Grass Valley
WE LOVE LIVE

WHITEPAPER

Developing an Intelligent Timing Model for Live Media Production in the Cloud

By Chuck Meyer, Larissa Görner-Mees and Chris Merrill

Introduction

In live media production, timing has two primary functions. It allows tracking of the contributions to the program so that the production team can tell a logical, linear story. And it synchronizes the processing equipment used in the production chain so that team members can do things like seamlessly switch between video elements.

The traditional method for timing a production uses video frames as a standard unit of measure. But new models are required for distributed live productions where part of signal switching occurs in the cloud.

This paper focuses on live production with a methodology for using new technology to keep each operator's user experience coherent for their location while enabling individual contributions to logically align in the final production. Using cloud service providers for asset storage, content sharing, or program playout and emission rely less on specific timing.

Timing Problems in Distributed Production

A common reason to adopt a cloud topology is to coordinate work across a geographically dispersed team with distributed processing. Problematically, geographic distance creates delay in information transmission. Even across the most advanced fiber optic network connections available, data can't travel faster than the speed of light. Therefore, individuals contributing to a production will have varying amounts of latency that depend on the physical distance between the operator and the data center where the processing occurs. The farther the separation, the longer the delay.

In reality, light-speed connections aren't available over long distances and they are generally very expensive. Instead, Wide Area Networks are typically employed. But network equipment also introduces additional delay. Individuals connecting over the internet have little control over the number of hops their data path may take before arriving at their destination, further increasing path-dependent latencies. As a result, times for receiving and returning information will vary for each member of a geographically-distributed production team.

While some aspects of a production may be accomplished in parallel, many steps in creating a live program feed must be sequential to maintain the thread of the creative process. This requires maintaining, for each team member, the perception that their contributions are part of a real-time sequence, regardless of when they receive and return their contributions. It also requires all contributions to the program to be correctly aligned before distributing the program to the audience.

This is what AMPP solves automatically.

Why Proposed Solutions Don't Solve the Problems

Signal standardization and content creation introduce latency

Most signals within a broadcast facility must be processed. Each processing step creates additional delay.*[1] For example:

- Initial synchronization of a "wild feed" = 1-2 frames
- Conversion of program inputs to production format = 2 frames
- Creation of video effects in production switcher = 2 frames
- Conversion of program to transport formats and distribution codecs = 2+ frames

Traditionally, the signals within the system are adjusted, or "back-timed" relative to a master clock to maintain alignment. In the example above, if any one signal followed the outlined path, all of the signals in the facility would need to be back-timed by a minimum of 8 frames.

But what is the master clock in a distributed system? It is possible to create a global master clock. Modern network technology is based on NTP, and the higher precision PTP, clock protocols that trace back to atomic clocks and ensure synchronization at global scale. But even a worldwide clock must respect causality. Some actions must follow others in sequence.

Transit times introduce additional variable latency

Because all contributions in a distributed system would have different transit times from the operator's workstation to the processing location, the system would have to treat them all as wild feeds, requiring a compensating adjustment for every network path and output of that source.

If all timestamps are forced to line up in a sequential fashion using a combination of the operator's workstation clock and a master clock, then the production will experience significant and growing delay over the course of the program as team members are forced to wait while other members return their contributions. The total time it takes to produce the content grows longer and longer.

How Does AMPP Provide a Solution That Works?

To provide a timing solution that works, we have to return to our fundamental reasons for timing

- Enabling the production team to tell a linear story
- Synchronizing the equipment in the production chain

Enabling the production team to tell the best story

Design for human realities

Our standard for system response times should match actual human realities. The de facto standard of measuring time on a frame-based clock, at 30 or 25 fps (33 or 40 ms), comes from a time when analog color carrier frequencies required this timing to ensure accurate delivery of color to the television set. This time base far exceeds what humans — and modern TVs — need.

Studies show that the fastest a person can react to an outside stimulus without any type of pre-cue is about 180 milliseconds. As shown in the timeline illustration, a minimum response time to receive a stimulus, take action and recognize a new state is about 240 milliseconds. The fastest replay operator in the world would take 240 milliseconds to see the action on the monitor, press the mark in button, and recognize that the clip record has started.

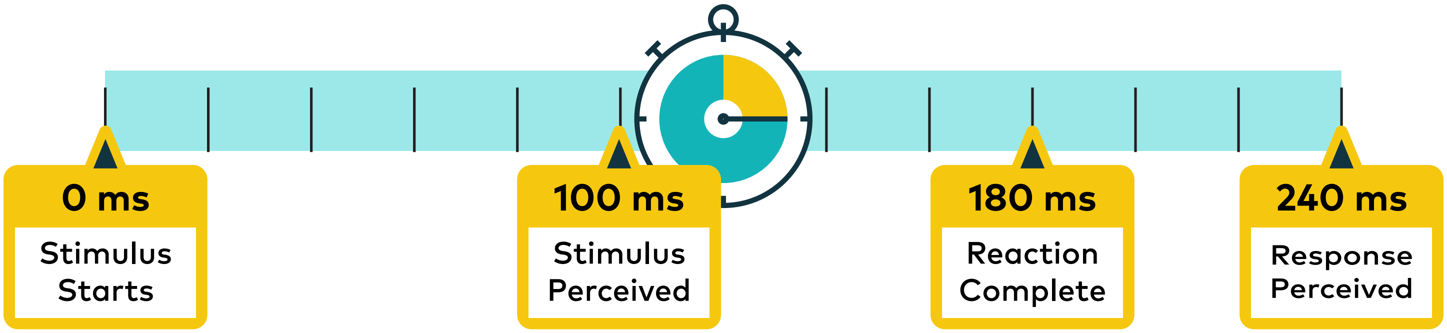


Figure 1: Human Reaction Timeline

*[1] For additional information on tracking delay through a broadcast facility see Wouter Kooij, Playout delay of TV broadcasting, 2012

More important than the actual time it takes to respond is the time it takes to notice a difference. People perceive differences in rates of change much faster than they perceive actual change*[2], as long as they differ by more than 20%. An ideal human will not be able to perceive the difference between 100 and 120 milliseconds of delay.

Finally, the synchronicity of interrelated stimulus must also be taken into account. Humans are most sensitive to changes between audio and video and least sensitive to changes between video and video. For an example, consider a multiviewer on a monitor wall with six video windows. If all six windows change within 120 (100+20%) milliseconds after the cue they will be perceived as in time. But if one changes first, the remainder will be perceived as late. Audio and video must change within 80 milliseconds of each other to be imperceptible. This is why lip sync is so noticeable.

To summarize, a system that feels live to the operator must have a response time of about 240 milliseconds (just under a quarter of a second) from the time the operator sees the cue to seeing the result of the action they have taken.

Make the production system feel "real-time"

It is important to remember that the user experience only needs to feel live to the operator. At the local workstation, audio, video, monitoring, intercom and control must all align within the tolerance ranges discussed above.

Unlike previous operations models, the local response times can be independent of any clock time. To feel live, they simply need to be coherent with each other. If the system manages the differential latency of the arriving essences at the operator's location, then back timing sources is not required.

➡ ***Humans are most sensitive to changes between audio and video and least sensitive to changes between video and video.***

For an operator, it is imperative to consider the relative latency of an essence, as well as its absolute latency. With AMPP, all creative decisions made by the operator and their associated processing time can be tracked relative to the operator's time. The order and local timing of the decisions are maintained. The operator experiences the phase-aligned environment they are used to. Yet the total environment is time-shifted relative to the source.

To maintain linear storytelling, the final result of the operator's work is time stamped with whatever offset time is best to synchronize the work across the production chain.

Synchronizing the production chain

To provide a unified timing model across the production chain, modern technology should follow a common design strategy:

a) For each audio, video or other essence stream entering the system, identify an essence landmark such as the video top of frame or audio time stamp.

b) Align all common essences based on their landmark with an established relationship between different essence types.

c) As editorial decisions are made, time-align the decisions with the essence.

d) Process the essence as orchestrated by the editorial decisions. The editor, or processing function, can be located anywhere.

e) Time stamp the final output based on any user-defined clock.

f) If required, a final NTP or PTP time stamp may be added.

Using these steps, the relative latency values of all essences are known, and differential adjustments can be calculated. Only as essences are aligned is a common time base is required. This may be any time base that is mutually suitable for all essences about to be processed.

*[2] M. Kanabus et al, Temporal order judgement for auditory and visual stimuli, 2002

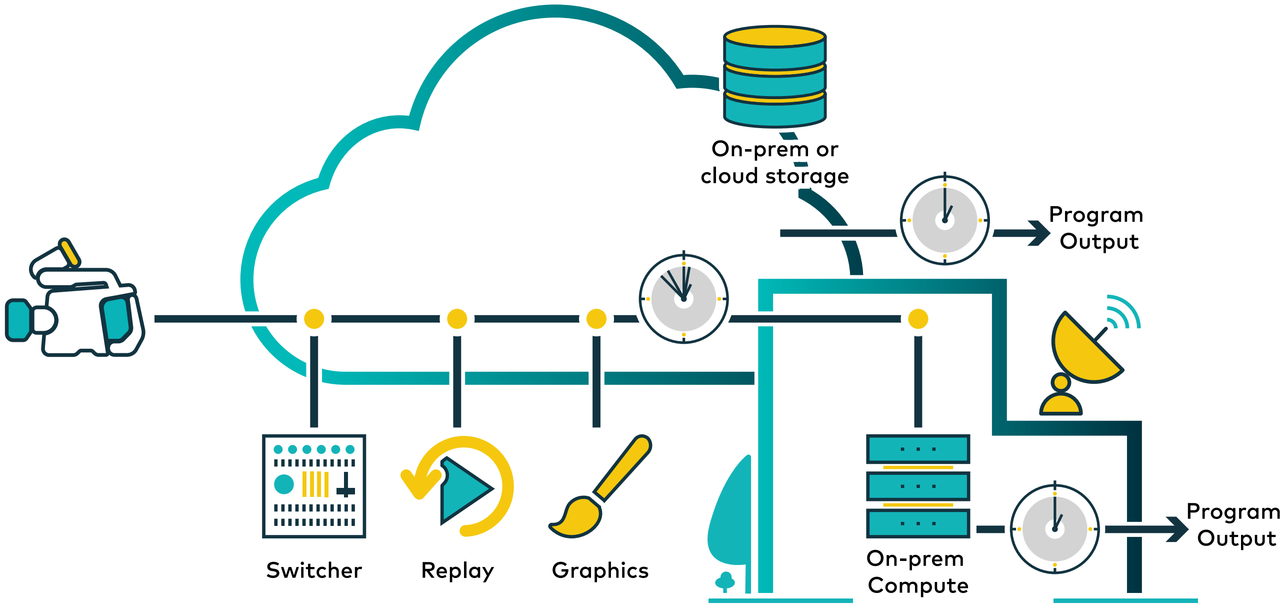


Figure 2: New technology can align contributions from multiple contributors

Unchaining individual workstations from external time is possible because today we operate faster than real time using technologies that did not exist when frames per second timing was implemented. Frame syncs are replaced by memory buffers. The AMPP Platform adjusts buffer depth to match the timing offset required for each essence.

- ***The AMPP Platform manages buffer size to compensate for latency.***

Following this design strategy, any live production task can be carried out in what feels like real time and assembled in a linear fashion to create programming that exceeds audience expectations. Even with complicated production tasks, total execution time is a few seconds. Compare this with today's traditional live broadcasts which, in the best of circumstances, still take as much as 50 seconds to get final emission delivery to the home.

The Production Latency Parameters in a Nutshell

It is informative to evaluate the sources of latency and then consider how latency is managed for the overall workflow as well as the needs of the operators in a workflow. Let's start by reviewing several potential sources of latency that occur during data transmission and routing.

Transport latency

The propagation delay for signal transmission includes the speed of light for a given medium as well as delay introduced by routing and switching equipment in the transport network. For fiber, the velocity of propagation delay is nominally 1.5 ns/ft. and coaxial cable is a bit less, at 1.2 ns/ft. These are approximate numbers as the index of refraction for each of these waveguides can vary based on the specific fiber or cable type.

To put this in perspective, 300 meters of coax will introduce around 1.2 µs of delay, and 300m of fiber will introduce 1.5 µs of latency. Perhaps it is more useful is to consider 1000 km of fiber. In this case the latency is 4.92 ms.

Delay is introduced by network routing equipment as it re-organizes data flows between ingress and egress ports. During this process, signals are buffered.

In addition, traffic dynamics can, and do, result in signals incurring a longer delay. Heavy network traffic results in jitter. To easily observe this effect, run a speed test of your connection from your device to a nearby data center. The value, or delay time, for "ping" will be shown, along with its jitter. Observe the values at different times during the day and you will note that rising delays in response to a ping are correlated to increasing jitter. Depending upon other traffic, peak jitter can be nearly 50% of the average value. Peak jitter must be accommodated to a reasonable degree. The delay can be considered as an average value, with a peak deviation, which is the expected peak jitter, or perhaps the peak jitter value which can be tolerated by the workflow.

- ***Transport latency = (1.5 ns*feet of fiber) + (switch buffer) + (peak jitter)***

It is not uncommon for transport latency, including network routing equipment, to be on the order of 5-10 ms, including peak jitter. This can be reduced with a private, or leased network, which is well groomed and managed to provide a better service level.

ARQ (SRT/RIST)

There are numerous protocols for moving data reliably over a network. UDP, TCP, ARQ, SRT, RIST, Zixie, FASP and more. These protocols, many of which include mechanisms for packet recovery, can introduce additional latency. The amount of latency varies by protocol.

This latency can be considered as a depending variable, which is known, based upon the protocol itself. AMPP is protocol agile, and when compensating for delay can consider this value as part of the total transport latency.

AMPP Streaming

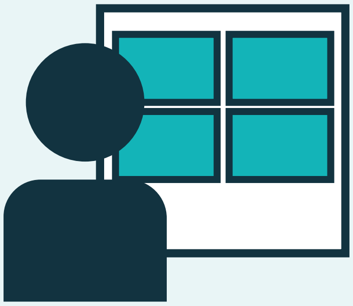
AMPP Streaming is another protocol for data transport. It has been integral from the inception of AMPP and offers very low latency. It is based on a particular profile of IET RFCs, and includes the ability to rapidly traverse firewalls. Additionally, once the secure connection is established, this fast transit capability is preserved.

Data Processing

Typically, compression technology is used when including the cloud to implement a workflow. There are a myriad of codecs available for use based on the application at hand. A GOP is chosen to trade between compression ratio and latency. For example, I-Frame has low latency, but only a low compression ratio, or gain. In some cases, available bandwidth is minimal and a higher compression ratio is required. AMPP supports a variety of codecs including NDI and FF MPEG I-Frame only codecs.

AMPP is codec agile. Based on the codec type, AMPP can determine the respective latency value as part of its calculations. Codecs can be at the edge as well as in the cloud. These codecs are typically implemented using software on CPU and GPU processors in the cloud. An edge device located on-prem might be CPU, GPU or FPGA based. Eventually, F1 (FPGA) instances may be widely available in the cloud. AMPP operates independent of codec implementation.

Not all cloud-enhanced networks require compression. There are private networks that are geographically wide; for example across the United States. Such networks can be provisioned to support full-bandwidth transport, thereby eliminating codec latency.



AMPP flow monitors, similar to a multiviewer, use cloud processing to provide an overall view of the essence signals in the workflow. They can be shown independently, or as a composite of multiple tiles, or PIPs. The final result, as in the case of a classical multiviewer, can then be provided as an output, within the cloud for other workflows, or transported out of the cloud to a remote viewpoint. On the input side, AMPP coordinates the individual latencies of each essence to create the full picture, which is coherent in time. At the output, the transported image will arrive at each viewpoint with whatever latency is incurred with each respective transport path.

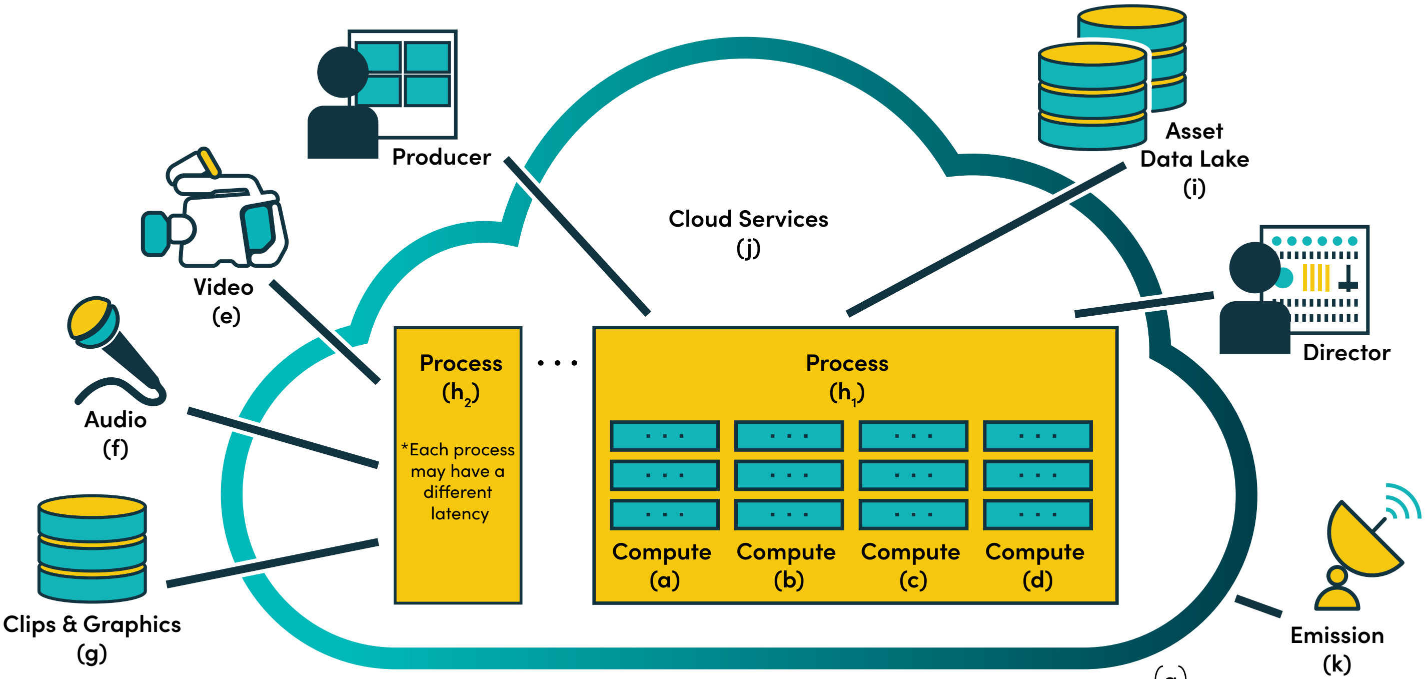
Essence Processing

Once the data path, including necessary encoding components is configured, the processing for the workflow is deployed. Some processing may occur on-premise, while other processes are carried out in the cloud. In either location, and in a hybrid configuration utilizing both locations, the supporting infrastructure for processing is likely to be common: a server with processing blades accessed with a NIC (network interface card).

Based on latency, some critical processing may occur on-premise. For example, audio production as illustrated by Zone 1 in Figure 4. When it is desired to minimize infrastructure on-site, then all the processing will be in the cloud.

Video processing will incur significant latency, usually one frame of delay, or more. Even when the processing requires only a few lines of latency, such as an anti-aliased squeeze or a basic, 1-alpha mix, using a GPU with frame-based processing is often most efficient overall, so a time base with granularity of one frame is simple and effective.

The parameters above are a few of the potential sources of latency. AMPP uses all of this latency information to orchestrate the timeline for production, ensuring that each process occurs in the correct order, and that each contributing essence is time aligned with the processing steps.



From input source(s) to program output, any path is ∑ latency e, f, i, j, $\left\{\begin{matrix}a\\b\\c\\d\\g\\h\end{matrix}\right\}$, k.

Figure 3: Factors that influence latency

Establishing Tiered Processing Requirements

As we've discussed, not everything in a production has to happen at the same time. Consider the case where response time might be layered.

Audio producers need very low latency when producing audio for video. These essence types must be coherent with latency of less than one video frame. The producer will be on-site, with some equipment. Perhaps they are located in the audio booth in the mobile truck. Their experience can be managed as an on premise configuration. The media is uncompressed, or very lightly compressed, and processing is carried out with the least amouont of latency possible.

This can be considered as a timing zone, with short delay — see Figure 4 Zone 1. This zone can then feed another zone with more relaxed latency requirements. This second zone can even be the same truck that houses both the audio booth and the video production theater. Zone 2 can include the cloud, configured for low latency production, and this second zone can feed a third zone, which is a longer latency cloud deployment used for studio production. In Zone 4 the production result is processed for transmission to the audience.

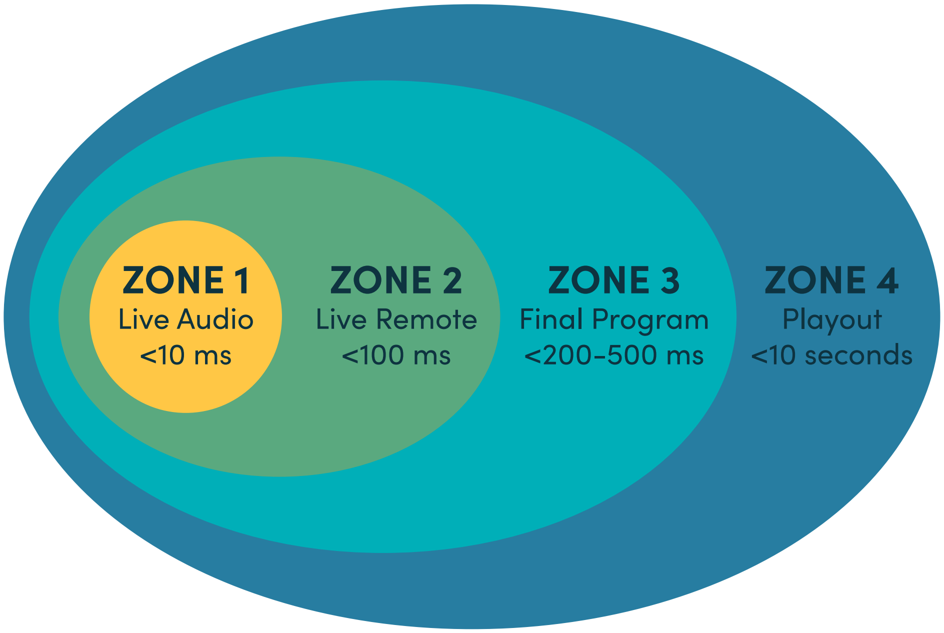


Figure 4: Establishing multiple zones of latency control

Distance Does not Matter

A deployment can simultaneously include more than one workflow and many endpoints that are either sources or consumers of media. While a range of latency can be calculated for a specific step in the production chain, that measurement of latency does not apply to the entire deployment. As noted, many of these latency sources are variables. This means that processing essence for the reporter on location and the commentator in the studio are two different calculations.

AMPP coordinates latency across the entire deployment. AMPP generates a sum for each desired endpoint and then time aligns the processing to the needs of an operator so they can manually control the workflow effectively in real time, without feeling the lag usually associated with remote operation. AMPP provides each person a comfortable experience, even though they may be thousands of miles apart. Causality cannot be violated, yet it can be masked in the range of 200 ms so that a human being does not notice the effect.

Grass Valley provides public notice of its allowed patents at www.grassvalley.com/patents.