METADATA DRIVEN GRID ENCODING

Dan Holden
Comcast Media Center

Abstract

This paper addresses a new transport methodology that will reduce bandwidth consumption on a cable plant. Cable faces a unique challenge — how to support heritage set top boxes (STB) concurrently with new Consumer Electronic (CE) devices that have exponentially greater capabilities which include: increased resolutions, frame rates, 3D, interactive applications, two-way/IP protocols, and the ability to reach beyond a traditional cable plant. The cable industry must find a way to maximize the video viewing experience, while minimizing the bandwidth and storage requirements associated with compressed video streams. The complexities associated with advanced services are daunting. Numerous standards try to address unique opportunities, but the industry currently lacks a unified methodology to associate all of these disparaging technologies into a single, integrated delivery mechanism. An approach to resolve this issue is dividing video, audio, and associated data into a grid. Rows are responsible for horizontal partitioning of enhanced video resolution, 3D, applications, data, subtitling, and audio tracks. Vertical partitioning through columns will allow for packetization along Group of Picture (GOP) boundaries, which will service IP streaming, ad insertion, start over, and other container-based services.

Current techniques for addressing the increased capabilities of STBs include simulcast for linear content and multiple file-based encodes for VOD assets. For 3D content, side-by-side and over/under frame compatible formats are being adopted, which will substantially increase storage requirement in the VOD plant. In addition, these 3D formats will require simulcast techniques when new formats such as 1080p60 are introduced into linear broadcast. Horizontal segmentation of compressed video, in conjunction with an enhanced transport container, will provide a new method to deliver 3D in a scalable fashion. A single video package will be able to contain all information required to produce multiple formats, e.g. 720p, 1080p24, 1080p30, 1080p60, 3D 1080p24, 3D 1080p30, and 3D 1080p60.

This delivery mechanism will allow for scalable video delivery that can expand with CE device capability. Allowing CE devices to make two-way requests of specific containers utilizing metadata will reduce storage costs and bandwidth consumption on the local loop. This generalized approach will significantly reduce VOD storage requirements. For linear broadcast, PID filtering and video processing at the STB will eliminate the need for simulcast.

INTRODUCTION

All Multiple Systems Operators (MSO) face similar challenges of how to implement advanced and re-usable business intelligence on heritage Consumer Premise Equipment (CPE). In the past three years, it has not been feasible for MSOs to move from MPEG-2 to H.264. EBIF (Enhanced TV Binary Interchange Format) was selected over tru2way for initial deployment at Comcast for one simple reason – there are tens of millions of deployed DCT2000s in the home. Implementation of 1080p will likely happen on MPEG-2 before H.264. The first release of 3D video will be using

MPEG-2. In order to stay competitive a paradigm shift in video encoding must take place. Video compression must not only save storage and transport bits, but there must also be a clear roadmap between existing deployed technology and the next generation of video display devices. Innovation has consistently been the key to cable success. The foundation of maintainable technology is based upon a clear upgradeable and supportable roadmap. Incompatibilities between core infrastructure and edge devices prevent the deployment of new technologies when they are available. By introducing an abstraction layer and proven Information Technology (IT) techniques, it will be possible to offer features in the home without the need to replace legacy devices. The future of encoding needs to target a system that allows for loose coupling of the encoder and decoder, and ensures all audio, video, and data is tightly bound. The proposed grid approach will allow compression experts to continue to do what they do best; save bits on the plant. This will allow other teams of experts to extend the functionality of CE devices in the home.

The cost to store and processes bits at the edge of a Hybrid fiber-coaxial HFC network is exponentially more expensive than in a super headend. Given this equation, STB tend to be basic devices with limited storage and processing power. Truck rolls for edge-device replacements are normally the option of last resort for MSOs because of the high cost and the number of affected households, which can be in the millions. For this simple reason, heritage devices tend to live on the network beyond their engineered life span. As a result, MSOs tend to deploy technology that meets the requirements of the simplest device on the network (currently a DCT2000). In order to stay competitive, a methodology must be adopted that allows new, innovative technology to exist concurrently on a shared network with heritage CE devices in the home. Like your older brother's hand-me-down clothes, old TVs and other CE devices never make it to the dumpster, they migrate from the living room to the bedroom, which has a dramatic effect on the network. The extensible nature of grid encoding will allow this bedroom TV to continue to generate revenue far beyond its life expectancy, without impacting innovation.

Grid encoding is a new science that will be deployed in the super headend in order to reduce the number of bits transported and stored at the edge of the network. It will reside between the encoders and decoders, and should not require significant changes to current encoding specifications or technologies. Video pumps and CE devices will need to be built in a fashion that will leverage the two-way infrastructure of the modern HFC plant. This cutting edge technology has the ability to extend video, audio, and interactive TV specifications; and most importantly, it will extend the life of CE devices in the home; therefore, reducing the operational cost and maintenance.

ENCODING

Blob Encoding

Current encoding technology relies on encoding assets to a single "blob." A blob encompasses video, audio, and other Packet IDs (PID) such as interactive TV applications that have been compressed and identified by metadata. If we examine a typical asset encoded using this methodology it might contain the following representative PIDs.

Table 1

PID Type	Values
Video	720p60, 1080p24, 1080p30, 1080p60, 1080i, Flash, WM9
Audio	AC3, AAC
Data	EBIF, tru2way, Packet Cable

Using the sample data in Table 1, it is possible to calculate the number of blob assets that would be generated by combining the different PIDs. A simple combination, without repetition can be expressed as:

number of blobs = $n! / [(n-r)!\ r!] = 13!/[(13-3)!*3!] =$
$(13 \text{ choose } 3)$
$= 286$ Blobs

If we were to take these 13 unique PIDs three at a time, we would derive 286 separate assets. Each asset would effectively be the same movie just implementing different compression or interactive TV technologies. Our single movie when encoded using blob technology would require unique encodings or packaging, in order to support seven different video 'coder-decoder' CODECs, two audio CODECs and three interactive TV data PIDs. Using this primitive encoding technique for 100,000 movies would result in 286,000,000 assets that would need to be managed, distributed, and streamed individually.

GRID ENCODING

Unlike current blob encoding technologies which do not leverage the capabilities of a two-way network infrastructure, grid encoding seeks to move the complexity of video transport off the CE device to the network. Old encoding technologies singularly focus on reducing the number of bits required for transport and storage. Now the focus shifts to solving issues surrounding migration paths from heritage technologies (MPEG-2, 1080i, 2DTV video) to 'better' technologies (H.264, 1080p, 3DTV). Significant improvements can be achieved by leveraging typical TCP/IP communications and information technology (IT) topologies. When a CPE device is granted the ability to announce itself on the network and broadcast its capabilities, it is possible to generate significant changes to the encoding specifications. It will effectively be able to tell a VOD system the decoder capabilities, and the VOD system will be able to respond to a request with a tailor made video encoding package.

Encode

Encoding is the process of creating a self-synchronizing stream of signals against a known timeline. For example, we will assume encoding with MPEG-2. The process will work equally well for VC-1, H.264, Flash, or any other CODECs that create horizontal separation of the data. In order to keep the example simple we will use “Theatrical Release” as the title of a 3D movie . For audio we will choose AC3 in order to ensure the encoded asset is compatible with currently deployed STB. A representative output stream from an encoder is depicted in Figure 3. This elementary stream will now be passed to the transcoding step of our process.

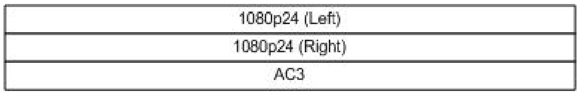


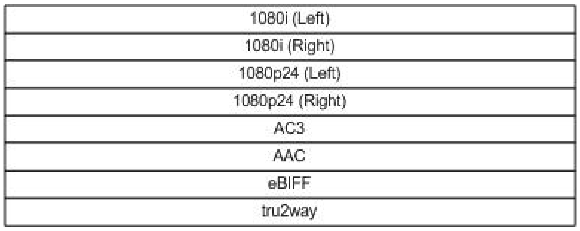
Figure 1

Transcode

During the transcoding process multiple CODECs, resolutions, and/or bit rates will be generated. This process will enhance the movie so that it can play on multiple CE devices at multiple bit rates. Support for multiple audio CODECs will also be added in the transcoding process. This process will prepare the asset for fragmentation at various bit rates.

iTV Striping

For this example the enhancement layers will carry EBIF and tru2way applications. These applications are carried in Package Identifiers (PID) which is added through a process called iTV striping. These data PIDs are bound to the video and provide new functionality on the CPE device. They add interactive features to the video stream such as voting-and-polling applications, advanced advertising, and other enhanced features. A representative output is depicted in Figure 2.

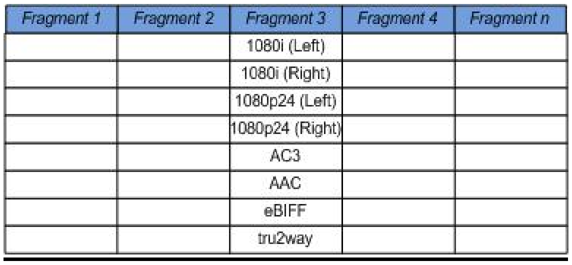


1080i (Left)
1080i (Right)
1080p24 (Left)
1080p24 (Right)
AC3
AAC
eBIFF
tru2way

Figure 2

Fragmenting

Fragmentation is the process used to separate the transport stream vertically in order to prepare the video for loading into the grid. Packetizing the video stream into PIDs is an effective means for separating the audio, video, and data into fixed or variable, size units. This process of placing the packets into fixed or variable duration units will be utilized for adaptive streaming. The number of frames loaded into each cell of the grid does not have to be consistent, as the timeline will be maintained in the client buffer. Breaking the video along the GOP Boundaries will help facilitate ad insertion. Each fragment will require a unique identifier to facilitate the loading of the grid and orderly retrieval of video, audio, and data for placement into the decode buffer. Multiple bit rates for the video PIDs will be provided in order to support adaptive streaming technology. This functionality is not expressed in Figure 3 for the purpose of simplicity.



Fragment 1	Fragment 2	Fragment 3	Fragment 4	Fragment n
		1080i (Left)		
		1080i (Right)		
		1080p24 (Left)		
		1080p24 (Right)		
		AC3		
		AAC		
		eBIFF		
		tru2way		

Figure 3

GRID

Let's start by examining a single fragment that has been produced by the fragmentation process (Figure 6.) It may contain multiple video CODECs. In reference to the "Theatrical Release" movie example, it contains two fragments with the same sequence of video frames. The left eye 1080i and the left eye 1080p24 cells are by definition not equivalent, even if they contain the same footage. They may contain the same number of frames and their time code must be synchronized with the audio cells and iTV cells. The grid represents a new type of structured video. Rather than thinking of the grid as fragmented video, it will be treated as a simple data structure that has been partitioned and described with metadata for optimized storing and retrieval of video objects. Each video cell in the grid is effectively equivalent. The timeline will be maintained in the buffer. It is natural to think of this type of data structure as an enhanced database that has been optimized for storing and retrieval of video, audio, and other objects of interest to a typical CE device.

Load the Grid

Examination of a single fragment (Figure 4) reveals multiple cells with video, audio and data. These cells can be loaded into the grid one column at a time. The first column of the grid represents time 0 on the timeline or the beginning of our sample movie title. Each cell of the grid is loaded until the last column is complete.

On video ingest, audio, and other components of the video stream are parsed and loaded into appropriate cells. These cells are then accessed by the VOD pump using Structured Video Query Language (SvQL) at the request of a CPE device.

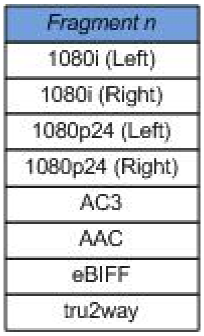
Query the Grid

After the data has been placed into the grid, it is possible to extract the data using SvQL. Below is an example that could be used to represent a 3D movie request from an STB:

SELECT * FROM movies WHERE
movie_title='Theatrical Release' and
left_eye_video=1080p24 (Left) and
right_eye_video=1080p24 (Right) AND
audio=AC3 and iTV=EBIF

In addition, this sample query could be used to retrieve a 2D movie on a Personal Computer without interactive features:

SELECT * FROM movies WHERE
movie_title='Theatrical Release' and
left_eye_video=1080p24 (Left) and
right_eye_video=NULL AND audio=AAC

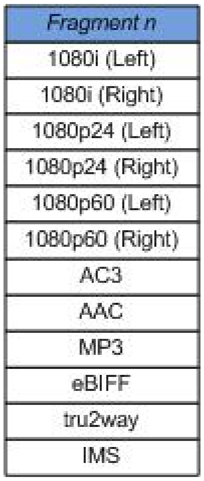


Fragment n
1080i (Left)
1080i (Right)
1080p24 (Left)
1080p24 (Right)
AC3
AAC
eBIFF
tru2way

Figure 4

Extend the Grid

Extending the grid is a straightforward process that can be achieved by simply adding additional rows to video before fragmentation. These rows will represent new columns in the grid. As a sample, 1080p video at 60 frames per second will be added. Additionally, IP Multimedia Subsystem (IMS) and MP3 audio will be added to the CPE device. The features will be added to the encoding, transcoding, and fragmentation farms. A representative fragment is shown in Figure 5.



Fragment n
1080i (Left)
1080i (Right)
1080p24 (Left)
1080p24 (Right)
1080p60 (Left)
1080p60 (Right)
AC3
AAC
MP3
eBIFF
tru2way
IMS

Figure 5

Extending the fragment has no effect on retrieving the data. The SvQL statements on legacy devices will not query the new, extended rows of the grid. It is possible to update the SvQL on legacy devices in order to extend their respective life and functionality. This type of extension will ensure CPE devices will expose the greatest amount of functionality and therefore will not lock the MSO into legacy technology.

Video On Demand (VOD)

Grid encoding for VOD leverages the inherent capabilities of a two-way infrastructure. Beginning at the left of Figure 6, the video is transformed, segmented, and loaded into the grid. From the right of the figure, the CPE device identifies itself and capabilities on the network, and requests only the cells of data that it knows how to process. The end-to-end VOD workflow loads the grid and exposes the data to the home.

VOD systems require significant innovation, and changes would be required to the CPE devices in order to take advantage of the new technology.

Late binding or polymorphism allows our CE devices to attach or request a stream that matches the exact features and hardware of the device.

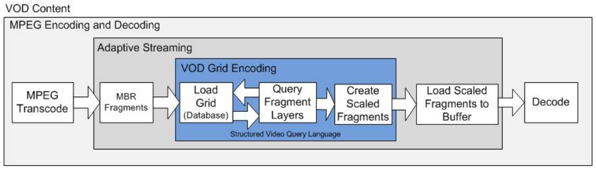


Figure 6

Linear

Linear content is not loaded into a static grid for a future query; rather the data is processed real-time and loaded into a "topic grid," which does not send video to a specific receiver. The published video is characterized into specific types of video without knowledge of which device will consume the video. This queue will live on the CPE device or another network location that is localized to the home. Subscribers then expresses interest in specific types of video (1080p24, AC3, EBIF) and receives only the video of interest, without knowledge of what, if any, video publishers that exist. This technique allows the live, streamed data to be loosely coupled and tightly bound to our encoding technology. All data retrieved from the grid is re-assembled as fragments and loaded into the buffer of the CPE device for decoding. Figure 7 represents end-to-end flow for linear content.

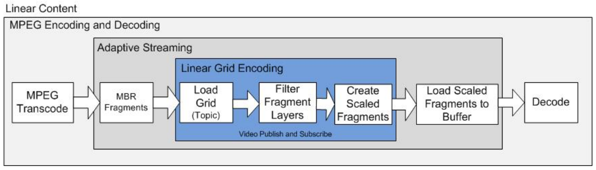


Figure 7

METADATA

Metadata is the key component to drive the entire encoding and decoding process. It is the glue that brings the architecture together. Original content is described by metadata. It is used to drive the orchestrations to encode, transcode, fragment and load the grid. All business logic is encompassed in metadata such as XML. The CE device utilizes metadata request relevant cells and the entire grid may be persisted on a device such as a DVR for future playback. Typical Electronic Program Guide (EPG) is another form of metadata that can be used to help parse the video and load it into the grid. PIDs use metadata to characterize their value to the ecosystem. Metadata is used to describe the grid and is integral in retrieving the correct components in the grid. The grid is self-describing metadata comprised of rows and columns similar to a database.

While SvQL may be used for VOD applications, linear content is in flight. A message bus should be utilized due to the data latency associated with database applications. In order to subscribe to the message bus, the following type of Java Message System (JMS) XML metadata could be utilized:

Sample Message Bean XML:

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<tv-ejb-jar
xmlns="http://www.objectweb.org/tv/ns"

xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
```

```
xsi:schemaLocation="http://www.objectweb.
org/tv/ns
      http://www.objectweb.org/tv/ns/tv-ejb-
jar_4_0.xsd" >
  <tv-entity>
    <ejb-name>VersusChannel</ejb-name>
    <jndi-name>VersusChannelHome</jndi-
name>
    <jndi-local-
name>ExampleTwoLocalHome</jndi-local-
name>
    <jdbc-mapping>
      <jndi-name>jdbc_1</jndi-name>
      <jdbc-table-name>MoviesTable</jdbc-
table-name>
      <cmp-field-jdbc-mapping>
        <field-name>MovieTitle</field-name>
        <jdbc-field-name>dbMovieTitle</jdbc-
field-name>
      </cmp-field-jdbc-mapping>
      <cmp-field-jdbc-mapping>
        <field-name>VideoCODEC</field-
name>
        <jdbc-field-
name>dbVideoCODEC</jdbc-field-name>
      </cmp-field-jdbc-mapping>
      <cmp-field-jdbc-mapping>
        <field-name>AudioCODEC</field-
name>
        <jdbc-field-
name>dbAudioCODEC</jdbc-field-name>
      </cmp-field-jdbc-mapping>
      <finder-method-jdbc-mapping>
        <tv-method>
          <method-
name>findByMovieTitle</method-name>
        </tv-method>
        <jdbc-where-clause>where
dbMovieTitle = ’Theatrical Release’</jdbc-
where-clause>
      </finder-method-jdbc-mapping>
    </jdbc-mapping>
  </tv-entity>
</tv-ejb-jar>
```

COMPARISON TO SCALEABLE VIDEO ENCODING (SVC) AND MULTIVIEW CODING (MVC)

SVC suffers from one inherent flaw: it adds significant complexity to the edge of the network, the very edge that is the most complex to maintain, upgrade, and support. The cost of a single byte of memory in a super headend is minimal. As this byte is propagated into the network the cost soars exponentially. For example, if the cost was $20 per GB, the memory would cost $20 if deployed in the super headend, or nearly $500MM if deployed at the edge. As indicated in table 2, the aggregate cost for installing a GB of memory in every set top box in a major market cable system could exceed $490 million, assuming a cost of $240 per set top box in a market supporting more than 200,000 set top boxes.

Table 2

Location	Cost for 1 GB RAM
Super Head	$20
Application Point of Presence (APOP)	$240
Headend	$3,000
Set Top Box	$ 490,000,000

The same calculations can be made for CPU, software support, and other technologies deployed in the STB. Moving complexity upstream saves capital expenditure and future support expenses. SVC video topology leads to significant cost increases for the operator. However, it does offer several advantages, for instance; SVC is a more effective approach than simulcast for reducing the bandwidth associated with the delivery of multiple CODECs. From a VOD perspective, it may be advantageous to use SVC to the video pump then to convert the stream at the pump to the specific CODEC requested by the CE device, which would be close to the grid encoding technique described earlier.

Multi-view coding (MVC) is an attempt to address encoding requirements for 3D and multiple camera views. Again, this technology is targeted at linear broadcasts and has limited application in an On Demand environment. While the cost of a general CE device may be relatively cheap, the cost to support and deploy the devices can be astronomical. MVC as a standalone technology may bring value to 3D content; it is certainly a better compression technology than the current side-by-side frame compatible approach that will be deployed in first generation 3DTV broadcasts.

Grid encoding can be used to enhance various types of encoding, including SVC and MVC. It may also be used as a replacement for both SVC and MVC encoding standards. SVC will utilize fewer bits to represent the video, but it does not currently have the ability to support adaptive steaming technology. SVC binds the encoder directly to the decoder which is a disadvantage. It will add complexity to the decoding process at the CPE device than grid encoding, which will drive up the cost of the CE device.

Grid encoding moves the complexity of decoding from CPE to the network, which should provide better architecture for scaling. Grid encoding is better suited for streaming On Demand assets as each CPE device has the ability to request only the information that is relevant to the device. VOD pumps are in a better position to handle the complexity of multiple CODECs than the STB. Pump upgrades are possible without intrusion into customer's homes. Grid encoding will require more bits on plant than the SVC. Note: the bits will be located on the "cheap" and easy-to-maintain distribution network, rather than the expensive and inaccessible network in the customer's home.

CONCLUSION

Grid encoding addresses a key component that is lacking in our current video compression architecture: how to make our services extensible. Within a few short years 1080p60 video will be available on STBs. Whatever technology is deployed today will ultimately have a very short life span. Ensuring a clear path to the future is a core component of any valid architecture that an MSO should consider deploying.

The Motion Picture Experts Group is highly regarded for "doing compression the best.," as evidenced by industry-wide adoption of such standards as MPEG-2, MPEG-4 and H 264. However, the group's' work on H.264 did not incorporate a mechanism that will allow MSOs to implement the technology. The technology addressed in this paper may belong to the Society of Cable and Television Engineers (SCTE) as they claim responsibility for everything in between the encoder and decoder. The future of technology is certainly in doubt, thus a flexible and extensible architecture is a vital component for a successful cable future.

Multiple CODECs will exist concurrently on the network. The network will evolve and building an architecture that allows for expansion is integral to our future success.