Automatic essence timing alignment in IP production
– final draft post-internal-review - Andy Rayner, 30th October 2021 SMPTE ATC

**Abstract:**

One significant challenge that has always existed in broadcast production is that of the achieving full synchronisation of multiple video, audio and other associated real-time media flows at any point through the processing chain. This paper will discuss these challenges and propose principles of a potential system architecture to facilitate automatic essence alignment within SMPTE ST 2110 based IP production systems.

**Background & challenge:**

SMPTE ST 2110 has been a phenomenal success and is now well established as 'business-as-usual' for high and middle tier IP production systems both large and small right across the world.

During the ongoing deployments of these SMPTE ST 2110 IP production systems, some 'system wide' challenges are coming to the forefront – timing is one of these. ST 2110 is an interface specification, and therefore, by design, does not directly attempt to specify or address system architecture.

The very heart of ST 2110 is that of transporting related media essences as separate flows and reconciling to a 'composite' when required. The challenge with this is that a mechanism is needed to allow for this reconciliation in a time-aware manner – and this does not currently exist.

In the design of ST 2110, key decisions were carefully made on how to handle media timing. The transport of video essences is based on IETF RFC 4175, an already-existing specification mapping uncompressed video into RTP. The timebase of the RTP timestamps is derived from PTP using ST 2059. RTP, defined in IETF RFC 3550, provides a 32-bit timestamp, which is used in SMPTE ST2110 to represent 'origin' or 'acquisition' time.

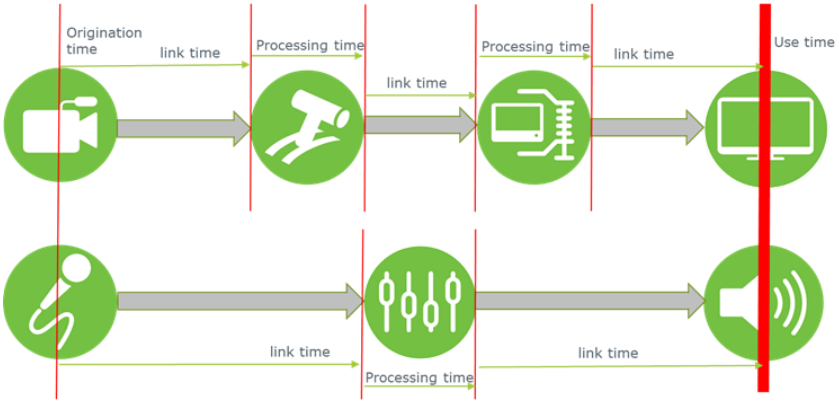
RFC 4175 defines the use of the timestamp with a constant timestamp value for all packets representing the samples from the same frame. (Interestingly, this appears to be a deviation from RFC 3550 that defines a continually incrementing timestamp with each packet transmitted). Considering this timestamp freezing across the frame, there is an additional challenge created here in that there is only one alignment point per frame – and that is at the head of the frame.

RFC 4175 also appears to mix up the concept of transport timestamps with media sampling timestamps. Some initial proposals that were evolved around the time ST 2110 was being developed included a media sampling timestamp in addition to the RTP transport timestamp. These proposals did not make it into the final ST 2110 specifications, which remains with the single RTP timestamp.

By the definition of the RTP timestamping in ST 2110, all sources presenting acquired video and audio will have common aligned timing (subject to any phase differences as per the sampling definitions). Synthetic sources (replay servers etc) that playback content with 'as now' creative intent will also have this same common aligned timing.

All processing devices immediately downstream of these sources (e.g. vision mixer) will be able to make use of the incoming timestamps as representing the origination time of the acquired essences. This allows alignment of the various sources, not only the 'phase' within a video frame but also the absolute

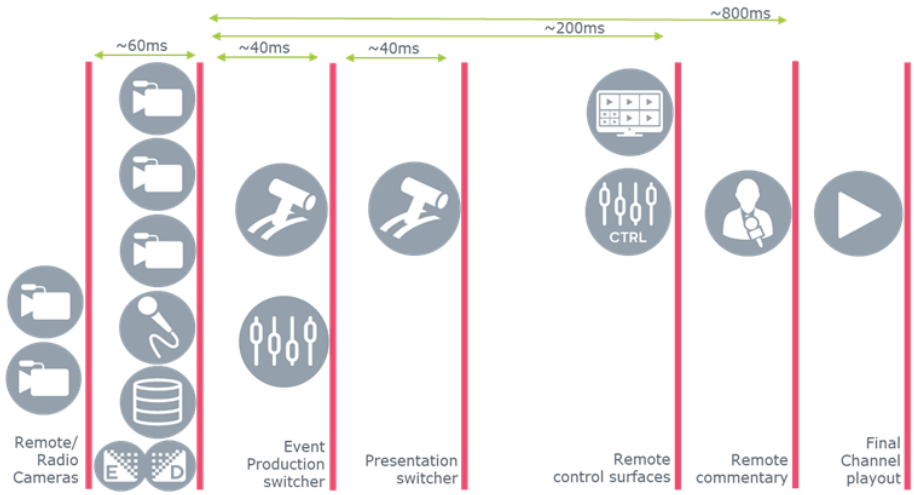
frames number. Note that prior to the recent revision of ST2110-10, there is no way of an upstream source declaring itself as being 'origin' timing.




The big challenge that arises once the media essences have been through the first piece of downstream processing since acquisition is that the meaning of the RTP timestamping egressing this downstream processor is *not necessarily* bearing relation to the origin capture time of the essences. The downstream device *may* be preserving the incoming RTP timestamp such that the same timestamp is reference to the same temporal element of essence, or the device *may* be restamping the egress flow with 'current time', inherently discarding the incoming timing information. The same principle applies to all processing elements in the chain. The latter of these is the default behaviour of most media processing equipment currently in the market.

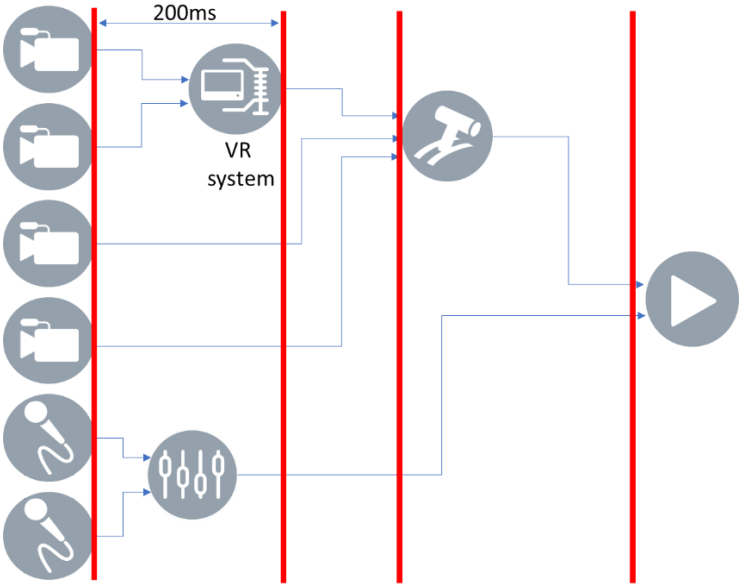
As a predecessor to SMPTE ST 2110, AES67 similarly uses RTP timestamps and a technique to define the presentation time of the stream on a per-link basis. This legacy does not preserve origination time but does provide a mechanism for defining and tracking delay.

**Different timing planes within the system**




As the media flows progress from the acquisition points through the chain, they are subject to gradually accumulating latency. Some of these latencies are only small (e.g. vision mixer) and some are bigger (e.g. Virtual/Augmented Reality Computer Generated processing – see example below), but the total cumulative delay at the end of the chain can be significant. As well as the processing latencies, there are other 'round trip' latencies within the production which can significantly extend the final cumulative delay. One of the most significant potential contributors is any form of public-internet-connected remote contribution that is involved in the production. An example of this is live multi-language remote commentary, often done using a low bit rate proxy compression of the program output. The return audio is then needing to be reconciled with the programme, incurring many hundreds of milliseconds.

**Recent real-world example**




A recent real-world example that demonstrates the need to align between different video flows and then with audio and temporal metadata is as follows: A Virtual Reality camera system will typically present the processed output several frames later than other non-VR related cameras on a set. It will also therefore be later than the audio flows. There therefore exists a requirement to align all of these flows - video flows in multiple timing planes and audios in yet another. Furthermore, metadata used within the VR has a different timing plane to metadata in the rest of the system.

**Moving forward in standards:**

In the recent review of SMPTE ST 2110-10, some additions have been proposed that will enable processing devices to state their output as being either a new timestamp, a preserved incoming timestamp or an 'origination' timestamp. This provides a mechanism to determine the level of preservation of origination time through a chain.

The review proposal also provides for processing elements to declare information on the latency they incur and recommends that they can be configured to present their egress at a pre-determined time (similar to link offset in AES67).

By combining the timing information from processing devices with the use of a control system to collate and arbitrate the incremental timing, it would be possible to create a system architecture that enables any element in a processing chain to be aware of the differential timing of the different essences.

Some processing elements in a production chain introduce very little latency whereas others can contribute significant latency. Other examples of high latency contributors include profanity delays, remote-contributed multi-language commentaries etc. The aim of the described architecture is to handle these significant latencies as well as the small latencies through processing devices such as video mixers etc.

**Software-based processing**

Until relatively recently, most elements of a live production chain have been entirely hardware appliance-based and are therefore usually both low delay and highly predictable in terms of processing latency. Moving forwards, however, most of the industry is looking towards using more and more software-based media processing functions for live production.

The world of fully linear transmission of media flows, initiated by the very physical behaviours and requirements of electro-mechanical television some 90 years ago has been perpetuated through all the broadcast interface specifications up to and including ST2110 and AES67 etc.

The reality is that native software compute processes are highly non-linear and inherently 'bursty' in the way they work - and software processes will typically work much faster than real time. Therefore, moving forwards, both the processing elements and the transfer of data between concatenated functions in a software workflow will be much less deterministic in terms of incurred latency than hardware-based appliances.

There will still be a drive (even more so in fact) to ensure that the processing, transfer and buffering of media essences software-based systems are optimised as much as possible to minimise end to end latency.

The timing planes approach described in this paper seeks to include latency management and relative timing alignment in software-based architectures within the scope of work.

**Fixed and variable latency processes:**

One challenge that is ever-present in a live production workflow is that of potentially variable latency processes. If these elements exist within a chain, then the system needs to have the margins and mechanisms to compensate. Fortunately, most modern appliance-based media processing elements feature both low and constant throughput latency.

AES67 addresses the issue of latency variability by specifying a presentation time using the link offset delay.

The variable latency of software processes (see previous section) also needs to be allowed for in the technique used for timing compensation.

**Different sources**

Sources can present in different ways from a timing perspective:

- Local 'live' acquisition sources that are PTP synchronised will inherently present as fully synchronous sources (allowing for the phase difference of native IP sources and SDI converted sources)
- Local 'synthetic' sources (e.g. replay servers) will be locked to PTP timing reference and will usually be playing back their content 'as live' and presenting with fully synchronous time.
- Some remote sources will be emanating from another location that has PTP absolute time lock so will have a constant phase offset but will need phase reconciliation. Absolute time tracking is possible here. Other sources may be inherently asynchronous or undefined in acquisition point and need both frequency and phase reconciliation to be used in the production. These will typically assume 'origin' status upon reconciliation.
- Some local sources such as radio-linked cameras that will typically be non-synchronous in acquisition. These sources are often reconciled to origin time, ignoring any encoding latency.

**User stories:**

As this paper is drafted (October 2021), work is about to commence on a ST2110 Engineering Guideline or Recommended Practice to define a system-level architecture that will meet the aim of facilitating automatic essence alignment in ST 2110 production systems.

As part of the work proposal to look at this system-wide technical recommendation, input would be appreciated in terms of user stories that will help to shape the outcome. Whilst any guidelines will not be able to cater for all scenarios, it is intended that the work is as comprehensive and all-embracing as possible.

**Conclusion:**

The system-level challenge of reconciling the timing planes of the different media essences in a production system is one that is as old as television itself. The advent of IP-based systems offers the potential for an evolution of a system timing architecture that can provide automatic alignment of flows at any point in the chain, whether it is comprised of appliance-based or software-based processing or a combination of the two. Any and all input is encouraged from the broadcast production community to help evolve a comprehensive and robust Engineering Guideline on this topic.

**References:**

SMPTE ST 2110: Professional Media Over Managed IP Networks suite of standards

IETF RFC 3550: RTP: A Transport Protocol for Real-Time Applications

IETF RFC 4175: RTP Payload Format for Uncompressed Video

AES 67: AES standard for audio applications of networks - High-performance streaming audio-over-IP interoperability