SILEX INSIGHT

HOW TO ACHIEVE LOW LATENCY
AUDIO/VIDEO STREAMING OVER IP NETWORK

White paper

How to achieve low latency audio/video streaming over IP network

Standard audio video interfaces such as HDMI and Display Port are well suited for short range connectivity of multimedia equipment. However, larger AV (Audio/Video) installations with multiple sources and displays, or installations spreading over several physical locations need more evolved connectivity. For those AV installations, it has become obvious that IP networks is the most standard and future proof way of transporting the signals.

Although IP-based networks enable the flexibility and scalability required by many applications, special care should be taken to keep the latency of the system sufficiently low for real-time, live use cases.

This whitepaper will first define the latency for audio/video transport. It will highlight the usual architecture challenges of an AV over IP transmitter/receiver. A deeper analysis is provided regarding the video compression which is often mistakenly considered as adding too much latency. Eventually, actual latency measurements of the Viper 4K HDMI to IP transmitter/receiver will be presented.

How is the latency defined for an audio/video transmission?

The latency of a system is the delay between the instant a sample enters the system and the instant it leaves the system. In an AV over IP system, it translates into the delay between the first pixel of a video frame entering the transmitter through the video input and the first pixel of the same video frame going out of the receiver on the video output.

The latency is naturally defined in seconds usually in the range of milliseconds for a real-time audio-video system.
Video experts also define the latency as the corresponding part of the video stream during that time. The latency is then described as a number of frames or lines of a video stream. In this case, the actual time varies depending on the frame rate of the video as shown in the table below.

	Latency of 1 frame (ms)	Latency of 1 line (ms)
720p 50fps	20	0.0278
1080p 30fps	33.3	0.0309
UHD 60fps	16.7	0.0078

Table 1 Latency equivalence of one frame and one line of video signal for different video formats.

This definition is very convenient for some image processing algorithm where the added latency is, for example one frame, whatever is the frame rate.

There isn't a unique definition of what should be the latency of an AV over IP system. "Low Latency", "Ultra-Low Latency", or even "Zero Latency" are commonly used terms to indicate that the latency is good enough for the intended application and end user expectation.

Applications that involve the interaction of the user (like Meeting Presentation, KVM or live events) are usually the most critical in terms of latency. Some users will be more sensitive to the latency than others, but keeping the latency below 30 ms is usually accepted. Some applications benefit from even lower latency for a seamless user experience.

Architecture challenges of a low latency AV over IP system

Special care should be taken to the architecture of the transmitter and receiver in order to achieve low latency AV over IP. The latency of the system directly comes from the buffering of the video/audio at the different processing stages. This buffering is necessary to enable some features, but should be kept to a minimum as described later.

Due to the high bandwidth of the video signal and the latency constraint, it is essential to use dedicated hardware processing from the video input to the IP network. Purely software-based solutions will inevitably increase the latency because of the memory transfers and the CPU load.

Although having dedicated hardware support for the video stream is mandatory, it needs to remain configurable and flexible.
For this reason, the software running on a CPU takes care of all the non-real-time tasks.
The following diagram shows the basic processing stages of an AV over IP platform.

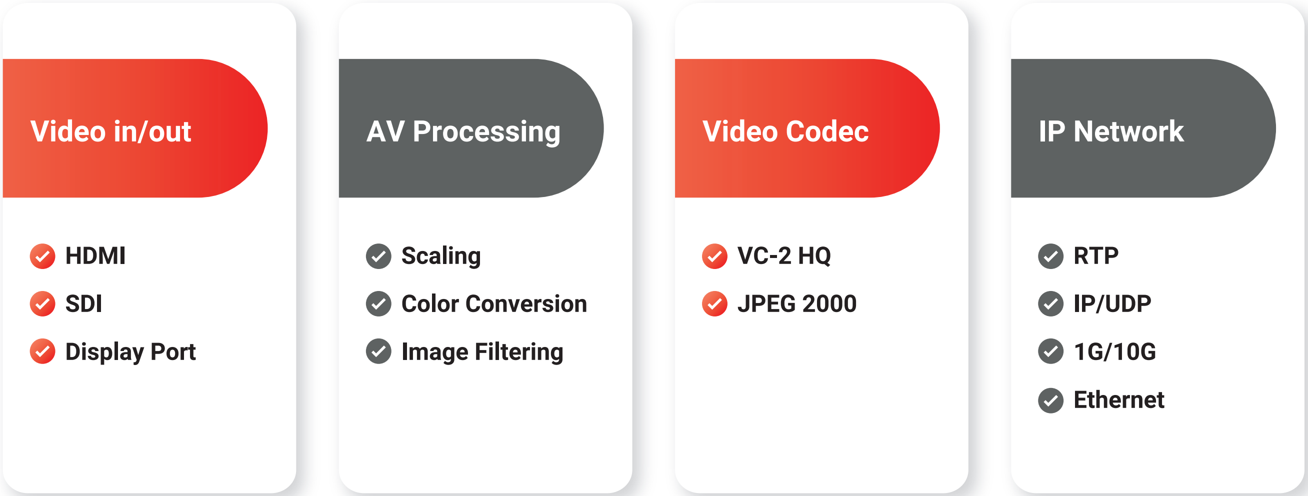

Figure 1 – Block diagram of the architecture of a typical AV over IP transmitter or receiver

A typical transmitter takes the video from its input, sends it through some video processing, video encoding and network processing before it outputs the stream on the IP network. A receiver does similar operations in reverse order. Each processing step can potentially add latency to the complete system and deserves a deeper analysis.

Video input and output

Receiving and transmitting on the video interfaces, such as HDMI, Display Port, SDI does not add latency to the system. A few frames may be discarded at startup during the initialization process of the input and output stages, but this does not add latency.

When the video content is protected with HDCP, there is an additional authentication phase that takes place when the cable is plugged in. After this authentication phase, the video can be encrypted/decrypted on-the-fly without any additional buffering, therefore without adding latency.

Another important aspect is that the video input and output of the AV over IP systems are located on two different devices connected together via the IP network. One of the challenges is that the video input of the transmitter board needs to run at the exact same frequency as the video output of the receiver board. If it wouldn't be the case, the receiver would have too many or not enough data to output on the video link making it quickly unstable. This issue is sometimes solved with a frame buffer at the receiver that can drop or repeat a frame when necessary, but this adds a frame of latency. The best approach is to implement a clock recovery mechanism over the network that will replicate the video clock of the transmitter at the receiver, guaranteeing synchronized operation.

Video processing

Transmitters and receivers often include video processing functionalities. It may include among others scaling, chroma up/down sampling, color conversion, frame rate conversion and image filtering.

Most of the video processing functionalities are described by a filter function. Each filter requires a certain amount of data to be buffered during the calculations, adding up to the total latency. If a filter uses pixels of a single line, the latency is negligible. The impact is more important if a filter uses a large part of the frame or even pixels from previous frame(s).

Video compression

Video compression is used to reduce the bitrate of the video. In the case of video transport over IP, reducing the bitrate has a direct positive effect on the network infrastructure costs. It also enables more video streams to be transported on a specific network installation without congestion. The following table gives an overview of the bandwidth of the raw video (uncompressed), together with the minimum compression ratio required to fit in 1G or 10G Ethernet.

Video source	Raw Bandwidth	Compression Ratio 1Gb Eth	Compression Ratio 10Gb Eth
3G-SDI 1920x1080, 60fps, 4:2:2, 10-bit	2.5 Gb/s	3:1	NA
Blu-Ray UHD 3840x2160, 60fps, 4:2:0, 10-bit	7.5 Gb/s	9:1	NA
HDMI 2.0 4096x2160, 60fps, 4:4:4, 8-bit	12.7 Gb/s	15:1	1.5:1

Table 2 – Example of compression ratio for several video formats. The raw bandwidth does not take into account the vertical blanking of the video signal.

It is often said that compression adds a huge latency to a system. This is simply not true if the right codec is selected. When choosing a compression algorithm, there are a lot of aspects to take into account like the compression ratio to achieve, the quality expectation, the complexity in hardware or software, the interoperability with other equipment and of course the latency. Each application has a different set of requirements that will lead to one or multiple possible codecs. The video compression topic is further explained in the next chapter.

Network transport

Before the audio and video data can be sent over the IP network, it needs to be encapsulated in several protocols. The audio and video is usually transported in RTP packets that are themselves encapsulated in UDP/IP packets. UDP protocol is used for the real-time transport as it allows broadcast/multicast. Moreover a connection-oriented protocol with packet acknowledgement and retransmission such as TCP would not work for a real-time and low latency transport. Each IP packet is encapsulated in an Ethernet frame. The payload of the Ethernet frame is limited to maximum 1500 bytes on general purpose networks. For this reason, each video frame is divided into many small packets for the transmission. The receiver reconstructs the video frames by concatenating the data of all the packets.

It is essential to process the packets in real-time to maintain the low latency, and not accumulate them in a buffer, for example, until a complete video frame would be ready. Using a hardware in-line packet engine, it is very easy to execute these tasks with negligible latency.

The transported audio uses much less bandwidth than the video and special care should be taken for the encapsulation in network packets. Indeed, audio samples are grouped in small amount of samples to avoid adding delay at the encapsulation. It is also necessary to keep a relatively small packet time for the clock recovery mechanism to be reliable.
The network infrastructure itself adds its own latency, but this is usually very limited on a local area network (less than a millisecond). Larger networks can also introduce some jitter at the packet level which needs to be properly handled at the receiver side.

In practice, the receiver has a small network packet buffer to compensate for the jitter and the granularity at which the video decoding can be done. This buffer should be configured to the minimum that guarantees the reliable operation of an installation.

How to achieve sub-frame latency with video compression?

There are many different video codecs for different purposes. Selecting the right video codec is always a compromise between the latency, compression ratio and quality. It is not possible to score well in all criteria at the same time. As an example, a codec like h264/h265 used for the video distribution over Internet is optimized to achieve the best compression ratio and image quality at the expense of high latency.

Neighboring pixels need to be involved in order to increase the efficiency of the compression. The pixels can be spatial neighbors (from the same frame), or temporal neighbors (from other previous or next frames). Most advanced video codecs in terms of compression ratio (like h.264/265) are called inter-frame codecs. They take advantage of this principle by using several frames before and after the current frame to encode it. This of course induces several frames of latency. In general, the codec latency is caused by the fact that future pixels are involved in the encoding of the current pixels. Some codecs also require several passes with complex calculations that can also increase the latency depending on the hardware/software implementation used.

Another aspect that affects the latency of the codec is the rate allocation mechanism that is used. The rate control of the encoder regulates the amount of compressed data to achieve the requested target bit rate on average. For a codec to be low latency, it is important that the bitrate is constant (CBR) on a small time window. The time window used for averaging the bitrate is important. For example, a codec could produce a stream at a bitrate that would be constant when averaged over 5 frames, but not necessarily constant when averaged over 1 frame. As the compressed stream is transported over a channel with limited bandwidth capacity, it is then required to use additional buffering and latency to smooth the transmission.

For this reason, a very low latency codec will generate a constant bitrate output when averaged over a few video lines.

Latency measurements of Viper 4K AV over IP

This section shows a practical measurement of a sub-frame latency solution for AV over IP. The equipment used for the measurement is the Viper OEM board from Silex Insight running the VC-2 HQ codec.

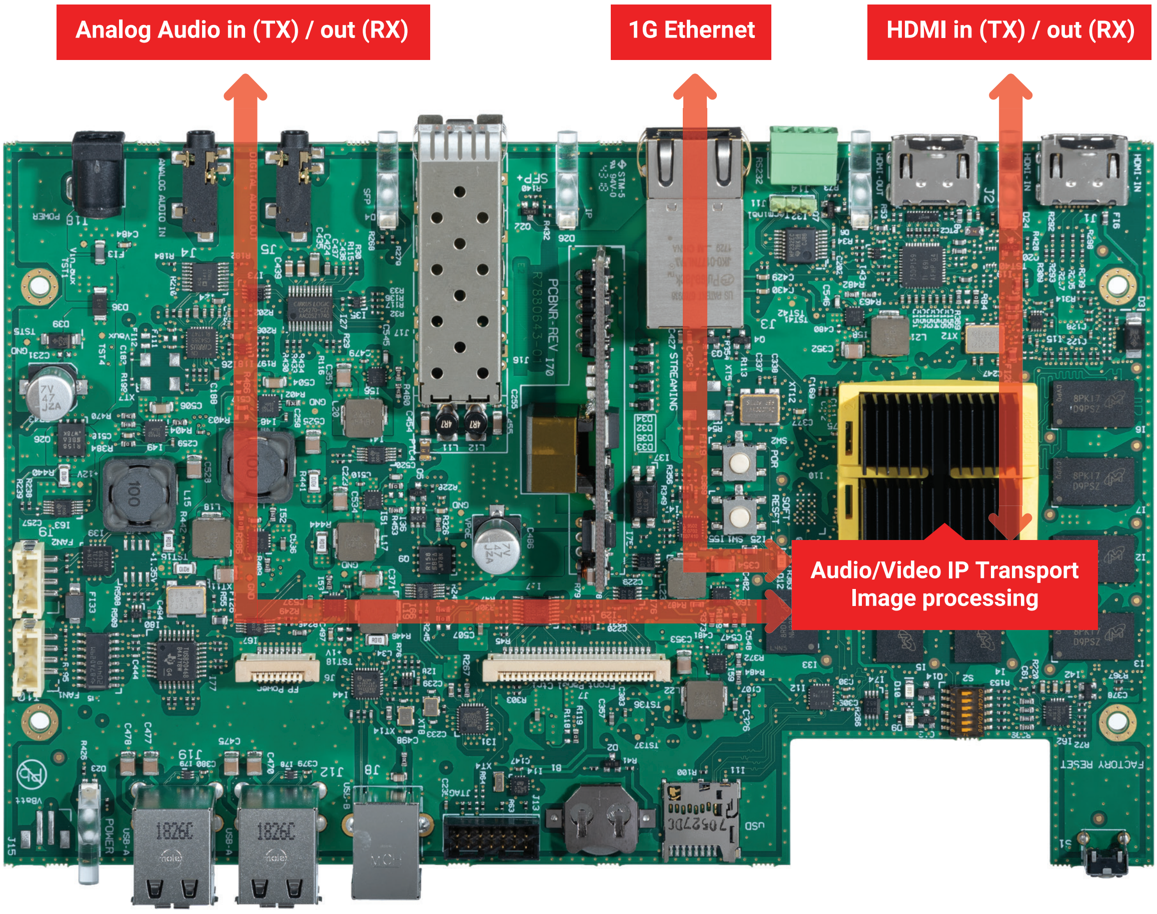

Figure 2 – Viper hardware board with HDMI, analog audio and Ethernet interfaces

Viper low latency architecture

The architecture of the Viper transmitter and receiver has been designed taking into account the principles that were described in the previous sections.

Neither the transmitter nor the receiver stores any significant amount of data (such as full video frame) during the conversion from HDMI to IP and vice versa. The HDMI output clock of the receiver is synchronized to the HDMI input clock of the transmitter over the network in order to avoid any overrun or underrun of the receiver.

On the video encoding side, the VC-2 HQ algorithm is used. VC-2 HQ is a SMPTE standard (SMPTE 2042) ideally suited for low compression ratios (up to 10:1. VC-2 HQ has low complexity, and its line-based wavelet allows a latency of a few video lines only. The ultra-low latency of VC-2 is below the millisecond just like another simple video processing functionality.

Another codec supported on the Viper boards is JPEG 2000. JPEG 2000 is a well-known JPEG standard that can achieve best quality for compression ratio up to 20:1. Full frame encoding/decoding has a latency of a few frames but it is permitted to encode/decode stripes (division of the frame in X lines) to bring the latency down to a few milliseconds.

Both the VC-2 HQ and JPEG 2000 encoder produce a constant bitrate stream when averaged over a few video lines making it ideal for the transport over a fixed bandwidth network. This guarantees that no extra buffering is required during the transport and decoding of the stream.

Latency measurement

The latency of an AV over IP system is typically measured from the HDMI input of the transmitter to the HDMI output of the receiver. The measurement could be done with dedicated test equipment. However, for this whitepaper, we decided to use the embedded features of Viper to make the setup simpler.

The transmitter and receiver are connected to each other via the IP network made up of a 1G switch. The subtlety of this test setup is that it uses the receiver to generate the audio and video HDMI stream that feeds the transmitter input. The latency measurement is done within the receiver device. The device generates a specific audio/video pattern that it is able to detect when it comes back after the transport over IP.

The receiver calculates then the delay between the instant when the pattern is generated on the HDMI output, and the instant it is received in return over the Ethernet interface including the decoding and image processing in the receiver.

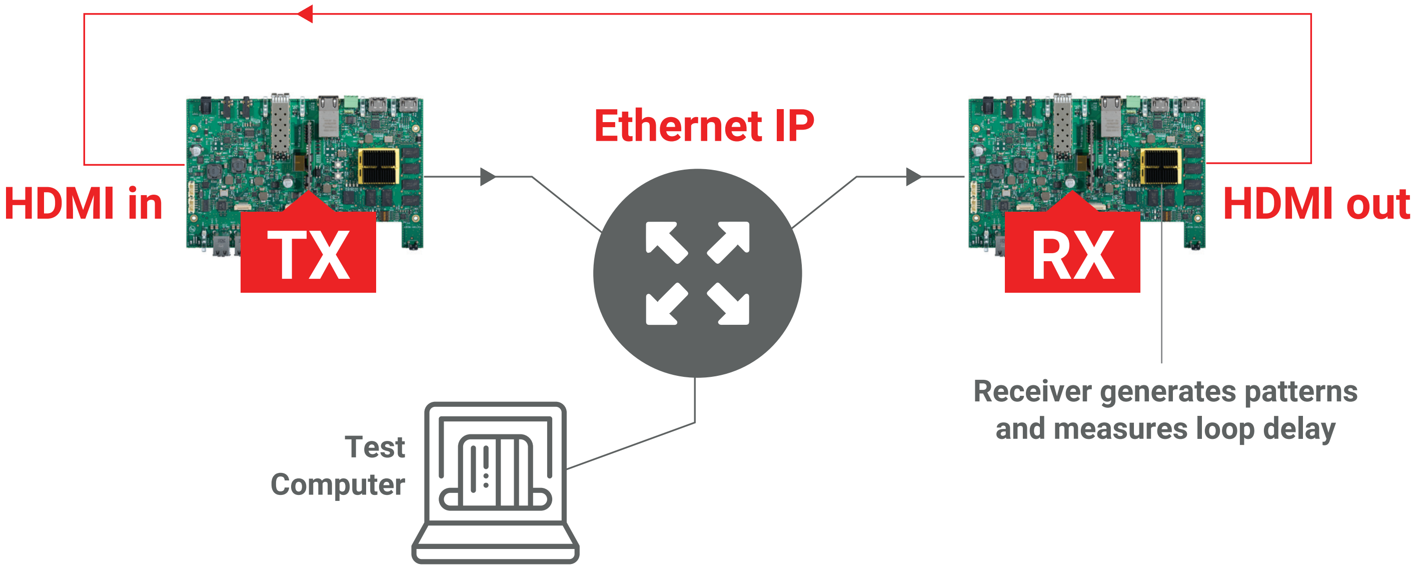

Figure 3 – Test setup used for the latency measurements.

The test setup is used to measure the latency of the audio-video stream in different conditions. The video format can be freely modified, including resolution, chroma subsampling and bit-depth. The target bit-rate of the compressed video can also be changed in order to analyze the impact on the latency. The latency measurements are summarized in the following table:

Test	Video format	Audio format	IP Bandwidth	End-to-end latency
1	1920x1080 60fps 4:2:2 10-bit	2ch PCM 16-bit 48kHz	900 Mbps	3.1 ms
2	1920x1080 60fps 4:2:2 10-bit	2ch PCM 16-bit 48kHz	300 Mbps	5.8 ms
3	3840x2160 30fps 4:4:4 8-bit	2ch PCM 24-bit 96kHz	900 Mbps	3.4 ms
4	3840x2160 60fps 4:2:0 8-bit	2ch PCM 24-bit 96kHz	900 Mbps	3.2 ms

Table 3 Measurements of the total latency of Viper system with VC-2 HQ compression including receiver and transmitter.

The measurements clearly show consistent very low latency around 5 ms or less in all tested use cases. The latency slightly increases with a lower bitrate. This is due to the network stream buffering method implemented at the decoder that guarantees smooth streaming. The buffer size represents a larger video stream duration at lower bitrate. It can also be noticed that the frame rate has a very small impact on the latency, for example, comparing 30 fps vs 60 fps. This is possible because all the video processing and encoding data path only uses very few lines of the video.

Conclusion

This whitepaper has given an overview of the latency challenges when implementing an AV over IP solution. The latency comes from the many processing stages of the system. It is important to take the latency aspect into account from the beginning of the product design and architecture. The video codec, when properly selected, has a very low impact on the latency. When the system is well design, as shown in the Viper case, the latency can be as low as 5 ms and the 4K HDMI 2.0 video fits within a 1G Ethernet cable. This demonstrates the possibility to transport UHD AV content over 1G Ethernet with a seamless user experience.

About Silex Insight

Founded in 1991, Silex Insight is a recognized market-leading independent supplier of two offerings; one is Security IP solutions for embedded systems, while the other offering is custom OEM solutions for AV over IP and video IP codec. The security platforms and solutions from Silex Insight include flexible and high-performance crypto-engines which are easy to integrate and a eSecure IP module which provides a complete security solution for all platforms. For custom OEM solutions for AV over IP and video IP codec, Silex Insight provides high-end image and video compression solutions for distributing low latency, 4K HDR video over IP. Development and manufacturing take place at the headquarters in Louvain-la-Neuve, Belgium and in Ghent, Belgium.

For more information

www.silexinsight.com

Silex Insight

@SilexInsight

SilexInsight

White paper V1.3