Stretch®
Extending the Possibilities® > www.stretchinc.com

The Stretch Scalable Video CODEC

The Trade-offs in Video

CODECs are used to compress video to reduce the bandwidth required to transport the streams, or to reduce the storage requirements to archive them. Successive generations of CODECs have achieved higher compression ratios resulting in smaller streams and file sizes. These improvements have come at a price, however, and that price is the computational requirements to perform the compression. Fixing the tradeoff between bandwidth and computational requirement has the effect of defining the minimum channel bandwidth required to carry the encoded stream and the minimum specification of the decoding device. In traditional video systems such as broadcast television, the minimum specification of a decoder (in this case a set-top box) is readily defined. Video is finding a home in increasingly diverse applications, however. As it does so, it encounters a correspondingly diverse set of client devices on which there is a desire to view it. From computers viewing Internet video to portable digital assistants (PDAs) and even the humble cell phone, device designers are striving to embrace the medium of video. Fixing the requirements for a single CODEC for use with all these devices defines a video stream that is viewable on an entry-level cell phone, but is totally unacceptable on a high resolution computer monitor.

If the video stream is to be made more compatible with a specific viewing device and channel bandwidth, it must be encoded many times with different settings. Each combination of settings must yield a stream that targets the bandwidth of the channel carrying the stream to the consumer and the decode capability of the viewing device. If the original uncompressed stream is unavailable, the encoded stream must be first decoded and then re-encoded with the new settings. This quickly becomes prohibitively expensive.

In an ideal scenario, the video would be encoded only once with a high efficiency CODEC. The resulting stream would, when decoded, yield the full resolution video. Furthermore, in this ideal scenario, if a lower resolution or bandwidth stream was needed to reach further into the network and target a lower performance device, a small portion of the encoded stream would be sent without any further processing. This smaller stream would be easier to decode and yield lower resolution video. In this way, the encoded video stream would be able to adapt itself to the bandwidth of the channel it was required to travel and to the capabilities of the target device. These are exactly the qualities of a scalable video CODEC.

H.264SVC

The Scalable Video CODEC extension to the H.264 standard (H.264SVC) is designed to deliver the benefits described in the preceding ideal scenario. It is based upon the H.264 Advanced Video CODEC standard (H.264AVC) and heavily leverages the tools and concepts of the original CODEC. The encoded stream it generates, however, is scalable temporally, spatially, and in terms of video quality. That is to say, it can yield decoded video of different frame rates, difference resolutions, or different quality levels.

The SVC extension introduces a notion of layers within the encoded stream that was not present in the original H.264AVC CODEC. A base layer encodes the lowest temporal, spatial, and quality representation of the video stream. Enhancement layers encode additional information that, using the base layer as a starting point, can be used to reconstruct higher quality, resolution, or temporal versions of the video during the decode process. In this way, a decoder can produce a video stream with certain desired characteristics by decoding the base layer and the number of the subsequent enhancement layers needed to achieve the desired result. Figure 1 shows the layered structure of an H.264 SVC stream. During the encode process, care is taken to encode a particular layer using reference only to lower level

layers. In this way, the encoded stream can be truncated at any arbitrary point and still remain a valid, decodable stream.

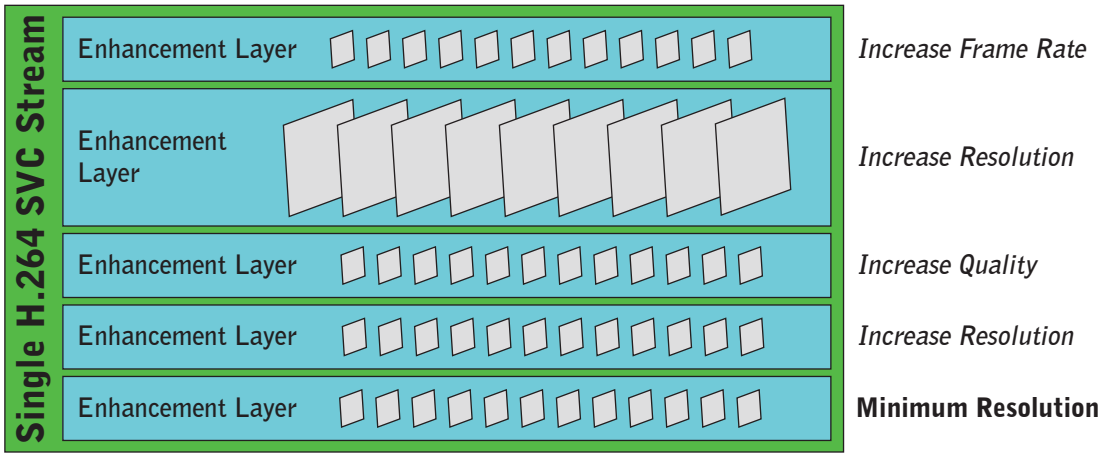

Figure 1: The H.264SVC Layered Structure

It can readily be seen that, by employing this layered approach, an encoded stream can be produced that can be truncated to limit the consumed bandwidth or decode computational requirements. The truncation process consists of simply extracting the required layers from the encoded video stream with no additional processing on the stream itself. The process can even be performed "in the network". That is to say, as the video stream transitions from a high bandwidth to a lower bandwidth network, it could be parsed to size the stream for the available bandwidth. An example might be as a video stream transitions from an Ethernet network through a WiFi link to a handheld device. Here, the stream could be sized for the bandwidth of the wireless link and the decode capabilities of the handheld decoder. Figure 2 shows such an example as a PC forwards a low bandwidth instance of a stream to a mobile device.

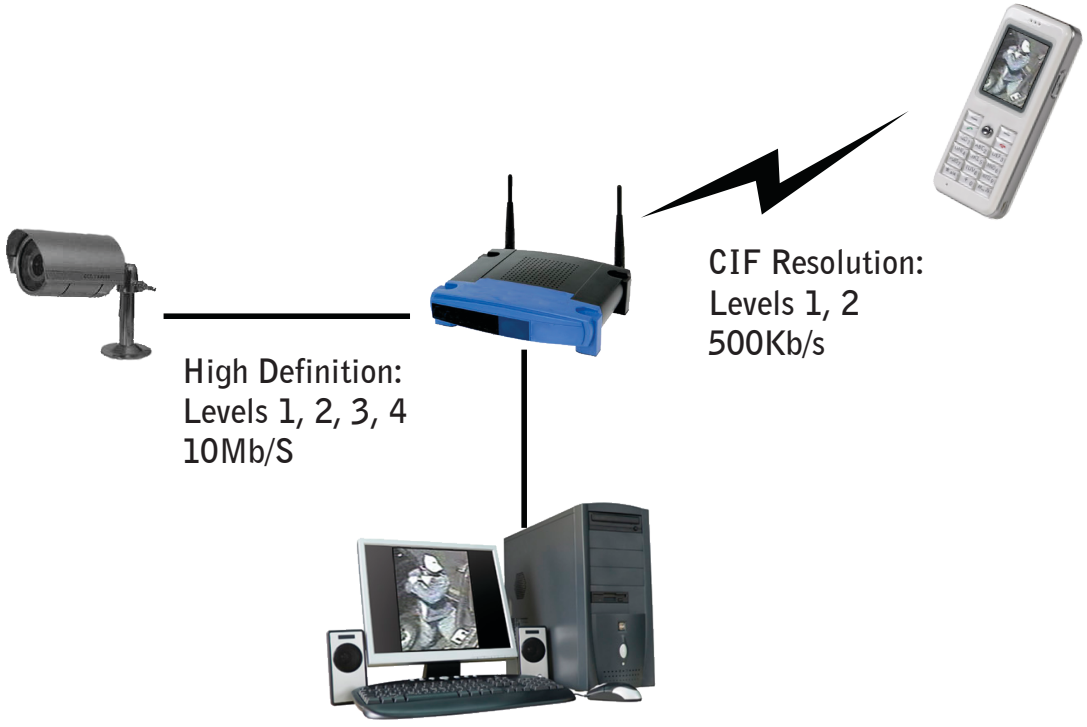

Figure 2: Parsing Levels To Reduce Bandwidth and Resolution

H.264SVC Under The Hood

To achieve temporal scalability, H.264 SVC links its reference and predicted frames somewhat differently than conventional H.264AVC encoders. Instead of the traditional Intra-frame (I frame), Bidirectional (B frame) and Predicted (P frame) relationship shown in Figure 3, SVC uses a hierarchical prediction structure.

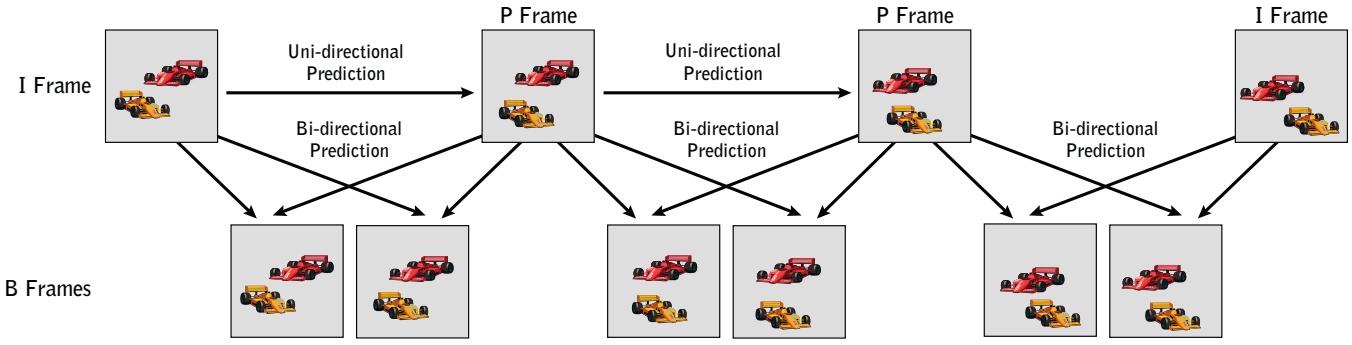

Figure 3: Traditional I, P, and B Frame Relationship

The hierarchical structure defines the temporal layering of the final stream. Figure 4 illustrates a potential hierarchical structure. In this particular example, frames are only predicted from frames that occur earlier in time. This ensures that the structure exhibits not only temporal scalability but also low latency

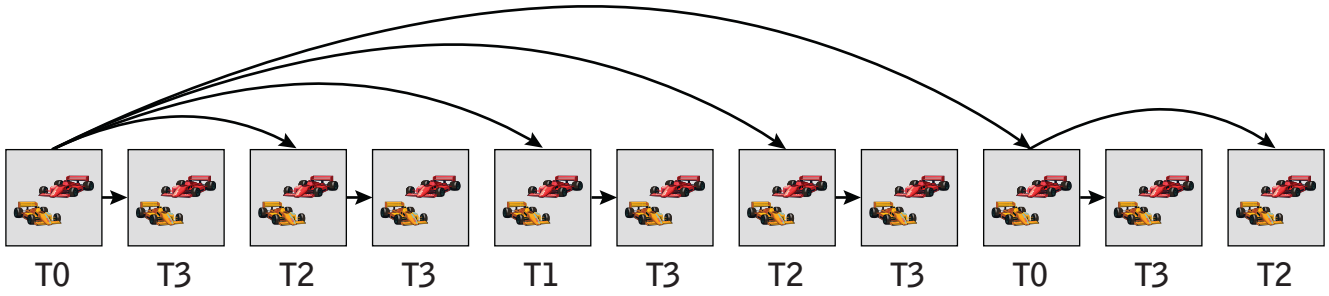

Figure 4: Hierarchical Predicted Frames in SVC

Here we see a scheme where there are four nested temporal layers T0 (the base layer), T1, T2, and T3. Frames that make up the T1 and T2 layers are only predicted from frames in the T0 layer. Frames in the T3 layer are only predicted from frames in the T1 or T2 layers.

To play the encoded stream at 3.75 frames per second, just the frames that constitute T0 need be decoded. All other frames can simply be discarded. To play the stream at 7.5 frames per second, the layers making up T0 and T1 are decoded. Frames in layers T2 and T3 can be discarded. Similarly, if frames that constitute T0, T1 and T2 are decoded, the resultant stream will play at 15 frames per second. Lastly, if all frames are decoded, the full 30 frame per second stream is recovered.

By contrast, in H.264AVC (for Baseline Profile where only unidirectional predicted frames are used), all the frames would need to be decoded irrespective of the desired display rate. To transit a low bandwidth network, the entire stream would need to be decoded, the unwanted frames discarded, and the stream re-encoded.

A similar philosophy is used to create spatial scalability. In this case, lower resolution frames are encoded as the base layer. Decoded and up-sampled base layer frames are used in the prediction of higher order layers. Additional information required to reconstitute the detail of the original scene is encoded as a self contained enhancement layer. In some cases, reusing motion information can further increase encoding efficiency.

Simulcast vs SVC

There is an overhead associated with the scalability inherent in H.264SVC. As can be seen in Figure3, the distance between reference and predicted frames can be longer in time (from T0 to T1 for example) than with the conventional frame structure. In scenes with high motion, this can lead to slightly less efficient compression. There is also an overhead associated with the management of the layered structure in the stream. Overall, an SVC stream containing three layers of temporal scalability and three layers of spatial scalability might be twenty percent larger than an equivalent H.264AVC stream of full resolution and full frame rate video with no scalability. If scalability is to be emulated with the H.264AVC CODEC, multiple encode streams are required, resulting in a dramatically larger bandwidth requirement or expensive decode and re-encode process throughout the network.

Additional SVC Benefits

Error Resilience

Error resilience is traditionally achieved by adding additional information to the stream so that errors can be detected and corrected. SVC's layered approach means that a higher level of error detection and correction can be performed on the smaller base layer without adding significant overhead. Applying the same degree of error detection and correction to an AVC stream would require that the entire stream be protected, resulting in a much larger stream. If errors are detected in the SVC stream, the resolution and frame rate can be progressively degraded until, if needed, the highly protected base layer is used. In this way, the degradation under noisy conditions is much more graceful than with H.264AVC.

Storage Management

The ability of an SVC stream or file to be truncated and remain decodable can be used as a feature after the file is stored, as well as during transmission. By parsing files stored to disk and removing enhancement layers, the file size is reduced with no additional processing on the video stream stored within the file. This would not be possible with an AVC file where an "all or nothing" approach would need to be taken to disk management.

Content Management

The SVC stream or file inherently contains lower resolution and frame rate streams. These can be used to accelerate the application of video analytics or cataloging algorithms. The temporal scalability also makes the stream easier to search in fast forward or reverse.

An Application Case Study

A typical application for H.264SVC is a surveillance system. Consider the case where an IP camera is sending a video feed to a control room where it is stored and basic motion detection analytics are run on the stream. The video feed is viewed at the camera's maximum resolution (1280 x 720) on the control room monitors, and is stored in D1 (720 x 480) resolution to conserve disk space. A first-response team also has access to the stream in the field on mobile terminals within the response vehicle. The resolution of those displays is CIF (352 x 240) and the stream is served at 7 frames per second.

In an implementation using H.264AVC, the first likely constraint would be that the camera serves multiple streams, one at the 1280x720 resolution and one at the 720 x 480 resolution. This places additional cost within the camera, but allows one stream to be directly recorded at the control room while another is decoded and displayed. Without this feature, an expensive decode, resize, and re-encode step would be needed. The D1 stream is also decoded and resized to CIF resolution to feed the video analytics run on the stream. The CIF resolution video is temporally decimated to achieve 7 frames per second and re-encoded so that it can be made available to the first-response vehicle over a wireless link. Figure 5 shows a potential implementation of the system using H.264AVC.

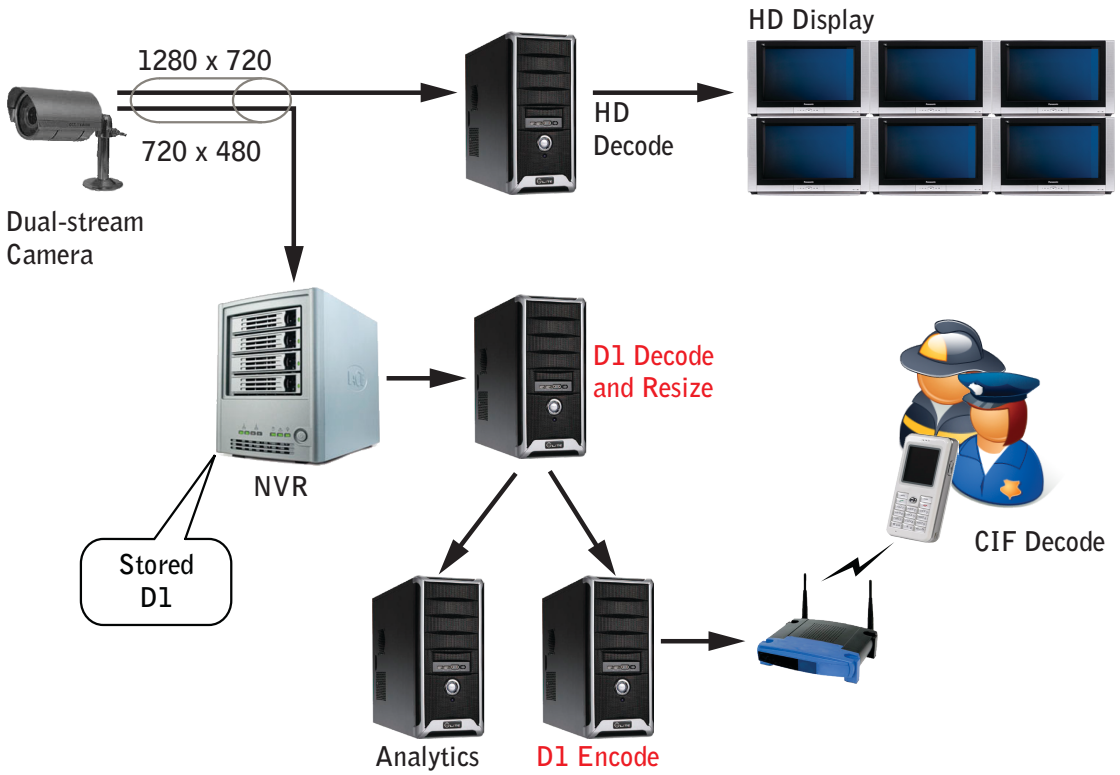

Figure 5: H.264AVC Video Surveillance Application

Using an H.264SVC CODEC, the multi-stream requirement placed on the camera can now be relaxed, reducing complexity and network bandwidth between the camera and control room. The full 1280 x 720 stream can now be stored on the network video recorder (NVR) in the knowledge that it can be easily parsed to create a D1 (or CIF) stream to free up disk space after a specified period. A CIF stream can be served directly from the NVR for analytic work, and a second stream at reduced frame rate can be made available to the first-response vehicle. Figure 6 shows a potential H.264SVC implementation.

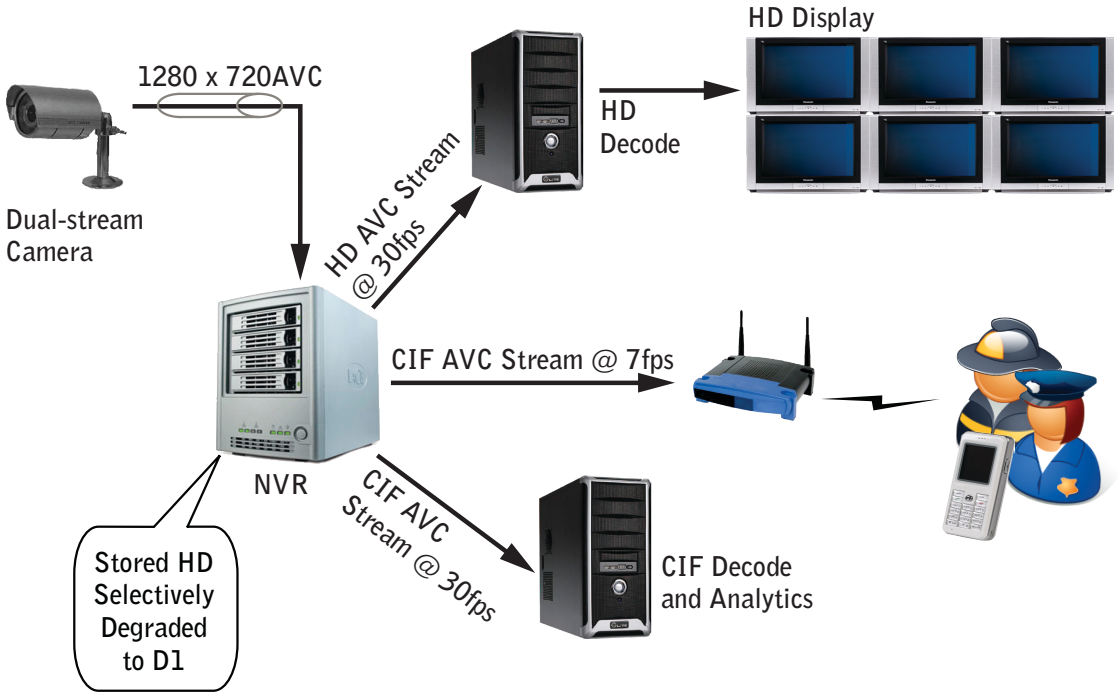

Figure 6: H.264SVC Video Surveillance Application

At no point is there a need to operate on the video stream itself, just the stored file. The advantages are clear::

- Reduced network bandwidth
- Flexible storage management
- Removal of expensive decode and re-encode stages
- High definition video available on the NVR for archive if needed

Conclusion

Scalable video CODECs have been in development for many years. The broadcast industry, strictly controlled by well-established standards, has been slow to adopt the technology. Advances in processor, sensor, and display technology are fuelling an explosion in video adoption. The internet and IP technology is seamlessly serving video to an ever more diverse and ever more remote community of display devices. Scalable video CODECS such as H.264SVC satisfy many of the demands of these systems, and are poised to become the catalyst of the next wave in video applications.

Stretch Inc.
1322 Orleans Drive
Sunnyvale, CA 94089
tel 408.543.2700 • fax 408. 747.5736
www.stretchinc.com