Video is quickly becoming a major component of the enterprise traffic mix. Both streaming and pre-positioned video has implications on the network that can substantially affect overall performance. Understanding the structure of video datagrams and the requirements they place on the network will assist network administrators with implementing a Media Ready Network.

Different Types of Video

There are several broad attributes that can be used to describe video. For example, video can be categorized as real time or pre-recorded, streaming or pre-positioned, and high resolution or low resolution. The network load is dependant on the type of video being sent. Pre-recorded, pre-positioned, low resolution video is little more than a file transfer while real-time streaming video demands a high performance network. Many generic video applications fall somewhere in between. This allows non-real-time streaming video applications to work acceptably over the public Internet. Tuning the network and media encoders are both important aspects of deploying video on an IP network.

H.264 Coding and Decoding Implications

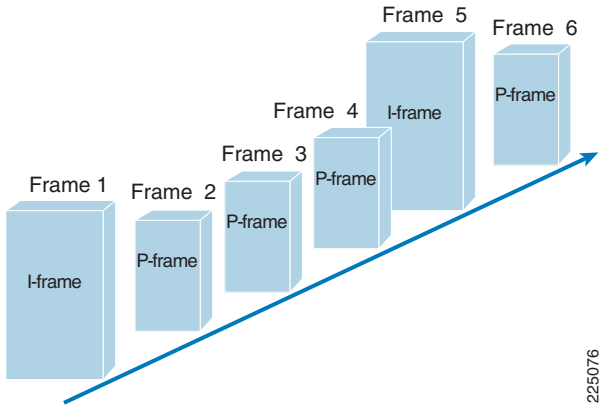
Video codecs have been evolving over the last 15 years. Today's codecs take advantage of the increased processing power to better optimize the stream size. The general procedure has not changed much since the original MPEG1 standard was released. Pictures consist of a matrix of pixels which are grouped into blocks. Blocks combine into macro blocks. A row of macro blocks is a slice. Slices form pictures which are combined into groups of pictures (GOPs).

Each pixel has a red, green, and blue component. The encoding process starts by color sampling the RGB into a luma and two-color components, commonly referred to as YCrCb. Small amounts of color information can be ignored during encoding and then replaced later by interpolatation (??). Once in YCrCb form, each component is passed through a transform. The transform is reversible and does not compress the data. Instead, the data is represented differently to allow more efficient quantization and compression. Quantization is then used to round out small details in the data. This rounding is used to set the quality. Reduced quality allows better compression. Following quantization, lossless compression is applied by replacing common bit sequences with binary codes. Each macro block in the picture goes through this process resulting in an elementary stream of bits. This stream is sliced into 188-byte packets known as a Packetized Elementary Stream (PES). This stream is then loaded into IP packets. Because IP packets have a 1500 byte MTU and PES packets are fixed at 188 bytes, only 7 PES can fit into an IP packet. The resulting IP packet will be 1316 bytes, not including headers. As a result, IP fragmentation is not a concern. An entire frame of high definition video may require 100 IP packets to carry all of the elementary stream packets, although 45 to 65 packets are more common. Quantization and picture complexity are the primary factors in determining the number of packets required for transmission. Forward error correction can be used to estimate some lost information. However in many cases, multiple IP packets are dropped in sequence. This makes the frame almost impossible to decompress. The packets that were successfully sent represent wasted bandwidth. RTCP can be used to request a new frame. Without a valid initial frame, subsequent frames will not decode properly.

Frame Types

The current generation of video coding is known by three names; H.264, MPEG4 part 10, and Advanced Video Coding (AVC). As with earlier codecs, H.264 employs spatial and temporal compression. Spatial compression is used on a single frame of video as described previously. These types of frames are known as I-frames. An I-Frame is the first picture in a GOP. Temporal compression takes advantage of the fact that little information changes between subsequent frames. Changes are a result of motion, although changes in zoom or camera movement can result in almost every pixel changing. Vectors are used to describe this motion and are applied to a block. A global vector is used if the encoder determines all pixels moved together, as is the case with camera panning. In addition, a difference signal is used to fine tune any error that results. H.264 allows variable block sizes and is able to code motion as fine as a ¼ pixel. The decoder uses this information to determine how the current frame should look based on the previous frame. Packets that contain the motion vectors and error signals are known as P-Frames. Lost P-Frames usually results in artifacts that are folded into subsequent frames. If an artifact persists over time, then the likely cause is a lost P-Frame.

Figure 1 P-Frames



H.264 also implements B-Frames. This type of frame fills in information between P-Frames. This means that the B-Frame will need to be held until the next P-frame arrives before the information can be used. B-Frames are not used in all modes of H.264. The encoder decides what type of frame is best suited. There are typically more P-frames than I-frames. Lab analysis has shown TelePresence I-frames to generally be 64K wide (50 packets @ 1316 bytes), while P- frames average 8K wide (9 packets at 900 bytes).

Motion JPEG (MJPEG)

Another type of video compression is MJPEG. Temporal compression is not used with this coding. There are some advantageous and disadvantages. First, the resulting video stream is larger but the packet sizes are more consistent at 1316 bytes (payload). Quantization can be used to mitigate the increased bandwidth but at the cost of picture quality. The advantage that MJPEG offers is that each frame of video is independent of the previous frame. If a several packets from a particular frame are dropped, the artifact is not carried forward. Another advantage is that a single frame is easily extracted from the stream without the need for the reference I-Frame and prior P&B Frames. This is useful in some applications such as video Surveillance where a single frame may be extracted and sent as a JPG via email.

Voice versus Video

Voice and video are often thought of as close cousins. Although they are both real time protocol (RTP) applications, the similarities stop there. They do not even use the same codec to encode audio information. Voice uses G.711 or G.729 while video

uses MP3 or AAC. Voice is generally considered well behaved because each packet is a fixed size and fixed rate. Video frames are spread over multiple packets that travel as a group. Because one lost packet can ruin a P-frame, and one bad P-frame can cause a persistent artifact, video generally has a tighter loss requirement than audio. Video is asymmetrical. Voice can also be but typically is not. Even on mute, an IP phone will send and receive the same size flow. Video uses a separate camera and viewer so there is no assurance of symmetry. In the case of broadcast video, the asymmetrical load on the network can be substantial. Network policies may be necessary to manage potential senders. For example, if a branch is presenting a video, the WAN aggregation router may be receiving more data than it is sending.

Video will increase the average real time packet size, and has the capacity to quickly alter the traffic profile of networks. Without planning, this could be detrimental to network performance.

Resolution

The sending station determines the video's resolution and consequently, the load on the network. This is irrespective of the size of the monitor used to display the video. Observing the video is not a reliable method to estimate load. Common high definition formats are 720i, 1080i, 1080p, etc. The numerical value of the format represents the number of rows in the frame. The aspect ratio of high definition is 16:9 which results in 1920 columns. There is work underway on 2160p resolution and UHDV (7,680 x 4320). This format was first demonstrated by NHK over an IP network and used 600 Mb/s of bandwidth. Video load on the network is likely to increase over time due to the demand for high quality images. In addition to high resolution, there is also a proliferation of lower quality video that is often tunneled in HTTP or in some cases, HTTPS, and SSL. Typical resolutions include CIF (352x288) and 4CIF (704x576). These numbers were chosen as integers of the 16x16 macro blocks that is used by the DCT (22x18) and (44x36) macro blocks respectively.

Table 1

Format	Resolution	Typical BW
QCIF (1/4 CIF)	176x144	260K
CIF	352x288	512K
4CIF	704x576	1 Mb/s
SD NTSC	720x480	Analog, 4.2Mhz
720 HD	1280x720	1-8 mb/s
1080 HD	1080x1920	5-8 mb/s h.264 12+ mb/s mpg2
CUPC	640x480 max	
YouTube	320x240	Flash(H.264)
Skype	Camera limits	128 - 512K+

Network Load

The impact of resolution on the network load is generally a squared term. An image that is twice as big will require four times the bandwidth. In addition, the color sampling, quantization and the frame rate also impact the amount of network traffic. Standard rates are 30 frames per second (actually 29.97) but this is an arbitrary value chosen based on the frequency of AC power. In Europe, analog video is traditionally 25 FPS. Cineplex movies are shot at 24 FPS. As the frame rate is decreased, the network load is also decrease and the motion becomes less life like. Video above 24 FPS does not noticeable improve motion. Finally the sophistication of the encoder has a large impact on video load. H.264 encoders have great flexibility in determining how best to encode video and with this comes complexity in determining the best method. For example, MPEG4.10 allows the encoder to select the most appropriate block size depending on the surrounding pixels. Because efficient encoding is more difficult then decoding, and because the sender determines the load on the network, low cost encoders will usually require more bandwidth then high end encoders. H.264 coding of real time CIF video will drive all but the most powerful laptops well into the 90% CPU range without dedicated media processors.

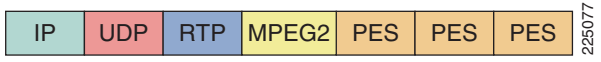
Multicast

Broadcast video lends itself well to take advantage of the bandwidth savings offered by multicast. This has been in place in many networks for years. Recent improvements to multicast simplify the deployment on the network. Multicast will play a role going forward. However, multicast is not used in all situations. Some applications such as multipoint TelePresence use a dedicated MCU to replicate video. The MCU can make decisions concern which participants are viewing each sender. The MCU can also quench senders that are not being viewed.

Transports

MPEG4 uses the same transport as MPEG2. A PES consists of 188-byte datagrams that are loaded into IP. The video packets can be loaded into RTP/UDP/IP or HTTP(S)/TCP/IP.

Figure 2 Transports



Video over UDP is found with dedicated real time applications such as teleconferencing or TelePresence. In this case, a RTCP channel can be setup from the receiver towards the sender. This is used to manage the video session, and is implementation specific. RTCP can be used to request I-frames or report capabilities to the sender. UDP and RTP each provide a method to multiplex channels together. Audio and Video typically use different UDP ports, but also have unique RTP payload types. Deep packet inspection (DPI) can be used on the network to identify the type of video and audio that is present. Note that H.264 also provides a mechanism to multiplex together layers of the video. This could be used to handle a scrolling ticker at the bottom of the screen without sending a continuous stream of motion vectors.

Buffering

Jitter and delay are present in all IP networks. Jitter is the variation in delay. Delay is generally caused by interface queuing. Video decoders can employ a play out buffer to smooth out jitter found in the network. There are limitations to the depth of this buffer. If it is too small, then drops will result. If it is too deep, then the video will be time delayed which could be a problem in real time applications such as TelePresence. Another limitation is handling dropped packets which often accompany deep play out buffers. If RTCP is used to request a new I-Frame, then more frames will be skipped over at the time of resync. The result is that dropped packets have a slightly greater impact in video degradation then they would have if the missing packet was discovered earlier. Most codecs employ a dynamic play out buffer.

Conclusion

A picture is worth a thousand words and video is 30 pictures per second. This can dramatically impact the performance of the network if planning does not properly account for this additional load. Cisco Systems offers guidance to assist customers in implementing a PIN based network platform to ensure video is a successful addition to an enterprise's information domain.