well as descriptive information about the objects. An “ID” field is part of all those sub-structures, which is necessary to uniquely identify each group, switch group or preset.

As part of the group definition (“mae_GroupDefinition”) the broadcaster can allow or disallow user changes of the gain level or position and can also restrict the range of interaction through minimum and maximum values for gain and position offset. A group can be “on” or “off” by default and can have a default gain value.

The group description contains other information about the group type (channels, objects or HOA), the kind of the content (e.g., dialog, music, effects, etc.), in the case of a dialogue object, the language, or the channel layout in case of channel-based content. It is also possible to include a textual description of the object, for instance “Network Commentary”.

A switch group is defined in the “mae_SwitchGroupDefinition”. It contains a list of the IDs of all groups that are member of the switch group. Only one of those members can be selected for playback. This way a switch group avoids confusing situations such as two dialog objects with different languages playing back at the same time. Additionally, one of the members of the switch group is marked as default to be used if there is no user preference setting.

The preset definition “mae_GroupPresetData” contains a list of group IDs of those groups that are part of the preset and if they are “on” or “off” in the preset. It is not necessary to include all groups in a preset definition. It is valid to only describe a sub-set of groups in a preset. Groups that are not part of a preset may be switched on or off. Additionally, descriptive information for the preset may be included.

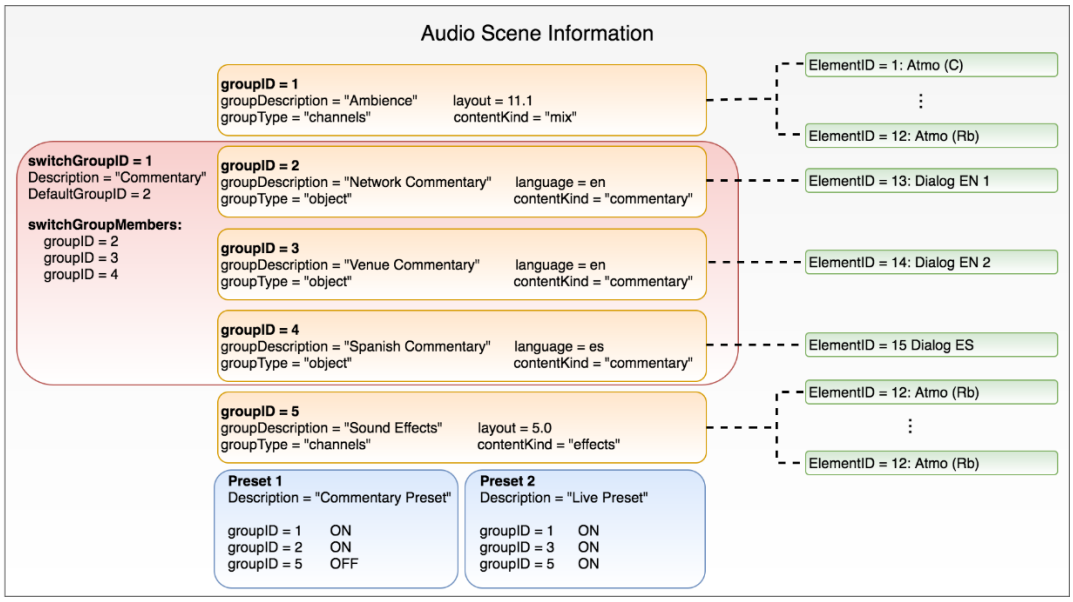



**Figure 4 – Example of an MPEG-H Audio Scene Information.**

Figure 4 contains an example of an MPEG-H Audio Scene Information with five different groups and one switch group. In this example, the switch group contains three commentaries to choose from, two different English commentaries and one foreign language. Additionally, the user may select the "sound effects" object. In this case, the sound effects object is not a single mono source but a multi-channel object with pre-rendered content.

A "commentary preset" for this Audio Scene could contain the groups "1", "2" and "5". The groups "1" and "2" are "on", group "5" is "off". A "live preset" would contain the groups "1", "3" and "5" and all those groups are "on".

### 3.3. MPEG-H Audio Stream

The MPEG-H Audio Stream (MHAS) format is a self-contained, packetized, and extensible byte stream format embedding the MPEG-H Audio data. Several MHAS packet types are specified for carriage of coded audio, configuration data, scene information, loudness information or control data. The packetized structure allows for easy access to configuration information and metadata, without the need to parse the complete bit stream. Moreover, it allows for insertion of additional MHAS packets in an existing MHAS stream.

As illustrated in Figure 5, an MHAS packet consists of a header, payload and stuffing bits for byte alignment. The header is formed by three bit fields: the packet type, which identifies each MHAS packet, the packet label and the packet length information. The packet label has the purpose to differentiate between packets that belong to either different configurations in the same stream, or different streams in a multi-stream environment. Therefore, all MHAS packets with labels in a specific range of values belong to the same stream (i.e., the label values 1 to 16 are used for the main stream, the label values 17 to 32 are used for the first auxiliary stream, etc.). This mechanism allows for the identification of MHAS packets with different labels within the same range of values (i.e., corresponding to one stream) as belonging to different configurations within the same stream.

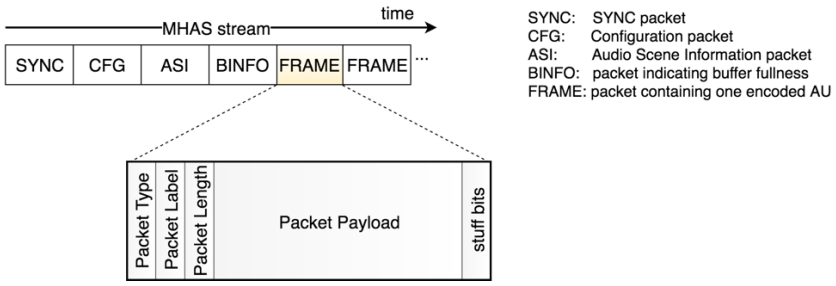



**Figure 5 – MHAS packet structure.**

The most important MHAS packet types are:

- MPEGH3DACFG (CFG) carrying the configuration data,
- MPEGH3DAFRAME (FRAME) carrying the byte-aligned access units (AUs) of the compressed audio bitstream,
- AUDIOSCENEINFO (ASI) carrying the Audio Scene Information,
- SYNC used for transmission over channels with no frame synchronization,
- BUFFERINFO (BINFO) used to indicate the buffer fullness of the encoded stream, and
- AUDIOTRUNCATION (TRNC) carrying the truncation information, used for discarding a certain number of samples at the end or at the beginning of one audio frame.

### *3.3.1. Random Access Point*

A Random Access Point (RAP) consists of the following MHAS packets, in the following order:

- SYNC
- MPEGH3DACFG
- AUDIOSCENEINFO (if Audio Scene Information is present)
- BUFFERINFO
- MPEGH3DAFRAME

The SYNC and BUFFERINFO packets are required only if the MHAS stream is encapsulated into an MPEG-2 Transport Stream. For ISOBMFF encapsulation, these packets can occur, but they will be ignored by the audio decoder.

Additional MHAS packets of different types can be present inside the sequence of packets of a RAP for example, MPEGH3DAFRAME packet may be followed by an AUDIOTRUNCATION packet, with one exception, the MPEGH3DACFG packet must be immediately followed by a AUDIOSCENEINFO packet.

The MPEGH3DAFRAME packet which is part of a RAP has to be encoded as a so-called Immediate Playout Frame (IPF). An IPF is an Access Unit (AU) that is independent from all previous AUs. It additionally carries the previous AUs information that is required by the decoder to compensate its startup 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.

### 3.3.2. Configuration Changes and A/V Alignment

A configuration change occurs in an audio stream when the content setup or the Audio Scene Information changes (e.g., the channel layout or the number of objects change). Usually, this happens at program boundaries, but it 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. This is not a problem in general, but some applications may require that audio and video streams are aligned at certain instants of time to enable stream splicing. Sample accurate stream splicing is very important for cable networks, for purposes of ad-insertion and alternative content, for example.

The MPEG-H TV Audio system 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 trough the AUDIOTRUNCATION packet, which provides information about the number of samples to be discarded and the location of the samples to be discarded (i.e., at the beginning or at the end of the frame). Additionally, the packet signals if the truncation should be applied or not. An AUDIOTRUNCATION packet indicating that the truncation should not be applied can be inserted at the time when the stream is generated and, due to easy access to the MHAS stream from the system level, the truncation can be easily enabled later on. This is only for applications that require support for sample-accurate configuration changes.

Figure 6 shows an example of a sample-accurate configuration change from an immersive audio setup to stereo inside one MHAS stream (the inserted ad is stereo, while the rest of the program is in 7.1+4H). The first AUDIOTRUNCATION packet indicates how many samples are to be discarded at the end of the last frame of the stereo signal, while the second AUDIOTRUNCATION packet indicates the number of audio samples to be discarded at the beginning of the first frame of the new immersive audio signal.

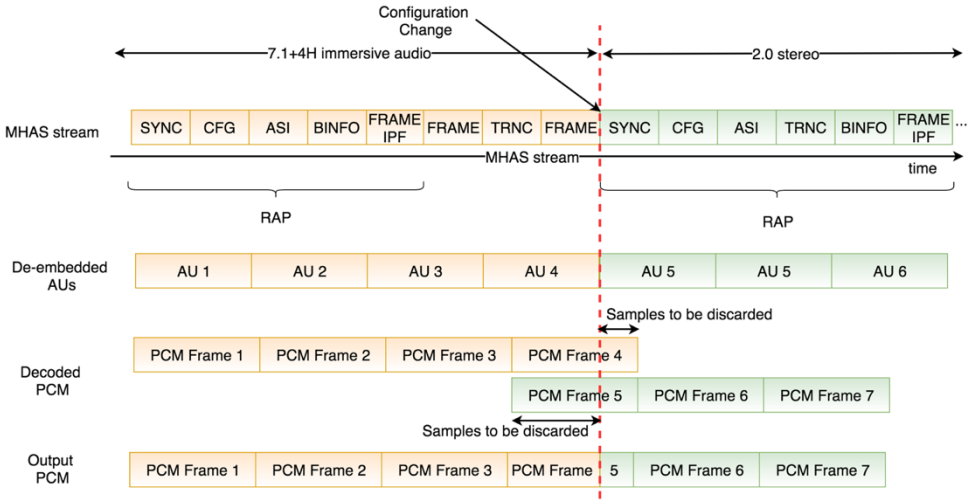



**Figure 6 - Example of a configuration change from 7.1+4H to 2.0 in the MHAS stream.**

### 3.4. Multi-Stream Environment

The MPEG-H TV Audio system can enable all NGA features described in the previous sections using only one stream. This is a much more efficient and robust solution compared to legacy codecs. For applications that involve a hybrid delivery system (i.e., a main MPEG-H stream delivered over broadcast and additional MPEG-H streams delivered over broadband), MPEG-H allows distribution of the audio components composing an audio scene over several streams. The MPEG-H metadata assists the decoder to correctly decode all streams and present the various signaled presets.

When the content is delivered only in a linear fashion, for example over MPEG-2 TS, it is recommended to use only one MPEG-H stream. There might be applications that could benefit from a multi-stream approach, even for such linear delivery systems. For example, if the content is produced for distribution over various platforms using a multi-stream approach, a cable operator can choose to re-use the streams for delivery over MPEG-2 TS without any additional processing.

The MPEG-H TV Audio system specifies for multi-stream delivery, independent of the transport format, a mechanism for receivers to merge several MHAS streams into a single stream. This is realized based on the metadata available on the MHAS stream level, without the necessity of decoding any audio data. In this case, the merged stream can be fed into a single MPEG-H TV Audio system decoder. Figure 7 provides a simple scenario with three MHAS streams:

- Stream 1 contains the channel bed and the main dialog in the original language,
- Stream 2 contains the main dialog in a different language, and
- Stream 3 provides the VDS service with audio description in original language.

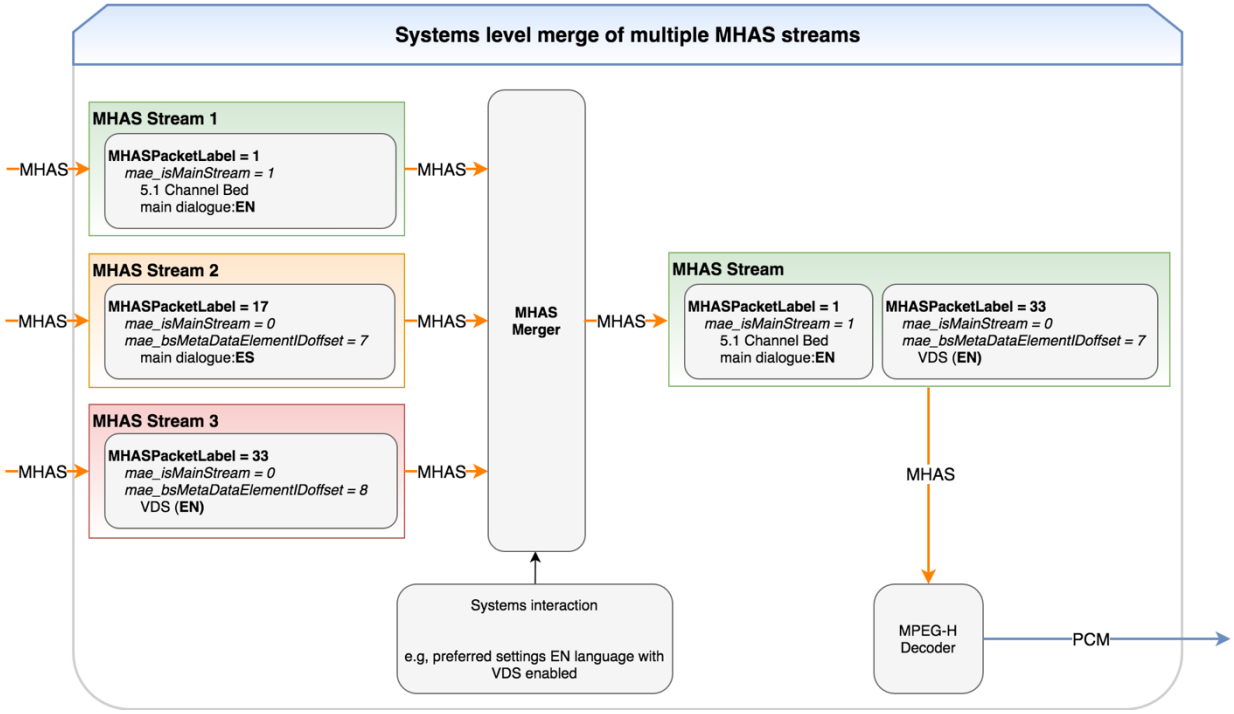



**Figure 7 – Example of selection and merge of multiple MPEG-H Audio streams.**

Assuming a receiver selects, on the systems level, the original language and the VDS service, the second stream can be discarded while the first and third streams can be merged into one single stream that is provided to the decoder. The MHAS packets belonging to different streams are identified based on the *MHASPacketLabel* field and the streams are merged based on the MAE information.

### 3.5. Distributed User Interface Processing

Next Generation Audio systems introduce interactivity options that, compared with existing systems, enable advanced User Interface (UI) features. Furthermore, the user can decide to connect the available devices in various configurations such as connecting a Set-Top Box to a TV over HDMI, or a TV connected to an AVR/Soundbar over HDMI or S/PDIF), while still have the user interface located on the preferred device (i.e., the source device).

For such use cases the MPEG-H TV 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 handled by the sink device. This is enabled by the packetized structure of the MPEG-H Audio Stream, which allows for easy parsing on system level and insertion of new MHAS packets on the fly.

Figure 8 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. The UI Renderer is responsible for handling the user interactivity and passes the information about every user 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 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.

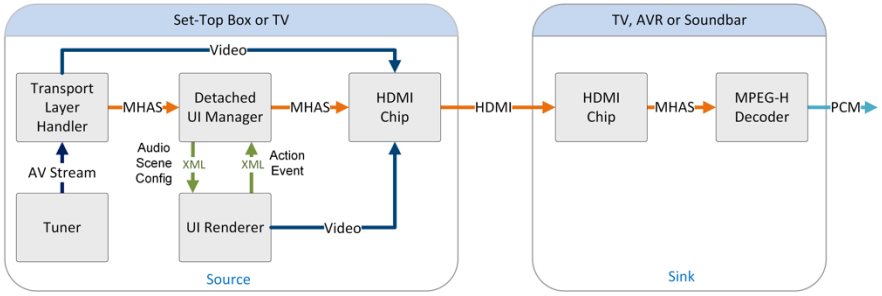



**Figure 8 – Distributed User Interface Processing with transmision of user commands over HDMI.**

## 4. MPEG-H TV Audio Transport Over Cable Networks

Currently, cable content networks delivery systems are based mainly on MPEG-2 Transport Stream (TS). Networks with IP-based delivery use MPEG DASH [12] and ISO Base Media File Format (ISOBMFF) as a container format.

For encapsulation into MPEG-2 TS and ISOBMFF, MPEG-H Audio uses as common intermediate elementary stream format, the MPEG-H Audio Stream, described in the previous section. A description of MHAS delivery over MPEG-2 TS and MPEG DASH is given in the following subsections.

## 4.1. MPEG-2 Transport Stream

### *4.1.1. MHAS Encapsulation into MPEG-2 Transport Stream*

MHAS delivery in MPEG-2 Transport Streams is specified in [13]. MHAS packets are first encapsulated into the payload of Packetized Elementary Stream (PES) packets, which are then encapsulated into Transport Stream (TS) packets. The PES packets consist of an extendable header and a payload. The PES header contains, amongst other information, an 8-bit field indicating the Stream ID, which can be freely set to any value between 0xC0 and 0xDF for all MPEG audio streams, including MPEG-H Audio Streams.

The PES packet payload is created by cutting the MHAS packets into segments at arbitrary byte boundaries, each segment forming a PES packet payload. Therefore, the PES payload may consist of one or more MHAS packets. Moreover, the first and the last MHAS packets in the PES packet do not necessarily need to be complete.

The header may carry information about Presentation Time Stamps (PTS) or Decode Time Stamps (DTS). If present in the PES header, the PTS value refers to the first complete MHAS packet in the PES payload. The PES header at a RAP needs to carry a PTS and, to further facilitate parsing at tune-in, the PES payload at a RAP should start with a complete MHAS packet, more specifically with the complete sequence of MHAS packets for a RAP as described in section 3.3.1. If this is not the case and the PES starts with a fraction of the previous MHAS packet, tune-in is still possible by first searching for the MHAS SYNC packet in the PES payload to identify the start of the MPEG-H Audio RAP.

The second step of encapsulation is to cut every PES packet into pieces that are further encapsulated into TS packets. TS packets have a fixed size of 188 bytes. The header in the TS packet is usually 4 bytes long and has an optional adaptation field that may extend the header. If the last piece of a PES packet is smaller than the payload of the TS packet, stuffing bytes need to be added to this TS packet (more specifically, to its adaptation field in the TS header), because every TS packet can only carry bytes of one PES packet and not of different PES packets. Figure 9 illustrates an example of how MHAS is encapsulated in PES packets and subsequently in TS packets.

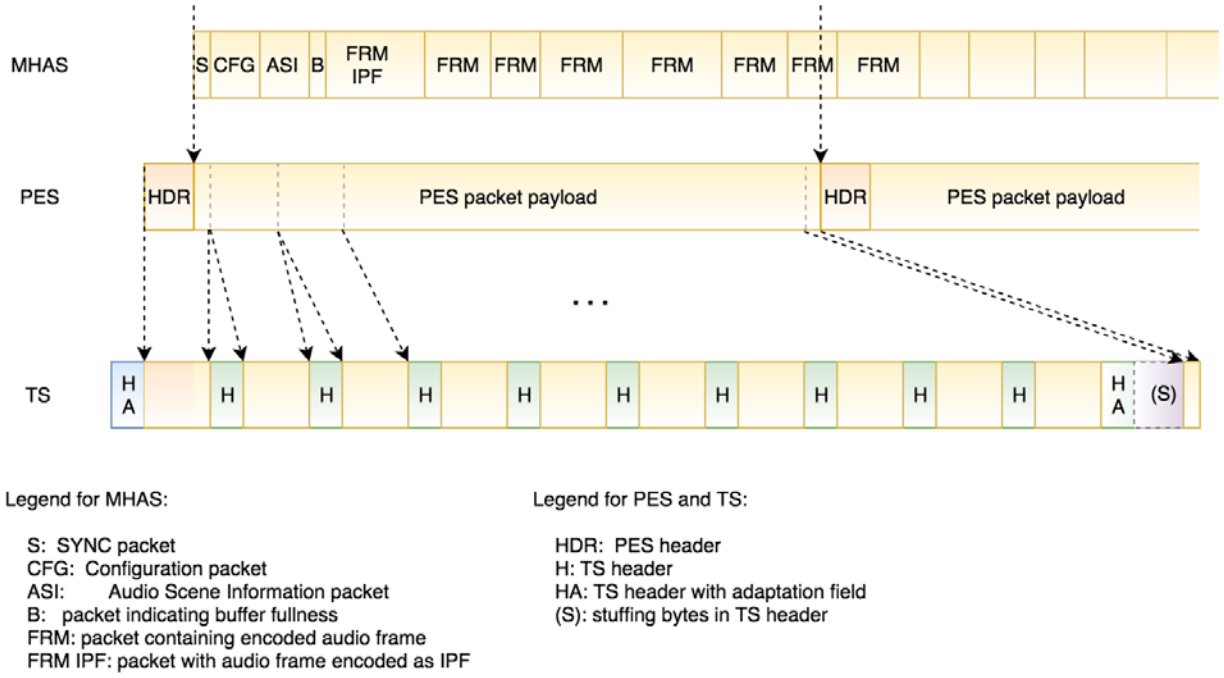



**Figure 9 – Example of MHAS encapsulation in PES and TS.**

MPEG-H Audio RAPs have to be inserted in the audio elementary stream at least once in every 2 seconds, with a minimum distance between two RAPs of 500 ms.

### *4.1.2. MPEG-2 Transport Stream Signaling*

MPEG-H Audio streams are identified in the Program Map Table (PMT) using the *stream_type* values 0x2D or 0x2E. The value 0x2D is used for indicating the main MPEG-H Audio stream, while the value 0x2E indicates auxiliary streams in case of a multi-stream environment.

The *MPEG-H_3dAudio_descriptor()* is specified in [13] and is used for providing the receiver with basic configuration information about the associated MPEG-H elementary stream, such as, codec profile an level indication. The descriptor is located in the PMT of the Program Specific Information (PSI) Tables [14].

NGA features are signaled using an additional descriptor, the *audio_preselection_descriptor()*, specified in [14]. The descriptor uses a more generic terminology for NGA codecs, which has a one-to-one relation to the MPEG-H Audio metadata, as described in [3]. For example, the MPEG-H preset concept is mapped to the Audio Preselection, while the smallest addressable unit on logical level, the MAE group, is mapped to the Audio Program Component.

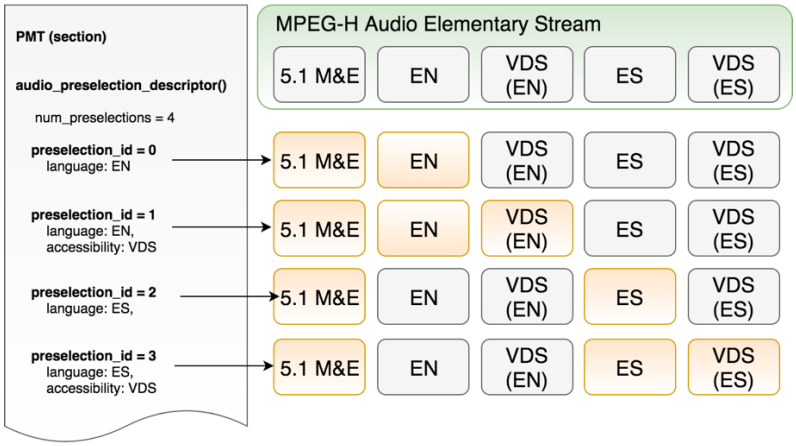



**Figure 10 – Example of Audio Preselection signaling for MPEG-H Audio.**

The descriptor contains information about the number of Audio Preselections and signals the available features for each Audio Preselection, e.g., accessibility features and languages, interactivity options and information about the preferred reproduction layout. This enables the receiving device to perform an early selection of the appropriate Audio Preselection.

Figure 10 shows an example of an MPEG-H Audio elementary stream containing several Audio Program Components (e.g., Music&Effects (M&E), dialog and VDS in English and Spanish languages) and four different Audio Preselections signaled in the *audio_preselection_descriptor().*

For multi-stream use cases, additional MPEG-H Audio streams may be required for one preset. In this case, the *audio_preselection_descriptor()* also provides signaling information for identifing the auxiliary streams. This means that for each preset it contains a list of *component_tag* fields used to identify the required auxiliary streams. Each *component_tag* corresponds to a *stream_identifier_descriptor()* associated with each of the auxiliary streams.

Figure 11 illustrates an example of how the three descriptors mentioned above can be used together in a multi-stream environment. The *audio_preselection_descriptor()* signals four preselections, distributed over three streams. If one of the first two preselections is selected, the main stream (PID 0034) contains all necessary components, while the third preselection requires an additional stream (PID 0035) and the last preselection requires all three streams.

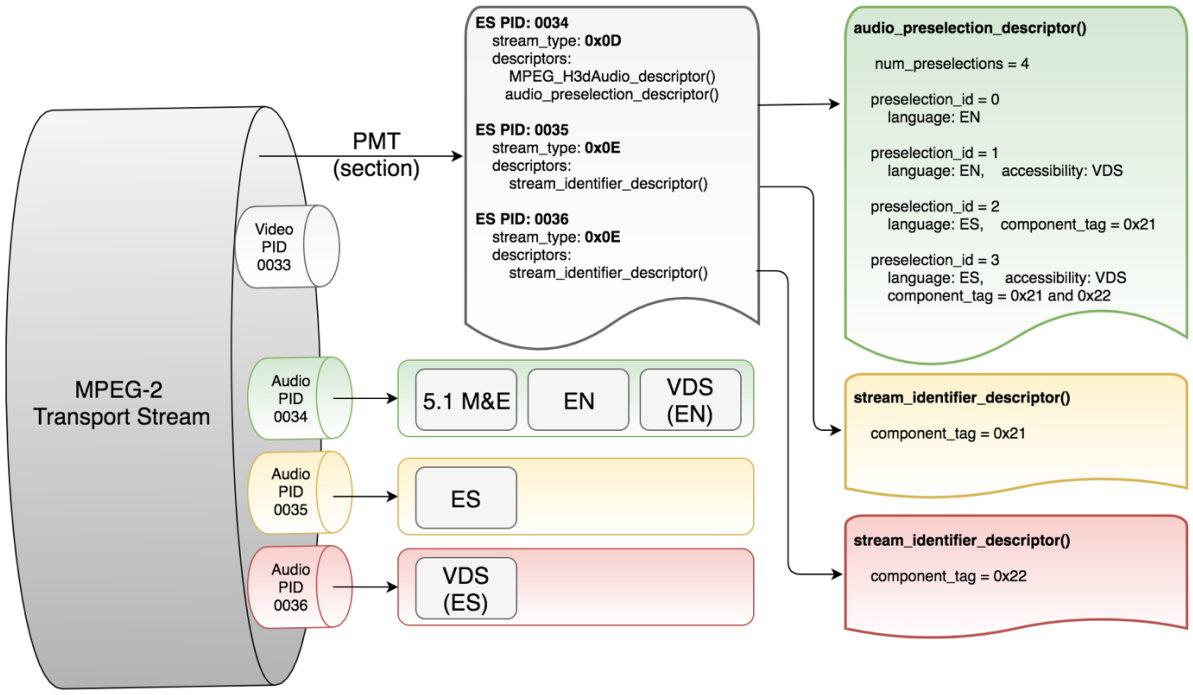



**Figure 11 – Example of MPEG-2 TS signaling for multi-stream MPEG-H Audio.**

ATSC 3.0 specifies in [15] an Emergency Alert System structure, providing several types of emergency information that can be delivered to TV receivers including audio/aural representation of the emergency information. This can be signaled for MPEG-H Audio streams delivered over MPEG-2 TS based systems using the *emergency_information_descriptor()* specified by SCTE/ISBE for NGA codecs.

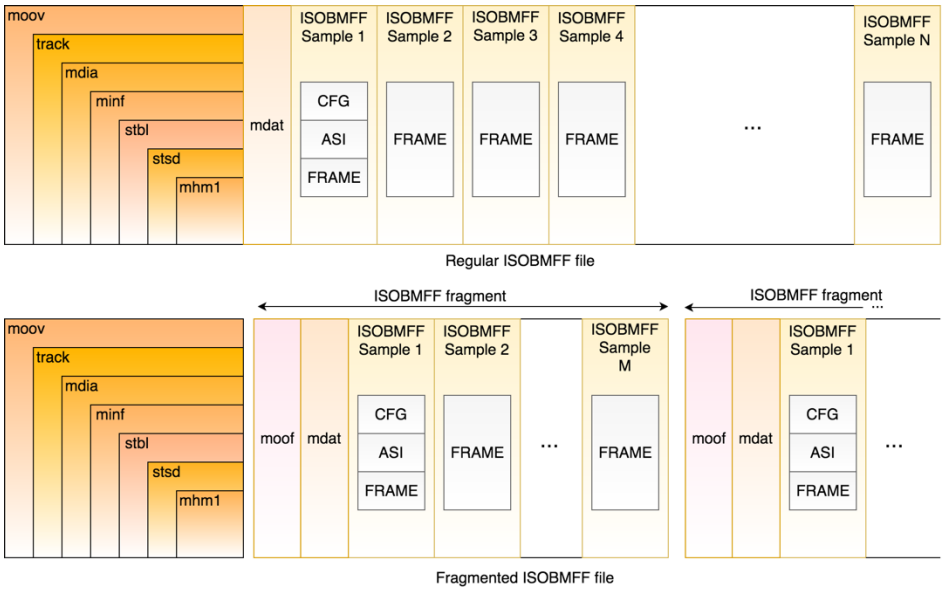
## 4.2. MPEG DASH

SCTE/ISBE has specified a suite of standards describing the usage of MPEG DASH in IP-based cable networks. Although packaging and signalling differs for this use case from MPEG-2 TS based delivery, both use the MPEG-H Audio Stream format (MHAS) as basic elementary stream format and also use the same constraints as defined for ATSC 3.0 [3], ensuring maximum interoperability.

### *4.2.1. MHAS Encapsulation into ISOBMFF*

For DASH delivery, MPEG-H Audio Streams are encapsulated into fragmented ISOBMFF files. Whereas a regular ISOBMFF file usually consists of the movie (moov) box that contains all configuration information (static and sample-related) and one encoded media data (mdat) box that embeds the encoded data of the complete file, a fragmented ISOBMFF file consists of a number of fragments, where each fragment is preceded by a movie fragment (moof) box followed by that fragment's mdat box. In case of a fragmented file the moov box contains only the static configuration information (i.e., the initialization segment), while all sample-related access information is contained in the moof box of each fragment. The data is stored in the mdat box of each fragment and together with its respective moof box comprises the smallest accessible entity [18]. Figure 12 illustrates the difference between a regular ISOBMFF file (upper side) and fragmented ISOBMFF file (lower side).




Fragmented ISOBMFF file

**Figure 12 - A regular MPEG-H ISOBMFF file vs a fragmented MPEG-H ISOBMFF file.**

An ISOBMFF file identifies the encapsulated media for each media track by a sample entry [18] and a respective four-character code (FourCC). The MPEG-H 3D Audio standard defines sample entries for encapsulation of plain Access Units ("mha1" and "mha2" [20]), and for encapsulation of MHAS packets ("mhm1" and "mhm2" [20]). "mha1" or "mhm1" are used for single stream delivery, while "mha2" or "mhm2" are used for multi-stream delivery.

For delivery over MPEG DASH in IP-based cable networks, only the Sample Entries storing MHAS packets in the mdat box (i.e., "mhm1" and "mhm2") are used.

Random Access Points (RAPs), as defined in section 3.3.1, are signalled by setting the "sample_is_non_sync_sample" flag to "0" for the respective ISOBMFF samples in the fragment header (i.e., the moof box). Every movie fragment has to start with a RAP in order to allow tune-in into a broadcast stream or bit rate adaptation in a streaming scenario.

### *4.2.2. MPD Signalling*

Similar to MPEG-2 TS the Preselection concept is used for signaling of NGA features on MPD level and for selection of different configurations offered in several Adaptation Sets. The Preselection element [23] specifies several attributes for providing information about the channel configuration, preferred rendering setup and available interactivity options. Furthermore, it signals the available languages, accessibility services, role or label information.

## 5. Conclusions

MPEG-H Audio has been developed and intensively tested for delivery of Next Generation Audio broadcasting services. Consequently, the MPEG-H TV Audio system has been adopted by DVB and ATSC. The first deployment of ATSC 3.0 for terrestrial Ultra High Definition TV had materialized in South Korea, where the TTA standards organization has adopted the MPEG-H TV Audio system [5].

For distribution of advanced TV services (i.e., UHD TV) over cable networks, SCTE/ISBE has technically finalized the standardization of coding and carriage constraints for NGA systems. MPEG-H Audio has been specified based on the ATSC 3.0 standard, with additional documentation for MPEG-2 TS carriage and signaling. Implementations on the encoder and decoder side as well as for the production of content are available from different vendors.

## 6. Abbreviations

| | |
|---|---|
| ADM | Audio Definition Model |
| ATSC | Advance Television System Committee |
| AU | Access Unit |
| AVR | Audio and Video Receiver |
| CICP | Coding-Independent Code Points |
| DASH | Dynamic Adaptive Streaming over HTTP |
| DE | Dialog Enhancement |
| DRC | Dynamic Range Control |
| DVB | Digital Video Broadcasting |
| DTS | Decode Time Stamp |
| HDMI | High-Definition Multimedia Interface |
| HE-AAC | High Efficiency Advanced Audio Coding |
| HOA | Higher Order Ambisonics |
| IGF | Intelligent Gap Filling |
| IPF | Immediate Playout Frame |
| ISBE | International Society of Broadband Experts |
| ISO | International Organization for Standardization |
| ISOBMFF | ISO Base Media File Format |
| LC | Low Complexity |
| LPD | Linear Predictive Domain |
| MAE | Metadata Audio Elements |
| MCT | Multichannel Coding Tool |
| MHAS | MPEG-H Audio Stream |
| MPEG | Moving Picture Experts Group |
| NGA | Next Generation Audio |
| UHD | Ultra-High Definition |
| UI | User Interface |
| USAC | Unified Speech and Audio Coding |
| PID | Packet Identifier |
| PES | Packetized Elementary Stream |
| PMT | Program Map Table |
| PSI | Program Specific Information |
| PTS | Presentation Time Stamp |
| RAP | Random Access Point |
| TCX | Transform Coding Excitation |
| TS | Transport Stream |
| TTA | Telecommunications Technology Association |
| VDS | Video Descriptive Service |
| SCTE | Society of Cable Telecommunications Engineers |

# 7. Bibliography and References


[1] "Information technology - High efficiency coding and media delivery in heterogeneous environments - Part 3: 3D audio," International Organization for Standardization (ISO), Geneva, Standard ISO/IEC 23008-3:2015, 2015.

[2] "Information technology - High efficiency coding and media delivery in heterogeneous environments - Part 3: 3D Audio Amendment 3: Audio Phase 2," International Organization for Standardization (ISO), Geneva, Standard ISO/IEC 23008-3:2015 Amd3, 2015.

[3] A/342 Part 3: ATSC Standard – MPEG-H System, March 2017.

[4] TS 101 154 v2.3.1: Digital Video Broadcasting (DVB) – Specification for the use of Video and Audio Coding in Broadcasting Applications based on the MPEG-2 Transport Stream. http://www.etsi.org/deliver/etsi_ts/101100_101199/101154/02.03.01_60/ts_101154v020301p.pdf

[5] TTAK-KO-07.0127R1: TTA - Transmission and Reception for Terrestrial UHDTV Broadcasting Service, Revision 1, December 2016. http://www.tta.or.kr/include/Download.jsp?filename=stnfile/TTAK.KO-07.0127_R1.pdf

[6] R. Bleidt et al. "Development of the MPEG-H TV Audio System for ATSC 3.0," in IEEE Transactions on Broadcasting, vol. 63, no. 1, March 2017.

[7] R. Bleidt et al. "Building the World's Most Complex TV Network: A Test Bed for Broadcasting Immersive and Interactive Audio", in SMPTE Motion Imaging Journal, vol. 126, no. 5, pp 26-34, July 2017.

[8] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "MPEG-H Audio - The New Standard for Universal Spatial / 3D Audio Coding," in Audio Engineering Society 137th Convention, Los Angeles, 2014.

[9] ITU R-REP-BS.2159-7-2015: Multichannel sound technology in home and broadcasting applications, 2015. https://www.itu.int/dms_pub/itu-r/opb/rep/R-REP-BS.2159-7-2015-PDF-E.pdf

[10] "Information Technology - MPEG Systems Technologies - Part 8: Coding-independent Code Points," ISO/IEC 23001-8:2016, 2016.

[11] "Audio Definition Model," ITU-R, Recommendation BS.2076, 2015. https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.2076-0-201506-I!!PDF-E.pdf

[12] SCTE 214, MPEG DASH for IP-Based Cable Services, Parts 1, 2 and 3.

[13] "Information technology -- Generic coding of moving pictures and associated audio information -- Part 1: Systems, Amendment 5: Carriage of MPEGH 3D audio over MPEG2 systems," ISO/IEC 13818-1:2015/Amd 5:2016.

[14] ETSI DVB BlueBook A038 (2017-06), Digital Video Broadcasting (DVB); Specification for Service Information (SI) in DVB systems (EN 300 468). https://www.dvb.org/resources/public/standards/a038_dvb_si_spec_june_2017.pdf

[15] A/331: ATSC Proposed Standard Signaling, Delivery, Synchronization, and Error Protection, May 2017. https://www.atsc.org/wp-content/uploads/2016/01/A331S33-174r7-Signaling-Delivery-Sync-FEC.pdf

[16] H. Fuchs, S. Tuff, and C. Bustad, "Dialogue Enhancement - Technology and Experiments," EBU Technology Review, June 2012. [Online]. https://tech.ebu.ch/docs/techreview/trev_2012-Q2_Dialogue-Enhancement_Fuchs.pdf

[17] "Information technology – MPEG audio technologies – Part 4: Dynamic Range Control," International Organization for Standardization (ISO), Geneva, Standard ISO/IEC 23003-4:2015, 2015.

[18] "Information Technology - Coding of Audio-Visual Objects -- Part 12: ISO Base Media File Format," International Organization for Standardization (ISO), Geneva, Standard ISO/IEC 14496-12:2015, 5th edition, 2015.

[19] "Information Technology - Dynamic Adaptive Streaming Over HTTP (DASH) -- Part 1: Media Presentation Description and Segment Formats," International Organization for Standardization (ISO), Geneva, Standard

ISO/IEC 23009-1:2014, 2th edition, 2014.

[20] "Information technology - High efficiency coding and media delivery in heterogeneous environments -- Part 3: 3D Audio Amendment 2: MPEG-H 3D Audio File Format," International Organization for Standardization (ISO), Geneva, Standard ISO/IEC 23008-3:2015/Amd.2:2016, 2016.

[21] DASH Industry Forum. (2017, Feb.) Guidelines for Implementation: DASH-IF Interoperability Point for ATSC 3.0. http://dashif.org/wp-content/uploads/2017/02/DASH-IF-IOP-for-ATSC3-0-v1.0.pdf

[22] "SMPTE Standard - Ancillary Data Packet and Space Formatting," Standard ST291:2006, 2006.

[23] "Information Technology - Dynamic Adaptive Streaming Over HTTP (DASH) -- Part 1: Media Presentation Description and Segment Formats, Amendment 4: Segment Independent SAP Signalling (SISSI), MPD chaining, MPD reset and other extensions" International Organization for Standardization (ISO), Geneva, Standard ISO/IEC 23009-1:2014/DAM 4.

# Forensic Watermarking Momentum Builds for Early Release Windows and Live Sports

A Letter to the Editor prepared for SCTE/ISBE by

Niels Thorwirth, VP, Advanced Technology at Verimatrix

Ali Hodjat, Senior Director Product Management at Verimatrix

# 1. Introduction

A notable trend in video revenue security during 2017 has been strong renewal of interest in forensic watermarking driven by two new use cases, early release window (EWR) movies and live sports streaming. There has been interest ever since watermarking came onto the agenda for protecting HD content over a decade ago, but this had subsided until the arrival of ultra HD (UHD).

The new era for watermarking really began in 2013 when MovieLabs specified the technology for transmission of UHD content in general, whether broadcast or streamed, and this will be an important use case for watermarking if services and available content ramp up over the next few years. Yet, since that move by MovieLabs, watermarking has been driven more by growth in live sports streaming and then, most recently, the imminent reduction in ERW.

The appeal of watermarking lies in its scope for tracing individual sources of content, whether that be from piracy through camcording of a movie or streams of illicit device transmission over the Internet. Indeed, no other technology has yet been conceived that is capable of effectively identifying sources of audiovisual content infringement.

# 2. Tradeoffs

One significant factor now is that watermarking technology has matured to the point where it can cater to the different use cases now calling for it, enabling varying tradeoffs to be made between the facets of the marks themselves. The key features of the marks are robustness against transformation, security against attack, imperceptibility to the viewer, scalability to large numbers of users or volumes, and ease of extraction from the audio or video. Greater robustness, for example, may also make the marks more visible or less imperceptible. There is also a distinction over implementation in terms of how and where marks are embedded into the content, which can be done on the client or server side and via a one-step or two-step process.

The growing importance of watermarking for premium UHD content has now been recognized by the Ultra HD Forum, which included the technology in updated guidelines published in April 2017. Over the last year or two, leading video security vendors have come up with different watermarking schemes to support diverging use cases, varying in their tradeoffs and marking processes. The use cases can be distinguished in several fundamental ways, between live and on demand or between broadcast and unicast streams, as well as the distribution of physical media. There is also a distinction by quality, such as UHD/4K resolution and or high dynamic range (HDR) versus HD and by the release window for the content.

Live sports have distinct watermarking needs that reflect the short life cycle of content of which value is compressed into a time frame confined to the duration of the event itself. The value has plummeted by the time the event is over, which also means that by then, any piracy has already succeeded in extracting revenue. The value of retrospective actions against pirates is limited, with the onus on content owners and distributors not just to detect violations quickly but also act upon them within a matter of minutes, rather than hours or days. Robustness against transformation and mark longevity are less important, while the ability to extract marks, identify sources and then act fast by taking infringing streams down is critical.

When it is the quality that is being protected, as in UHD services, it is clearly vital that the marks be absolutely imperceptible to the viewer even on large screens. However, robustness against transformation

is less important because if the video is downgraded to a lower resolution then by definition it is no longer UHD content and so no longer needs protecting under that service category.

By contrast, early release movies will be issued at an intermediate quality, with the aim of creating a second wave of engagement around UHD or HDR after subsequent full release. Robustness of the watermarking against downgrading or other forms of transformation is important for ERW movies since the marks must survive in the event of retransmission in different formats and plausible attacks, including camcording.

## 3. Changing Landscape

The major studios have long wanted to reduce the window between theatrical release and availability online, DVD and Blu-ray from the long-standing three months on the grounds that this would open up new revenue to compensate for a gradual decline at the box office over recent years. However, when serious discussions last took place around 2011, the effort foundered on strong opposition from the cinemas, which feared major loss of revenue.

However, several factors have changed since then, one being the rise of subscription video-on-demand (VoD) providers, especially Netflix and Amazon, which have become major funders of content production, which they then release simultaneously to theater and TV, or even just the latter. Another factor is that while in 2011 studios were focused on distribution in a highly secure walled garden environment to expensive set-top boxes, today they are seeking to reach as many target devices as possible from smart phones to tablets to consoles and cast dongles. The objective is to milk that second window to its maximum potential before it closes. This has led to the third and most conclusive factor, which is that in order to obtain this window, the studios are now prepared to share some of the associated revenue with the theatres as compensation for the lost audience to online.

So while there are still arguments over how much the window will be slashed, there is a consensus now that it will happen across the board for all Hollywood produced movies. The upshot is that whatever the window will be, it will drive demand for forensic watermarking to ensure adequate protection for the content when distributed to all these online devices. This content will not be UHD, but it will be widely distributed on demand as unicast streams, which brings its own particular requirements for server-side watermarking now being worked on by the major security vendors.

## 4. Selecting the Right Watermarking Approach

All watermarking systems are specific to the user or client receiving either a stream or multicast/broadcast transmission, given that the objective is to trace illegal redistribution back to its source. So whether the marks are embedded in the client or server, the solution is session-based and specific to the user. In all cases the embedded marks must be unique and as resistant against both transformation and attack as possible.

Under client-side watermarking, the marks can be embedded securely in decompressed video during playback in the set-top box or viewing device, within the secure video path. This avoids the need for any content pre-processing at the head-end and is now pre-integrated into around ten of the leading system-on-chip platforms, which means that the watermarking process takes place in hardware with virtually zero latency or overhead in processing or storage. As such, client-side watermarking can be used for live streaming for which the latency budget is very tight.

It is also the ideal solution for broadcast and multicast applications that share the transmission path from head-end to client between multiple users so that it cannot uniquely identify a given source of illicit redistribution. In that case, only by embedding the marks at the client end can a subsequent retransmission be pinpointed to its source, whether it involves camcording the screen or directly capturing streams using readily available tools that circumvent the protection provided by high-bandwidth digital content protection (HDCP) over HDMI interfaces to TVs.

Apart from seamless support for broadcast content and suitability for live streaming, client-side watermarking is independent of the codec or containers because the embedding occurs after decoding. Yet for that reason, it is important that access to this clear content is protected, for example, by being integrated tightly into a secure video pipeline, as well as with the decryption and digital rights management system.

One disadvantage of client-side watermarking is that it only works with clients capable of embedding marks. Partly for this reason, server-side watermarking has been proposed instead for unicast on-demand services delivering content to legacy devices. Server-side watermarking may also require integration with encoders, as well as content packagers or CDNs prior to deployment so that marks are embedded in encoded video.

The big advantage of server-side watermarking is that it can work with any device, and as a result, it is a perfect match for the impending ERW, given the desire to reach as many viewing clients as possible. It is worth noting though that if and when UHD content becomes available through ERW, this advantage fades away because only newer devices will be capable of playback at the higher quality and these will support client-side watermarking anyway.

## 5. One-Step Vs Two-Step Processes

The other major distinction between watermarking systems lies in whether the marks are embedded in one or two distinct steps. While one-step watermarking is often used for client-side embedding and two-step with server-side, each can be combined with either depending on the requirements for scale and processing locations.

One-step watermarking is often deployed with client-side watermarking because unique marks are then embedded in the baseband video as a single step during decode and playback in the client. However, the access to baseband video and watermark embedding can also be integrated with the encoder on the server-side, although, since this would require a separate encode for every watermarked file, it is not scalable beyond a small number of uniquely marked copies. It is therefore only suitable for limited use cases.

In contrast, two-step watermarking has been used mostly with server-side watermarking, in which the second step is used to prepare a uniquely marked file for delivery to clients that lack watermarking capabilities. This exploits the first step, which has already pre-computed variants of the watermark for application to different parts of the video asset. The second step then assembles unique combinations of these variants for embedding into each unicast stream.

The reason for separating the process into two steps is that it isolates computationally intensive processes in the first step so that the second embedding step can be as light and fast as possible. It is this separation that enables the second step to be distributed if desired to clients where processing resources are limited. This could be useful when content is delivered in broadcast channels or physical media that is the same

for all recipients and a secured client player would be able to embed the marks with the appropriate software without undue processing overhead.

For ERW streaming, however, two-step watermarking can be deployed entirely at the server side. A key development here has been enhancement of the process to work with adaptive bit rate (ABR) streaming, including both MPEG-DASH and Apple HLS. The embedding is executed during delivery in such a way that each will receive a uniquely marked stream. Successive ABR segments would be assembled to yield a unique binary code identifying each stream.

This requires integration in an eco-system that may involve encoder, packager and content delivery to maintain efficiency, scale and codec independence, but the industry is working on solutions to facilitate and unify workflow and interfaces between different vendors.

One important distinction between one- and two-step watermarking lies in the frequency and size of the marks that work best with the process. One-step systems apply small modifications to a lot of frames, which make them ideal for UHD services for which high imperceptibility is essential. Two-step watermarking tends to make larger modifications to fewer frames, which allows chaining of modifications in order to individualize the embedded information.

One further important development with respect to streaming is support for the new Common Media Access Framework (CMAF)) standard being introduced to unify at last the world of ABR, bringing Apple and supporters of DASH together behind a common format. Although there are some issues to be resolved regarding having two different encryption schemes respectively for DASH and HLS, full support for CMAF within server-side watermarking is an important goal that will help bring about secure services that enable the ERW.

# Ultra HD Forum Promotes New Guidelines for Forensic Watermarking to Enable the Release of Premium UHD Content

A Letter to the Editor prepared for SCTE/ISBE by

Laurent Piron
Principal Solution Architect
Chairman of the Ultra HD Forum Security Working Group
NAGRA
Route de Genève 22-24
1033 Cheseaux
Switzerland
laurent.piron@nagra.com
+41 21 732 05 37

# 1. Introduction

Forensic watermarking is now being required by holders of premium rights to movies and increasingly also live sports. This trend has been gathering force ever since MovieLabs announced it was recommending forensic watermarking be mandated for content distributed by its member studios in Ultra HD formats in April 2014. This now means content created at High Dynamic Range (HDR), Wide Color Gamut (WCG) and High Frame Rate (HFR) as well as 2160p "4K" resolutions. Forensic watermarking will be required for premium content with these attributes.

Furthermore, although the Ultra HD Forum is naturally focused on the security of UHD content, it recognizes that forensic watermarking will also be required for early release window movies, as well as to provide a new line of defense against illicit redistribution over the Internet in general as this is becoming an easy means for accessing content [1]. Indeed, the Forum has made sure that its guidelines for deployment of forensic watermarking are equally applicable to all content, irrespective of the format.

# 2. Ultra HD Forum Security Working Group Makes Forensic Watermarking a Priority

The Ultra HD Forum Security Working Group (WG) has dedicated its first efforts to forensic watermarking because the technology is still at a formative stage and yet, at the same time, is now an absolute requirement for UHD content from the studios, which will not release the most premium content without that protection. This is because the main piracy threat in the Internet era is illicit content redistribution of content using various IP-based technologies, which can only be dealt with by identifying individual streams and tracing them back to their source. This in turn requires insertion of some unique identifier at the source or during distribution. A watermark can be designed to be tamper resistant and at the same time to have minimal impact on the video or audio quality, although there is a balance to be struck between robustness, performance and transparency to the user.

# 3. Forensic Wartermarking is an Absolute Requirement for Premium Content

The Ultra HD Forum (Forum) has identified strong demand from the infrastructure community that it represents, not just for information about use cases, applications and the technology itself, but also for guidance on deployment and integration. Technology providers are under pressure to incorporate forensic watermarking in their products from the ground up because the MovieLabs mandate has prompted a change in mindset towards security in general so that it is taken seriously again and no longer bolted on almost as an afterthought. Now it should be one of the first considerations at the specification and design stages.

This immediately means that the different watermarking systems should be compatible as much as possible and interoperate within delivery ecosystems. That is why the Ultra HD Forum has defined guidelines that specify not just a common vocabulary and systems architecture, but also key integration points, including the encoder, CDN and client device, where marks can be embedded. The guidelines also describe requirements on the ecosystem side, again with a view to ensuring that forensic watermarking is catered for from the beginning.

[1] https://torrentfreak.com/millions-of-north-american-households-use-kodi-with-pirate-add-ons-170504/

The Forum is fortunate to have many of the principal forensic watermarking technology providers as members: Content Armor, NexGuard (acquired by NAGRA in 2016), Verimatrix and most recently, Irdeto. The guidelines therefore represent their collective wisdom and cater for all possible use cases and deployment scenarios. At the same time these vendors can ensure their products are aligned with security requirements for Ultra HD content.

The focus of the current guidelines is very much on deploying forensic watermarking now, but it is still a work in progress with more to be done over the coming months. The WG has had interesting and valuable feedback from various parties, especially from streaming providers as well as content creators such as Sony Pictures, and some of that will be incorporated in the next version of the guidelines. That will make it even easier to integrate forensic watermarking through a unique interface, which currently is still being worked out with the different technology providers.

## 4. Conclusion

The underlying point is that OTT and IP Video distribution has changed the security game, presenting new threats and making unauthorized content available in many different places for easy access by users, many of whom are normally law abiding but still happily consume pirated material. UHD ups the ante by increasing the value of the content but also by making it easier to redistribute at high quality over the Internet. Existing content protection mechanisms can readily be bypassed, possibly by direct camcording from a screen, although a greater threat is through direct capture of streams using illegal tools that circumvent the protection provided by HDCP over the HDMI interfaces to TVs.

In this context, forensic watermarking has become a critical component that must be integrated within the ecosystem to ensure end-to-end content and revenue protection. The Ultra HD Forum guidelines are designed to facilitate and accelerate this integration and are available on www.ultrahdforum.org starting April 24th.

*(Earlier revision of this was published by VideoNet.)*

SCTE · ISBE

Society of Cable Telecommunications Engineers, Inc.
International Society of Broadband Experts™
140 Philips Road, Exton, PA 19341-1318
T: 800-542-5040 F: 610-884-7237

www.scte.org | www.isbe.org