Guidance for Conforming Track and Referencing Them

Encoding Best Practice

This practice defines how to align tracks when timecodes are not necessarily aligned.

NOTE: No effort is being made by the Motion Picture Laboratories to in any way obligate any market participant to adhere to this specification. Whether to adopt this specification in whole or in part is left entirely to the individual discretion of individual market participants, using their own independent business judgment. Moreover, Motion Picture Laboratories disclaims any warranty or representation as to the suitability of this specification for any purpose, and any liability for any damages or other harm you may incur as a result of subscribing to this specification.

REVISION HISTORY

Version	Date	Description
1.0	TBD	Initial publication

1 TIME AND TIMECODES

1.1 Introduction

It is astounding how something as seemingly simple as a timecode can be some complicated. Some of this is rooted in history, and much is rooted in the industries inability to come to agreement on practices. There are common practices, but practitioners inevitably encounter subtle deviations that ultimately result in media components not lining up in time.

While this paper has no illusions about providing a universal solution to aligning audio, video, timed text, and metadata we hope to provide some practices for using Media Manifest/Media Manifest Core (MMC) to resolve some ambiguities in timelines.

1.2 Scope

This paper address various conditions where time needs to be aligned across media and time-based metadata. This generally works because parties are using an (implicitly?) agreed upon set of assumptions. With these assumptions it is possible to know how to align everything. However, not everyone has (or had) the same assumptions which means alignment might be non-obvious.

This paper addresses those case where tracks or time-based metadata come from different sources that use different assumptions.

By media, we primarily mean video, audio, and timed text. Time-based metadata covers everything else that might be aligned to media. Time-based metadata may be information about the assets (e.g., quality control (QC) comments or OCR text detection) to metadata that modifies the media (e.g., HDR metadata such as DolbyVision or HDR10+). When data that is usually static varies over time, it could also be considered time-based metadata.

When media tracks and associated data can be aligned in time, they are said to *conform*. This just means that there is some way to align them and figure out when they belong.

Note that his paper is only addressing timecodes used in file-based workflows (i.e., not linear broadcast), and only computer-readable data (i.e., not burned-in timecodes).

This paper only addresses simple conform (i.e., time align tracks), not more complex conforms, such as conforming in the presence of commercial black, or conforming to a specific edit.

This series of papers is focused on the MovieLabs Digital Distribution Framework (MDDF), but the concepts apply elsewhere as well.

1.3 Time and Timecodes

The simplest model for time has the beginning of the program starting and 0 and ending at the time of the end. This is sometimes referred to as media time.

The elegance of media time is that it is independent of encoding artifacts such as framerate. With sufficient precision, events can be tied to content as long as the playback speed does not change. If playback speed is scaled, time can be scaled accordingly.

From the perspective of story, media time is the only valid time representation. This means that unless you know the encoding, only media time can be used. If events are tied to the edit (e.g., as referred to by an EIDR edit ID), media time is the only valid solution.

In accordance with this argument, Media Manifest/MMC originally only allowed media time which forced parties to map timecodes to and from media time. Eventually, excessive begging led to the addition of other timecodes, resulting in the necessity for this paper.

1.4 Timecode

Timecode is a general term for any encoding of time in media. It generally refers to SMPTE Timecode which can be encoded in video, audio (i.e., AES/EBU timecode), and timed text formats. Most commonly, events are tied to video which means SMPTE timecode.

1.4.1 Media Time Timecodes

Generally, media time is expressed in seconds, including fractional seconds. With sufficient accuracy, frame accuracy can be achieved (even without knowing framerate). Media Manifest and Common Metadata (Section 3.24) describe the necessary assumptions.

1.4.2 Starting at the Beginning

Media often starts with material that is not part of the program. For example, video might have slates or bars. Audio might have 2-pop lead-in. The convention for knowing when the program starts is to use “hour-based” timecodes. This means the program starts at 1:00:00:00 and everything before it is lead-in.

However, there can also be “zero-based” timecode where the program begins at 00:00:00:00.

It is important to know which one applies.

1.4.3 Dropframe and Frame Math

In a 30fps video, one might think that frame 29 is followed by frame 0. Unfortunately, to avoid interference between audio and video carriers in color NTSC video the video frame rate was set to about 29.97 frames per second (30 x 1000 / 1001). “1000/1001” is referred to as the multiplier.

However, playback systems are based on seconds and although playback is 29.97 frames per second, using timecode with incrementing frames will result in drift. Let’s say we have a 2 hour video. It has 120 minutes x 60 seconds/minute x 30 frames/second * 1000/1001 = ~215,784 frames. However, if you assume 30 frames/seconds you’d have 216,000 frames, an error of 216 frames or 9 seconds. So, during that two hours, you’d have to drop 216 frame in the timecode. That’s about a frame every 33.33 seconds. If you drop two frames timecodes a minute, you’d drop 240 timecodes (too much). So, every 10th minute, don’t drop two frames and you’re at 216. In short, the first two frame timecodes are dropped every minute (i.e., not 1 frame every 30 seconds), except every 10 minutes where nothing is dropped.

The makes ‘frame math’ a little complicated because you can’t assume that the next consecutive frame is the one with the next frame number. I might go frame 29 to frame 00, or it might go from frame 29 to frame 02. If trying to conform to an edit by timecode, all this must be taken into account.

Conformed subtitles must use the same timebase as the video to which it is conformed. If the subtitle is based on non-dropframe and the video dropframe, events will be seconds off by the end of the program.

Non-dropframe (NDF) is generally formatted with a colon before the frame (hh:mm:ss:ff) while dropframe (DF) is encoded with a period or semicolon (hh:mm:ss;ff). MDDF requires explicit statement of dropframe through an attribute (@dropframe) rather than relying on formatting.

1.5 Time in Media Manifest and MMC

Within Media Manifest / Media Manifest Core (MMC), time and timecode are used in Presentation, Playable Sequence, and Timed Events. The key is to align time between and within these objects. For example, it must be possible to conform tracks within a Presentation, and Markers and Chapters must use time that is consistent with that conform.

A later section will describe best practices for doing this.

Note that Media Manifest/MMC originally supported only media time which forced parties to map timecodes to and from media time.

2 TIME ALIGNMENT

It is always possible to associate time between objects if there is enough information. This information is sometimes assumed, which is acceptable given that all parties are using the assumptions.

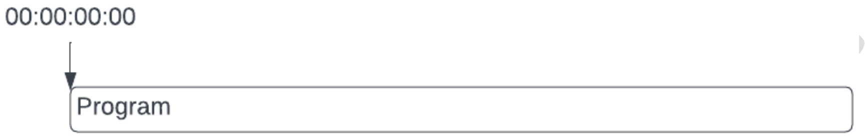
We provide the guidance for stating those assumptions and for expression exceptions.

2.1 Program Start

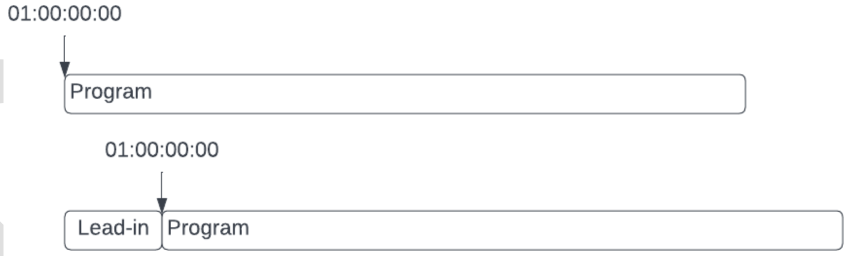
Program start is where the content begins. For example, the beginning of the movie, TV episode, or advertisement. As mentioned above, there might be material in the audio or video that precedes the start of the program.

The most common methods are

- ‘Zero-based’ – The beginning of the program is the beginning of the asset. There is no material prior to the start of the program



- ‘Hour-based’ – The beginning of the program corresponds with the timecode associated with one hour. There may be lead-in material prior to the 1-hour timecode such as slate, color bars, or 2-pop.



There are analogies for image sequences (one image per frame). The zero-based frame sequence starts at from 0000. More often, the start frame is 1000 with frames with lower frame numbers being lead-in material.

Program Start for a Presentation is expressed in Presentation/StartTimecode. This indicates the start time across the entire Presentation. It is assumed that

2.2 Playback speed

Content is generally created at particular speed, at the discretion of the creative artists. We’ll call this real-time, even though a work might be artistically sped up or slowed down. For video, this would correspond with 24 fps for modern “film”, and 25 or 30 fps for TV. High framerate video can be 48 fps, 60 fps, 120 fps, or something else. For audio, this would correspond with 44.1 KHz or

48KHz sample rate. Regardless, of the video fps or audio sample rate there is a rate at which the content was intended to be played.

If the playback speed changes, for example exhibiting a 24 fps film on 25 fps PAL TVs, all time must be adjusted accordingly.

Note that while dropframe manipulates timecodes, it says nothing about playback speed. Playback is 30 fps for both DF and NDF.

2.3 Conforming

For a set of tracks to be conformed, it is necessary that at every point in time the content or metadata corresponds with the content or metadata in all other tracks.

It is not necessary for all tracks to be the same length for them to conform, although this is often the expectation.

For tracks and metadata to conform, it must be possible to align both start time and playback speed.

2.3.1 Conform terminology

Conform can be used as a verb (e.g., “Conform those tracks”), or a noun (e.g., “There are five tracks in the conform.”). “Conformed” is an adjective (e.g., “The tracks are conformed.”). The phrase “simple conform” generally applies to aligning timecodes in different tracks so they conform. The phrase “complex conform” typically applies to conforms where other operations are required, such as aligning audio and video on different timelines (e.g., different edits, or removing commercial black).

2.3.2 Playback speed

MDDF assumes conformed tracks are all the same playback speed. There is currently not way to state an exception. Simply put, if they are not at the same playback speed, they should not be in the same Presentation.

Note that Video/Picture/OriginalPicture will define the original framerate of the video before transcoding to its current state. It says nothing about the current video, other than explaining framerate transcoding artifacts.

Note that we had the option to add this to the Inventory, but opted to avoid this because of the high likelihood of error.

2.3.3 Presentation Timeline

A Presentation has a single timeline. This is necessary because this timeline is referenced by Markers, Chapters, Playable Sequence Clips, and Timed Events.

Presentation Timeline can be explicitly stated or inferred. The ability to specifically stated was added in version 1.12, so it must be inferred in previous versions.

2.3.3.1 Explicitly Stated Timeline

Presentation timeline is explicitly stated using the PresentationStart element in Presentation/PresentationTimeline. PresentationStart is the timecode of the beginning of media time. StartTime can be represented in any of the supported timecode formats (i.e., seconds.milliseconds, dropframe timecode, non-dropcode timecode, etc.).

Typically, PresentationStart is either zero or 1-hour (i.e., zero-based or hour-based).

More formally, the PresentationTimeline-type defines timecodes for the Presentation and allows individual tracks to be conformed to that timeline.

PresentationStart represents the offset from zero-based time that is encoded in the Presentation's tracks. When PresentationStart is present, it indicates the start timecode of the conformed set of tracks.

If absent, PresentationStart corresponds with the start of the program in the media tracks, typically zero or 1 hour. If this is ambiguous, PresentationStart should be used.

Note that PresentationStart is intended address content encoded with hour-based timecodes (i.e., start timecode of tracks start with 01:00:00:00), while still allowing markers to be expressed in media time.

2.3.3.2 Inferred Timeline

The Presentation timeline is determined as followed based on tracks referenced in Presentation/TrackMetadata:

- When there is a video track and it has timecode information (i.e., timecode track)
  - If there is one video track, use its timeline
  - If more one video track is present, use the first listed video track with the lowest TrackMetadata/TrackSelectionNumber.
- When no video tracks are present or the video track has no timecode, and audio contains SMPTE timecode (i.e., AES-EBU embedded timecode)
  - If there is one audio track, use its timeline
  - If there is more than one audio track, use the first listed audio track with the lowest TrackMetadata/TrackSelectionNumber
- When there is no timecode present in any video or audio track, the Presentation timeline is zero-based. Time references should be media time.

If there is no timecode reference, the Presentation Timeline should be referenced by media time. If the Presentation Timeline has a reference to SMPTE timecode (e.g., via a video track), and Start Time is known, either SMPTE Timecode or media time can be used to reference the Presentation Timeline.

2.3.4 Default Conform Assumption in Presentation

Media Manifest assumes, by default, the following for tracks referenced in a Presentation.

- All tracks are played at the same speed
- The content of each track aligns with all other tracks
    - All tracks have the same lead-in. For example, if video has slate and bars, audio will start at the same time as the video.
    - Subtitles timecodes correspond with the Presentation timeline (which is typically a video track).
    - Ancillary tracks are conformed to the track they reference (e.g., a Dolby Vision track is conformed to the referenced video track).
    - Markers, Chapters, and Timed Events align to the Presentation timeline

Furthermore, the following assumptions apply to aligning time-based metadata including Markers, Chapters and Timed Events:

- All time-based metadata is aligned to the Presentation Timeline.
- If time references use media time they are relative to program start.
- If time references use timecode, they refer to the corresponding timecode in the Presentation timeline.

2.3.5 Simple Conform when Default Assumptions are not valid

When any of the default assumptions are not met, additional action is needed to conform the tracks. These are still 'simple' conforms because tracks and metadata can be conformed with timecode transforms. In particular, if one has a common reference point and playback speed is the same, it is possible to align the tracks/data in time.

2.3.5.1 All media and metadata reference Presentation Timeline

The key is to establish the Presentation Timeline as a reference and then define offsets for each track and datum relative to the Presentation Timeline.

2.3.5.2 Media Tracks may need to be adjusted to Presentation Timeline

If tracks or metadata are not conformed (time-aligned) with each other their timecodes must adjusted for each track until they conform. This only works if the only problem with conformance is time-offset (i.e., simple conform). If not, see Complex Conform in Section 2.3.6.

AudioTrackStart, VideoTrackStart, and SubtitleTrackStart instances are used to 1) avoid ambiguity of track start time, and allow for simple conform of tracks. Start times are the timecode within individual tracks that correspond with PresentationStart. By default, track start values correspond with the start of program in the track.

If a track start is different from PresentationStart, timecodes in the track are adjusted by the difference. For example, if PresentationStart is 1 hour, and VideoTrackStart is zero, then 1 hour is added to timecodes within the video track. If PresentationStart is zero and VideoTrackStart is 1 hour, then 1 hour is subtracted from timecodes in the video track.

Presentation-relative timecodes such as found in Chapters and Markers should be relative to StartTimecode for frame-encoded timecodes. Media time is relative to PresentationTime, so time-encoded timecodes (i.e., seconds.milliseconds) should be zero-based relative to PresentationTime.

If media tracks are shifted in time, Markers, Chapters, and timed events might require adjustment if their point of reference (e.g., a video track) has been adjusted.

Note that track starts are offsets and there need not be a frame or event with that exact timecode.

2.3.6 Complex Conform

The term “Complex Conform” refers to taking actions necessary to conform tracks when any of the default assumptions are not met, and a simple timecode alignment is not possible.

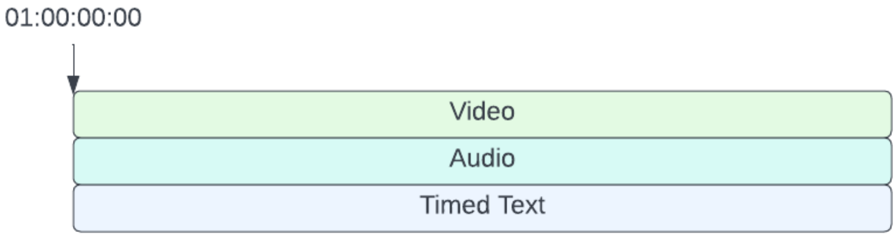
Media Manifest does not provide information specifically designed to facilitate complex conform.

3 USE CASES / EXAMPLES

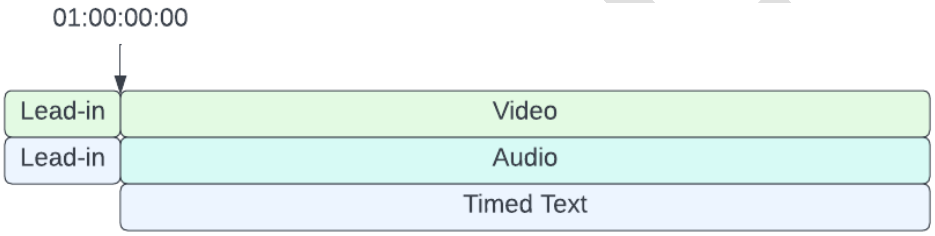
3.1 Conformed Asset

3.1.1 Assumed hour-based

In this use case, all assets are aligned to an hour-based timeline. Media time is assumed to start at 1 hour. Any zero-based tracks will have to be mapped to hour-based (i.e., add 1 hour).



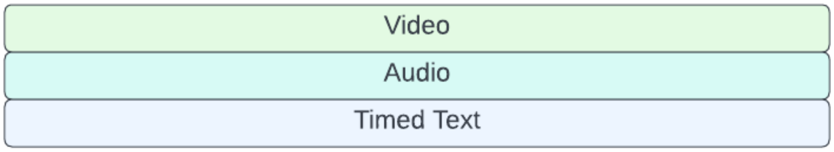
If tracks have timecode, any material before 1 hour is assumed to be lead-in.



3.1.2 Assumed zero-based media-time

In this case, all assets start at time 0 with nothing prior to program start. That is, the program starts at the beginning of the tracks.

This can occur in cases where there is no lead-in, regardless of whether timecode is hour-based, zero-based, or absent. Hour-based tracks will have to be mapped to zero-based.



3.2 Using StartTime to force to hour-based or zero-based

Presentaion/PresentationTimeline can be used for correct time.

3.2.1 PresentationStart

PresentationTimeline/PresentationStart is used to explicitly state the start time for the Presentation as a whole. This can force hour-based to zero-based, force zero-based to hour-based, or establish a start time when it is unclear from the tracks.

PresentationStart of zero (i.e., 00000.000, 00:00:00:00) will force to zero-based.

PresentationStart of 1 hour (i.e., 3600.000, 00:01:00:00) will force to hour-based.

@dropframe will correspond with whether dropframe time is assumed for the Presentation.

@format must match the format of PresentationStart, and also defines the expected timecode format for time references. Note that dropframe is indicated by @dropframe, not by the semicolon.

```
<manifest:PresentationTimeline>
    <manifest:PresentationStart dropframe=’true’ format=’hh:mm:ss:ff’>00:01:00:00</manifest:PresentationStart>
</manifest:PresentationTimeline>
```

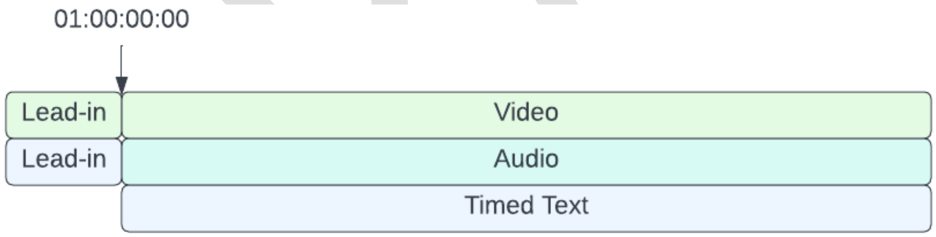
3.3 Aligning tracks to StartTime

Once StartTime is established, individual tracks can be adjusted to align.

PresentationTimeline/{Audio|Video|Subtitle}TrackStart define the start time for individual tracks. There is an element for each track type. Each references a particular Track ID in Presentation/TrackMetadata. For example, AudioTrackStart/@audioTrackID corresponds with an instance of TrackMetadata/AudioTrackReference/AduioTrackID.

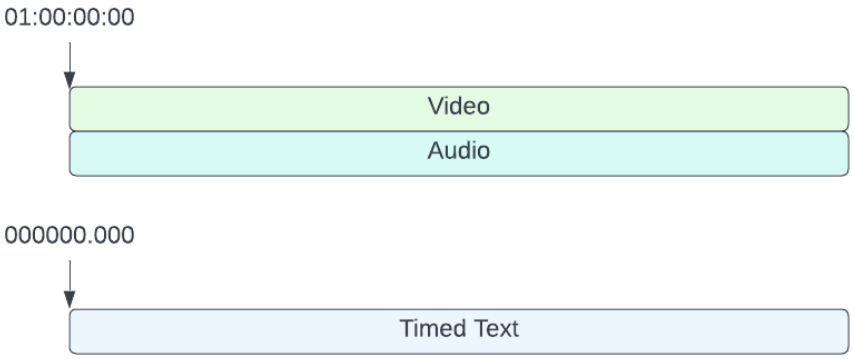
{Audio|Video|Subtitle}TrackStart references a time within that track that constitutes the start of the program.

For example, if the start time of a video track is 1 hour, then VideoTrackStart would be 01:00:00:00.



```
<manifest:PresentationTimeline>
    <manifest:PresentationStart dropframe=’true’ format=’hh:mm:ss:ff’>00:01:00:00</manifest:PresentationStart>
    <manifest:AudioTrackStart dropframe=’true’ format=’hh:mm:ss:ff’>00:01:00:00</manifest:AudioTrackStart>
    <manifest:VideoTrackStart dropframe=’true’ format=’hh:mm:ss:ff’>00:01:00:00</manifest:VideoTrackStart>
    <manifest:SubtitleTrackStart dropframe=’true’ format=’hh:mm:ss:ff’>00:01:00:00</manifest:SubtitleTrackStart>
</manifest:PresentationTimeline>
```

Let’s say there is one track that zero-based.

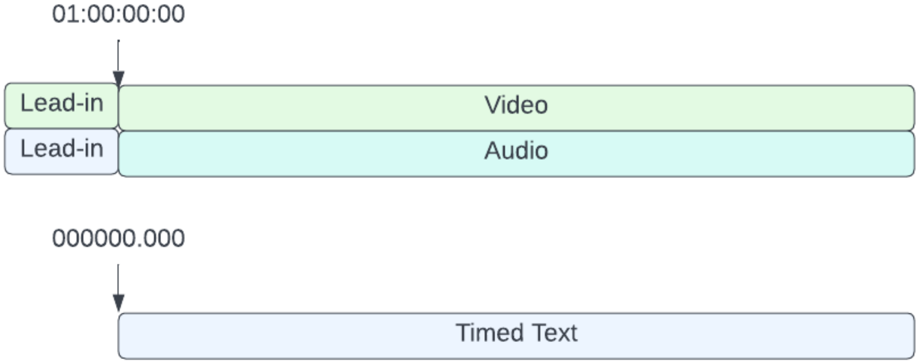


This would be corrected simply by stating at track starts at 0.

```
<manifest:PresentationTimeline>
    <manifest:PresentationStart dropframe='true' format='hh:mm:ss:ff'>00:01:00:00</manifest:PresentationStart>
    <manifest:AudioTrackStart dropframe='true' format='hh:mm:ss:ff'>00:01:00:00</manifest:AudioTrackStart>
    <manifest:VideoTrackStart dropframe='true' format='hh:mm:ss:ff'>00:01:00:00</manifest:VideoTrackStart>
    <manifest:SubtitleTrackStart dropframe='true' format='hh:mm:ss:ff'>00:00:00:00</manifest:SubtitleTrackStart>
</manifest:PresentationTimeline>
```

Note that the Presentation timeline is hour-based as specified in PresentationStart, so when calculating timecodes, one hour must be added to each timed text timecode.

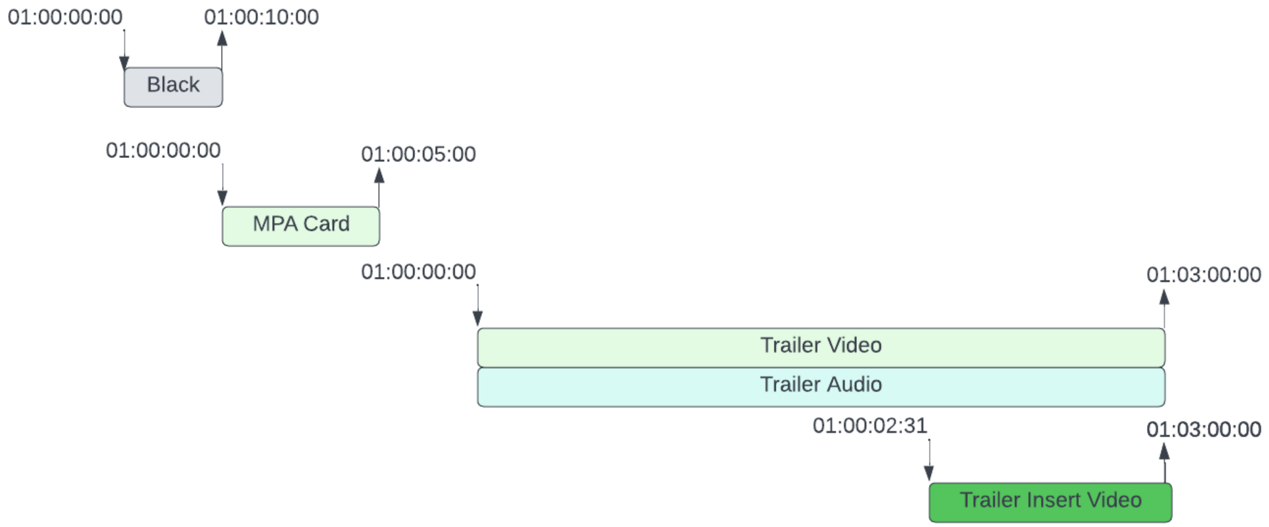
Note that this is the same regardless of whether there is lead-in.



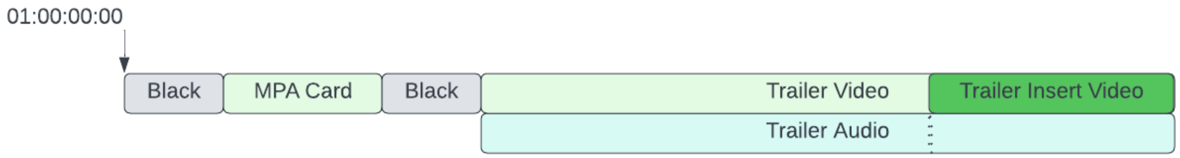
3.4 Replacing a Segment of Video

This example covers a case where there is an insert. In this particular example, want to cut in the last ~15 seconds of video.

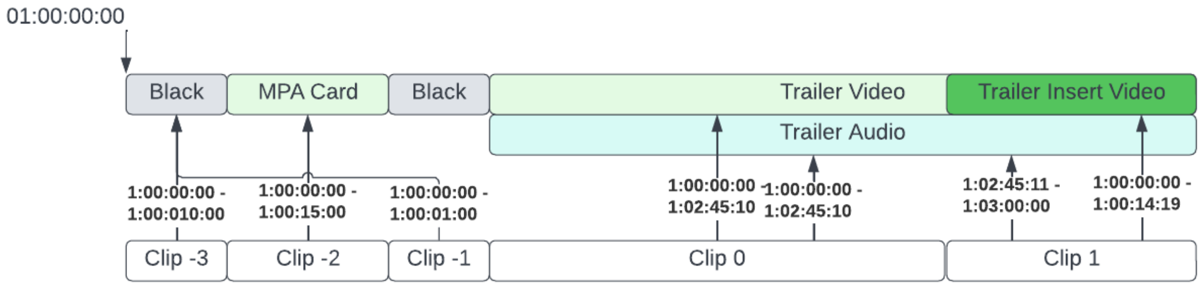
Following are the assets that need to be put together.



The goal is to look like this:



Playable Sequences are created as follows:



Notes

- You only need one Black clip. In this example, it is referenced twice. Going forward, special ImageID values will be supported so it will be unnecessary to include this video.
- For Clip 1 audio, @seamless is true to indicate there is no gap in the audio. The recipient of this Manifest should realize that both clips are referencing the same audio and the just play it through.
- To start at 1:00:00:00, set SequenceTimeline/PresentationStart to “01:00:00:00”.