Bandwidth Allocation Strategies for Transporting Variable-Bit-Rate Video Traffic

Marwan Krunz
Department of Electrical and Computer Engineering
University of Arizona
Tucson, AZ 85718
Tel. (520) 621-8731
krunz@ece.arizona.edu

Abstract

Large-scale deployment and successful commercialization of digital video services over computer networks strongly depend on the cost effectiveness of these services. Network bandwidth is one of the major factors that impact the cost of a video service. In this paper, we survey various approaches for reducing the bandwidth requirement for transporting compressed video traffic over high-speed networks.

1 Introduction

Broadband communications networks are expected to support a wide range of multimedia applications, including entertainment video-on-demand (VOD), high-definition TV (HDTV), and multimedia teleconferencing. These applications generate video and audio streams that must be transported in a timely manner to ensure coherent reception and playback at the receiver. Video streams are typically compressed before being transported over a network. Compression is needed to reduce the storage and bandwidth requirements of digital video. Its efficiency, however, depends on the video dynamics as well as the underlying compression technique. For constant-quality video, this means that the encoder will generate a sequence of variable-size compressed frames. When the frame generation rate is constant (e.g., 30 frames/second in the NTSC format), the output of the encoder constitutes a variable-bit-rate (VBR) stream.

Transporting VBR-coded video streams while guaranteeing a required level of quality of service (QoS) is a challenging problem that has received a lot of attention in recent years. Conventional data networks (e.g., the current Internet) have not been designed to transport traffic streams with QoS requirements. While ATM networks are more suited to real-time and guaranteed QoS communications, the complexity of VBR video combined with the diversity of its QoS requirements make it difficult to transport video traffic in a cost-effective manner. This has been a major hindrance facing the economic viability of digital video services over computer networks. Unless efficient bandwidth allocation approaches are devised, the cost of digital video services will prevent their widespread acceptance among potential customers.

The goal of this article is to survey the various approaches that have been proposed for reducing the bandwidth requirement of digital video. After describing the various types of video traffic, we discuss its transport requirements and network support for these requirements in both IP and ATM networks. We then present the three fundamental approaches for bandwidth reduction: statistical multiplexing, temporal smoothing, and multicasting.

2 Video Traffic

Several factors impact the nature of video traffic and its transport requirements. Chief among these are the target quality (constant or variable), the compression technique, the coding time (on-line or off-line), the adaptiveness of the video application, and the supported level of interactivity. These factors have important consequences on the choice of an appropriate transport mechanism for video traffic.

2.1 VBR Versus CBR Coding

As indicated before, the compression efficiency strongly depends on video dynamics. To maintain constant-quality picture, the encoder must generate more bits during high-action scenes than during low-action scenes. Figure 1 illustrates the tradeoff between image quality and the output bit rate for a typical encoder. In this figure, video quality is measured by the distortion in a video frame. Distortion is a measure of the lossyness of the compression algorithm, i.e., the difference between the quality of the original image before encoding and the one obtained after decoding. One common approach to control the level of distortion in a frame is to adjust the quantization values that are used to encode that frame. Note that most video compression algorithms are lossy in nature, so some level of distortion is inevitable.

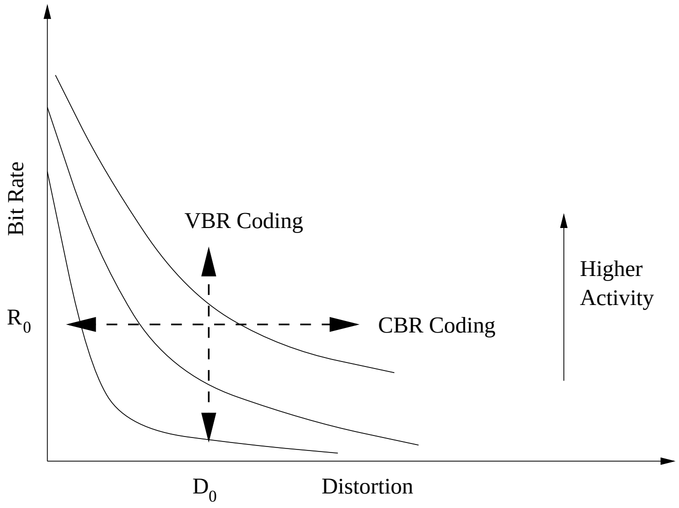

Figure 1: Tradeoff between quality and bit rate in video compression.

The rate-distortion curves in Figure 1 define a region of operation for video encoders. This region is bounded by two orthogonal lines of operation: (1) VBR coding (vertical line), which maintains constant quality throughout the video session, and (2) CBR coding (horizontal line), in which a constant bit rate (or frame sizes) is maintained throughout the video session. VBR coding results in widely varying frame sizes (see Figure 2). From the standpoint of traffic management and resource allocation, CBR (or near-CBR) coding is much easier to handle. Practical video encoders often operate somewhere between CBR and VBR coding, with VBR coding being the choice for high-quality video (e.g., HDTV). Instead of producing a CBR traffic on a frame-by-frame basis, some encoders maintain the CBR on a block-by-block basis, where each block consists of several frames. This reduces the possibility of severe degradation in video quality.

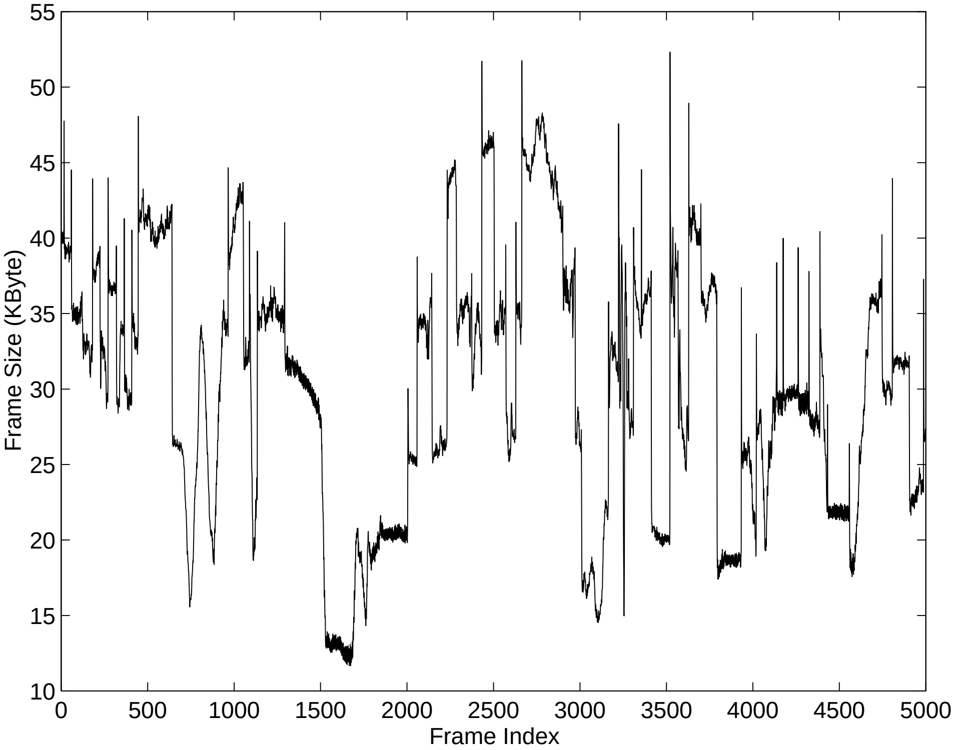

Figure 2: Fluctuations in the bit rate of a VBR-coded video sequence.

2.2 Compression Scheme

Besides scene activity, the VBR behavior of a video source is strongly dependent on the underlying compression technique. Several compression techniques are available today, although many of them are not standardized. In recent years, the standardization committees have been working on providing a set of generic compression standards that can be used for a variety of video applications. These include H.261 for video teleconferencing, JPEG for still images, and MPEG for full-motion video. In particular, MPEG has been widely accepted as the standardized platform for many digital video services, including HDTV, video-on-demand, and multimedia wireless communications (MPEG-4). It enjoys a generic structure whereby various modes of compression can be combined in different ways to achieve different tradeoffs between compression efficiency, encoding/decoding delay, and error resilience. In fact, JPEG can be regarded as a special case of MPEG with intra-coded frames. This justifies the particular emphasis on devising efficient techniques for transporting MPEG-coded video traffic over high-speed networks.

2.3 On-line Versus Off-line Compression

In real-time video applications (e.g., video conferencing), compression is performed on-the-fly while frames are being transported. Resource allocation is more difficult in this case, since frame sizes are not known in advance. In contrast, other applications involve off-line video compression, whereby a video sequence is encoded and stored on an appropriate storage medium, and is later transported over the network. VOD and digital libraries are examples of archived (or prerecorded) applications. Clearly, the availability of the traffic profile of archived video streams makes it easier to develop appropriate bandwidth allocation schemes for this type of traffic.

2.4 Interactivity

Another important issue that impacts the nature and transport requirements of video traffic is user interactivity. Video services differ in their support for interactivity. On one extreme, interactivity is limited to establishing and terminating video sessions. On the other extreme, a video service may offer a full range of VCR-like interactive functions, such as fast-forward, rewind, pause, jump, etc. Interactivity brings with it a host of issues in terms of transport and decoding demands. Some forms of interactivity are relatively easy to accommodate, since they do not require more bandwidth than normal playback (e.g., a stop, jump, or pause function followed by resume). In contrast, fast-scanning (FS) functions, which involve displaying frames at several times the normal rate, impose huge demands on network bandwidth and decoding speed. Backward FS is even more difficult to support in compression schemes that involve motion interpolated frames such as MPEG B frames, where all the reference frames in a Group of Pictures (GOP) must be decoded before B frames of that GOP can be played back in the reverse order. The large demand for network bandwidth in fully interactive video services is compounded by the lack of prior knowledge about the pattern of interactivity that is to be expected from a user.

2.5 Adaptiveness

Some video applications are designed to be adaptive to fluctuations in network performance. These applications require little QoS support from the network. Based on explicit or implicit feedback from the network, the application continuously adjusts its bit rate by varying video quality. Different approaches can be used to reduce the bit rate of the encoder. For example, the encoder could vary the quantization factors that are used in frame encoding. It could also reduce the rate at which frames are generated. Some compression techniques (e.g., MPEG-2) provide several modes of scalability that can be exploited in rate adaptation. It should be noted that high-quality video applications are inherently non-adaptive, as they often require stringent, deterministic QoS guarantees.

3 Quality of Service in Video Communications

QoS is a generic notion that can be defined at various levels in the protocol stack, particularly the application and network levels (and sometimes the network adaptation level). Application-level QoS is essentially visual and is hard to measure in an objective manner [6]. It depends on the viewer's sensitivity to glitches in the video (which can be caused by lost packets), variability in the frame rate, variability in the distortion factor, etc. Assuming that application-level QoS requirements have been determined in one way or another, the next step is to map them into network-level QoS requirements, which are specified in terms of throughput, delay, and packet (or cell) loss.

Throughput Requirement

Video is produced by displaying frames at a fixed frame rate known as the playback rate. This rate varies from one video format to another. Several standardized video formats are available, including NTSC (30 frames/second) and PAL (25 frames/second). To ensure continuous streaming of video, the rate at which frames are transported over the network must, on average, be no smaller than the playback rate. For VBR-coded video, this means that the minimum required throughput is given by the mean bit rate, which is known for many compression techniques (e.g., 1.5 Mbps for MPEG-1 and 5 Mbps for MPEG-2). However, because of the fluctuations of the bit rate, the actual throughput requirement is typically higher than the mean rate, and it depends on many factors, including the degree of smoothing at the end systems and in the network, whether statistical multiplexing is involved, and the stringency of the delay and loss requirements of the video stream. For CBR-coded video, the throughput requirement is simply given by the constant bit rate, which depends on the target level of distortion. This rate is often greater than the mean bit rate that is obtained from VBR coding of the same video.

Delay Requirements

Interactive communications have stringent requirements in terms of the maximum delay and delay variation (jitter). Acceptable values for the one-way delay lie in the range 150 to 400 msec (one way) [6]. For the jitter, the requirement depends on the amount of smoothing at the sender and the receiver. Assuming a constant playback rate of f frames/sec and no smoothing, the receiver requires a frame every $1/f$ seconds (33 msec if $f = 30$). In addition, for synchronization between video and audio, it requires that the transfer delays of video and audio data are within 80 msec for one-way communications [6].

Loss Requirements

In packet video, losses are primarily caused by buffer overflow in the network. These losses may have considerable impact on the perceptual quality, and must thus be kept under control. The acceptable loss rate depends on several factors, including the compression scheme, the duration of the loss interval, the relative importance of the lost data, and the error concealment mechanism (if any). In general, the loss requirement at the network level is expected to be in the range 10^{-2} to 10^{-6}. The network may provide a video stream with multiple loss rates, depending on the importance (i.e., priority) of the transmitted data.

4 Network Support

4.1 ATM Networks

ATM provides several options for transporting video traffic. Of the five network services specified by the ATM Forum, two can be readily used for video transport: the CBR and real-time VBR (rt-VBR) services. The CBR service provides a constant-bandwidth pipe that can be used to support stringent QoS guarantees on the cell loss rate (CLR), the maximum cell transfer delay (maxCTD), and the cell delay variation (CDV). Resources are dedicated for each connection, ensuring deterministic QoS guarantees (statistical multiplexing is not used). The rt-VBR service provides guarantees on the same QoS parameters. But since it optionally uses statistical multiplexing, the resulting guarantees are probabilistic in nature. For adaptive video applications, some researchers suggested using the available-bit-rate (ABR) service, which guarantees a minimum cell rate (MCR) in addition to a CLR [9]. It is assumed in this case that video sources can adapt their rates according to network feedback that is is conveyed using an explicit-rate-based congestion mechanism. Rate adaptation is performed by adjusting the quantization values in the encoding algorithm at the expense of variable-quality video. To limit the variation in the quality, the ABR-based video transport mechanism can be enhanced by adding a smoothing function at the source.

VBR-coded video is known to exhibit multiple-time-scale variations. In particular, the bit rate is strongly modulated by scene changes, which occur at a time scale of several seconds (hundreds of frames). As can be observed from Figure 2, the bit rate fluctuates in small amounts within some neighborhood that corresponds to a scene. Between scenes there is a significant change in the bit rate. To accommodate such behavior, some researchers have suggested introducing a Renegotiated CBR (RCBR) network service [5], in which the bandwidth for a connection can be renegotiated several times during the lifetime of the connection. If supported by the standards, this approach promises to provide CBR-like service at a much lower cost. Naturally, RCBR will require a more complicated traffic management than a typical CBR service.

4.2 IP Networks

The IP-based Internet has not been designed to support QoS guarantees, simply because the original Internet applications (e.g., email and FTP) are data oriented, with little or no need for stringent guarantees. In the new multimedia era, this situation is rapidly changing. There is a growing interest within the telecommunications industry in providing voice and video services over IP networks. This trend is paralleled by a phenomenal growth of the World Wide Web (WWW), with voice and video being further integrated into the design of Web pages. Given these trends, the Internet Engineering Task Force (IETF) has been exploring various possibilities for supporting voice and video over IP.

4.2.1 Real-Time Protocol (RTP)

RTP is a first step towards supporting voice and video over the Internet. It is a session layer protocol that runs on top of UDP, interfacing it with the audio or video application. This means that RTP is transparent to network routers, and cannot be used to support network-level QoS guarantees. However, it provides several functions that are useful for real-time communications, including sequence numbering, timestamping, and payload type identification (i.e., type of video or audio encoding). A sampled voice or video signal is encapsulated into an RTP packet, which is then encapsulated into a UDP packet. Lost RTP packets are not retransmitted. The receiver can detect these losses based on the sequence numbers of the received RTP packets. There are various ways for assigning RTP *streams* to media sources. For example, in a teleconferencing application, one may assign an RTP stream to each voice or video source in each direction. Alternatively, when voice and video are bundled (as in the case of MPEG), both sources are assigned one RTP stream in each direction.

The timestamp field in the header of an RTP packet contains the relative sampling time of the video or voice signal. The time unit is taken as the inverse of the sampling rate (e.g., frame rate for video). Thus, if every N samples are sent in one RTP packet, then the timestamp advances in increments of N. The timestamp continues to be incremented even when the source is idle. This way, the receiver can determine the correct playback time of the packet relative to the most recent packet. By comparing this time to the actual interarrival time, network jitter can be computed. Negative jitter (i.e., packet arriving before its playback time) can be removed by means of buffering. RTP relies on the real-time control protocol (RTCP) to convey various types of information, including the number of transmitted packets and the number of lost packets. This information can be used by senders to adjust their compression parameters, reducing the bit rate if necessary. In other words, RTP is best suited for adaptive video applications.

4.2.2 Integrated Services (Intserv)

One of the notable efforts to support QoS over the Internet is being undertaken by the Integrated Services Working Group of the IETF. This Working Group has developed a framework for *Integrated Services* (Intserv) that can support some form of QoS guarantees over the Internet. As of the writing of this paper, the Intserv framework (which initially consisted of five services) consists of three service offerings: Guaranteed Service, Controlled Load Service, and the conventional best effort service. Of these three services, the Guaranteed Service is best suited for video transport. In this service, each traffic flow is characterized at the entry to the network by five parameters. Two of these parameters define a token bucket (a token generation rate and a bucket size). The other three parameters are the peak rate, the maximum datagram size, and the minimum policed unit. A traffic flow provides these parameters and requests a certain amount of bandwidth. By guaranteeing this bandwidth, the network guarantees an associated maximum packet delay (packet loss is guaranteed to be zero).

4.2.3 Differentiated Services (Diffserv)

The Guaranteed Service relies on a reservation protocol such as RSVP to signal reservation requests and to coordinate between network elements. Recently, there has been some concern regarding the scalability of RSVP to the large, distributed Internet. In particular, the need to maintain the soft reservation state of each RSVP session at each router represents a significant overhead for routers at the core of the Internet. Accordingly, the IETF has recently started to investigate another framework for QoS based on the *Differentiated Services* (Diffserv). In contrast to the Intserv model, Diffserv does not require a reservation protocol. It relies on bilateral agreements between ISPs and between an ISP and an end system to provide QoS guarantees. Classification of packets is performed based on the Differentiated Services (DS) field in the header of an IP packet (the TOS octet in IPv4 and the Traffic Class octet in IPv6). It is not clear yet whether the Diffserv model can be used to support video services over the Internet.

5 Reducing Bandwidth Requirements

The viability and economical feasibility of packet video are largely contingent upon the ability to reduce its bandwidth requirement. For this reason, an extensive amount of research is being conducted on mechanisms for reducing the bandwidth requirement of video traffic. Invariably, the approaches that have been proposed rely on one or more of the following strategies:

1. Statistical multiplexing.

2. Temporal smoothing.

3. Multicasting.

5.1 Statistical Multiplexing of Video

Statistical multiplexing (SM) is a mechanism for reducing the bandwidth requirement of bursty and VBR traffic sources. It has been used for decades over the Internet to improve network utilization, but without providing any performance guarantees [6]. It is being used in ATM networks in conjunction with QoS guarantees. In essence, SM is a spatial aggregation mechanism by which several individual streams are asynchronously superposed and transported over the same channel. The resulting aggregate traffic exhibits smoother bit rate behavior (i.e., less variability) than the original streams. Bandwidth is allocated to the aggregate traffic, resulting in a reduction in the per-stream allocated bandwidth. This reduction is proportional to the burstiness of the multiplexed sources.

In addition to its use inside network switches and routers, SM can also be employed at end systems. For example, a remote video server may use SM to maximize the number of simultaneous video connections transported from the server to a head-end (HE) switch over a fixed-bandwidth pipe (Figure 3). In this scenario, multiplexing is performed among streams that are destined to the same HE switch, which in turn demultiplexes the traffic into several streams.

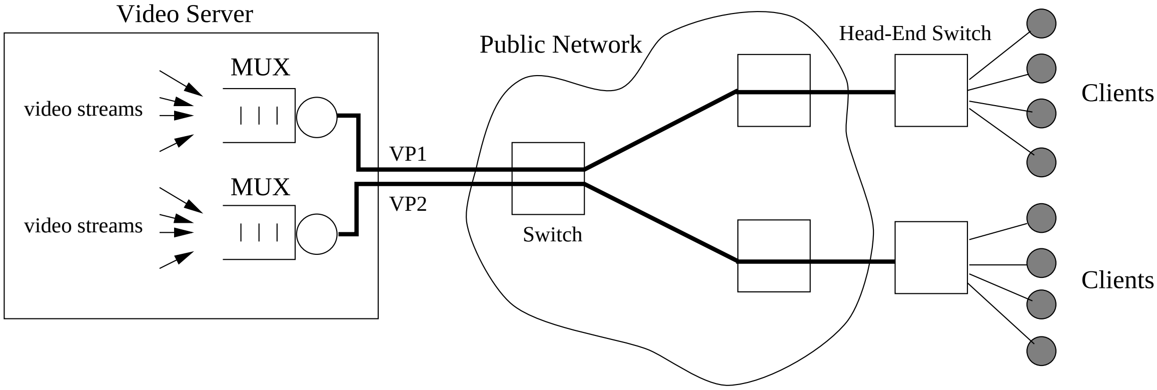

Figure 3: An example of statistical multiplexing at a video server.

Other scenarios for SM are possible. For example, the distribution network may consist of several local video servers in addition to a central remote server that maintains a repository of compressed videos. A subset of these videos are periodically multicast to the local servers. Each local server connects to a HE switch that serves a pool of clients. In this scenario, SM can be performed over the high-speed link connecting the remote server to each local server, and also over the link between a local server and the associated HE switch.

5.1.1 SM with Statistical Guarantees

Three different approaches can be used to provide statistical guarantees through SM. They differ in the type of the source model under which SM gain is evaluated.

First Approach

The first, and most common, approach relies on stochastic source models. This approach is useful in assessing the transport performance and determining the required resources for *real-time* video, whose traffic profile is not known at the time of admission control. In this case, video sources are characterized by *detailed* stochastic models that describe the fluctuations of the bit rate. Bandwidth gain is attained by allowing the sum of the peak rates of the input streams to exceed the service rate of the multiplexer. This results in cell queueing and possible buffer overflow, the amounts of which are determined by evaluating the queueing performance under the presumed traffic model. An extensive amount of literature has been reported on the modeling of VBR-compressed video (see [1, 4] for surveys). These models can be classified according to the persistence of their autocorrelation structure into:

1. Renewal models (no correlations).
2. Markovian and autoregressive models (exponentially decaying autocorrelations).
3. Subexponential models (the decay of the autocorrelations is slower than exponential but faster than a power function).
4. Long-range dependent (LRD) models (autocorrelations decay as a power function).

From the queueing analysis of a video traffic model, the next step is to determine the minimum amounts of bandwidth and buffer that are needed to guarantee a certain level of QoS (i.e., "effective" bandwidth and buffer). Usually, this is done numerically, although in certain cases one may provide explicit expressions for the effective bandwidth under a video model.

Despite the amount of gain that can be achieved from SM under stochastic models, there are still a number of obstacles that limit the applicability of this approach. First, it is currently not possible to apply this approach in a multi-hop scenario due to the difficulty of characterizing the departure traffic from a statistical multiplexer. Second, many video models are too involved to lend themselves to queueing analysis, hence limiting the applicability of the model to off-line dimensioning problems (which can be treated via simulations). Third, when queueing analysis is feasible, it is typically done on an asymptotic basis, i.e., over an infinite-time horizon and for very large buffers. It is still questionable whether the asymptotic results would apply to the more realistic finite-horizon connection holding times. Finally, accurate video models often require specifying more traffic parameters than what is currently supported by the standards. The use of video models in on-line traffic control is still an open research issue.

Second Approach

Instead of detailed stochastic models, video sources can be specified by *stochastic bounds*. For example, if $X(t)$ is the number of cell arrivals in any interval t, then a stochastic bound could be given by $\Pr[X(t) > \theta_t] = \epsilon_t$ for several values of t, θ_t, and ϵ_t. Aside from these bounds, no assumptions are made on the actual arrival pattern. The end-to-end guarantees are obtained by first obtaining stochastic bounds on the traffic at the edge of the network, which are then used to bound the departure traffic at that node. In turn, the bounded departure traffic of one node is used to bound the arrival traffic at the next node, and the procedure is repeated for all nodes along the path. A drawback of this approach is that the bounds become loose as more nodes are traversed.

Third Approach

In contrast to the above two approaches, a video source in this approach is characterized by a deterministic time-invariant traffic envelope, such as the D-BIND envelope that was proposed in [7]. When several of these traffic envelopes are statistically multiplexed, the multiplexing gain depends on the relative phase shifts of these envelopes. Since the envelopes are time invariant, their phase shifts can be assumed to be random, giving rise to statistical QoS guarantees.

5.1.2 SM with Deterministic Guarantees

A different application of SM is described in [8] for prerecorded MPEG video streams. In MPEG compression, three types of compressed frames can be generated: Intra-coded (I), Predictive (P), and Bidirectional (B) frames. On average, I frames are larger than P frames which, in turn, are larger than B frames. Frame types in an MPEG sequence are determined according to the Group of Pictures (GOP) pattern, which is applied repetitively during the encoding of the MPEG sequence. This pattern defines the sequence of P and B frame types between two successive I frames. Typically, the GOP pattern exhibits a regular structure that can be specified by two parameters: the number of frames between two successive I frames (N), and the number of frames between an I frame and the subsequent P frame in a GOP (M). For example, in the GOP pattern `'IBBPBBPBBPBB'`, $N = 12$ and $M = 3$.

The repetitive application of the GOP pattern during the encoding process produces periodic traffic behavior that can be exploited in resource allocation. In particular, each MPEG stream can be characterized by a periodic *time-varying* traffic envelope that is specified by five parameters: the largest frame in the sequence (I_{max}), the largest P frame (P_{max}), the largest B frame (B_{max}), N, and M. Figure 4 depicts an example of the MPEG traffic envelope with $N = 6$ and $M = 3$. More sophisticated variants of this envelope can also be constructed (see [8] for details).

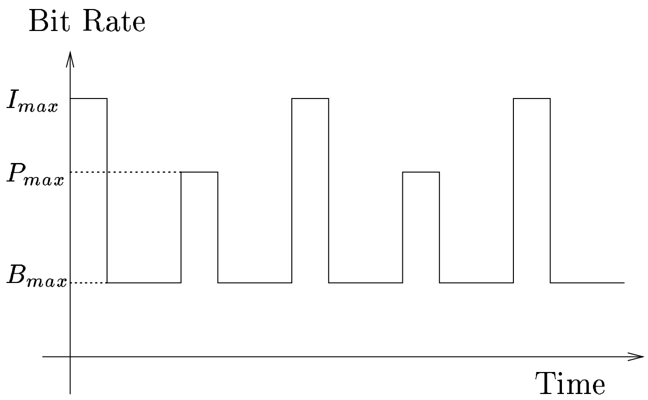

Figure 4: Traffic envelope with $N = 6$ and $M = 3$.

Under the traffic envelope characterization, MPEG sources can be statistically multiplexed at the video server and transported to a HE switch with minimal bounded delay and no losses. The server maintains an active table of the traffic envelopes of all ongoing sources. From these envelopes, the server computes and maintains an "aggregate" envelope that represents a time-varying bound on the bit rate of the aggregate traffic. This envelope is updated when a new stream is added or an ongoing one is terminated. The server allocates bandwidth to the multiplexed traffic based on the peak value of the aggregate envelope. By manipulating the phase shifts between the MPEG sources (through appropriate stream scheduling), the allocated bandwidth can be much less than the sum of the source peak rates. To illustrate the idea, we consider a simple example of two multiplexed sources (Figure 5). We assume that the two sources have the same traffic envelope with $N = 6$ and $M = 3$. However, they differ by a phase shift of one frame period, implying that I frames of one source overlap in time with B frames of the other source. Assume that frames are transported at a constant frame rate and that the bit rate during one frame period is constant (given in the number of ATM cells per frame period). The peak rate of the aggregate envelope, taking the phase difference between the two sources into account, is given by $I_{max} + B_{max}$. The per-stream allocated bandwidth (PSAB) is given by $C = (I_{max} + B_{max})/2$, which is much less than the source peak rate I_{max}. By allocating this amount of bandwidth on an end-to-end basis (i.e., over the pipe from the server to the HE switch), the end-to-end delay is guaranteed to be less than $1/C$ frame periods. Since frames are delivered at about the same playback rate, no smoothing is needed at the receiver.

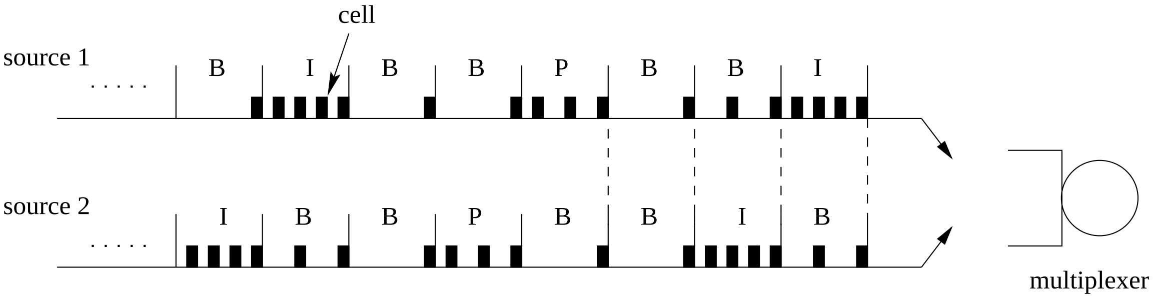

Figure 5: Example of statistical multiplexing of two MPEG sources.

For a given set of MPEG streams, the PSAB is a function of their relative phases, which in turn depend on the relative starting times of the MPEG streams. Let $\boldsymbol{u}$ be the vector of relative phases of n multiplexed streams. This vector, which is known as the *arrangement*, completely specifies the synchronization structure of the multiplexed MPEG streams with respect to their GOPs. Thus, $\boldsymbol{u}$ can be optimized by allowing the server to control the starting times of new streams for the purpose of minimizing the PSAB. This type of stream scheduling comes at the expense of delaying the initiation of a new stream by a maximum of a GOP period (1/2 second). Let $C(\boldsymbol{u}, n)$ be the PSAB for n multiplexed streams with arrangement $\boldsymbol{u}$, and let $C_{min}(n)$ be the minimum PSAB over all possible arrangements of these n streams. An optimal arrangement $\boldsymbol{u}^*$ is the one for which $C(\boldsymbol{u}^*, n) = C_{min}(n) = \min_u C(\boldsymbol{u}, n)$. For homogeneous multiplexed streams (i.e., identical envelopes), a closed-form expression for the optimal *arrangement* $\boldsymbol{u}^*$ was provided in [8]. However, no tractable expression is available for the optimal *arrangement* of heterogeneous streams. Instead, a computationally feasible suboptimal scheduling scheme, known as Minimal-Rate Phase (MRP), was provided for the heterogeneous streams case. According to MRP scheduling, given n multiplexed streams, a new stream (the $(n+1)$th) is scheduled for multiplexing in a phase for which the aggregate bit rate is minimal. The MRP policy is described in Figure 6. It has been shown that as n increases, the PSAB that can be achieved using MRP scheduling approaches $C_{min}(n)$ (i.e., MRP scheduling is asymptotically optimal).

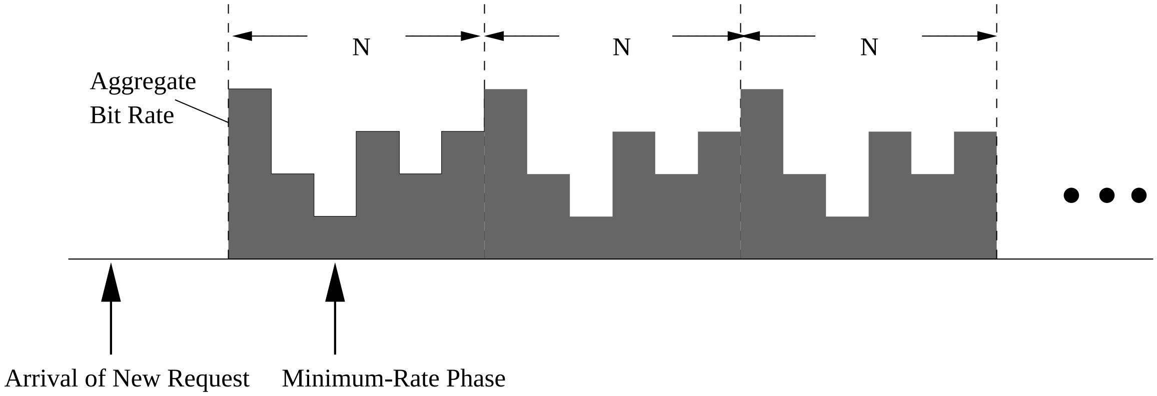

Figure 6: Minimal-Rate Phase (MRP) scheduling for heterogeneous envelopes.

It should be noted that even if no scheduling is performed, some bandwidth gain can still be realized by multiplexing MPEG streams under time-varying traffic envelopes. In this case, $\boldsymbol{u}$ has an arbitrary structure, which is determined only by the times of video requests and are not controlled by the server.

5.2 Temporal Smoothing

The variability in video traffic can also be reduced by means of temporal smoothing on a stream-by-stream basis. The general idea in video smoothing is to introduce a buffer in the path of the stream, either at the sender or at the receiver. Smoothing at the sender is typically used for real-time video, which requires stringent delay and delay jitter guarantees. Smoothing at the receiver is often used for archived video, with the buffer being placed inside the client's set-top box. Either way, the smoothing buffer acts as a low-pass filter by averaging the bit rate over a time window whose length is determined by the size of the buffer and its drain rate. It should be mentioned that smoothing of traffic streams (including video) is frequently used inside the network as part of various control functions, including policing and shaping.

For archived video, the availability of the traffic profile (i.e., frame sizes) makes it possible to combine video smoothing with prefetching (or *working-ahead*), for the purpose of achieving optimal smoothing. Video frames are transported to the client prior to their playback times. A *transmission schedule* is used for this purpose, which consists of a set of successive fixed rates. The client maintains a buffer that temporally stores (and smoothes out) the received frames. This buffer is drained at a constant frame rate. A build-up delay is needed to accumulate a certain amount of video data in the set-top buffer before the commencement of the movie.

A significant aspect of video smoothing research deals with the design of an appropriate transmission schedule that ensures that starvation (underflow) and overflow of the set-top buffer will not occur. Consider a VBR coded video sequence that consists of N frames. Let f_i be the size of the ith frame, and let B be the size of the buffer at the set-top box (in bytes). Assume that the client drains this buffer at a constant frame rate. The time unit is taken as a frame period. Let $X(t)$ be the total amount of video data that the client receives up to time t. To prevent starvation at the client buffer, one must ensure that $X(t) \geq \sum_{i=1}^{t} f_i$ for all $t = 1, 2, \ldots, N$. Similarly, to avoid buffer overflow at the client, we must have $X(t) \leq B + \sum_{i=1}^{t} f_i$ for all $t = 1, 2, \ldots, N$. The above two bounds define a region of feasible transmission schedules, as shown in Figure 7. The next step is choose an appropriate schedule that is optimal in some sense. Several criteria for optimality have been used, including minimizing the peak transmission rate, the number of rate changes, the variance of rate changes, and the startup delay [3].

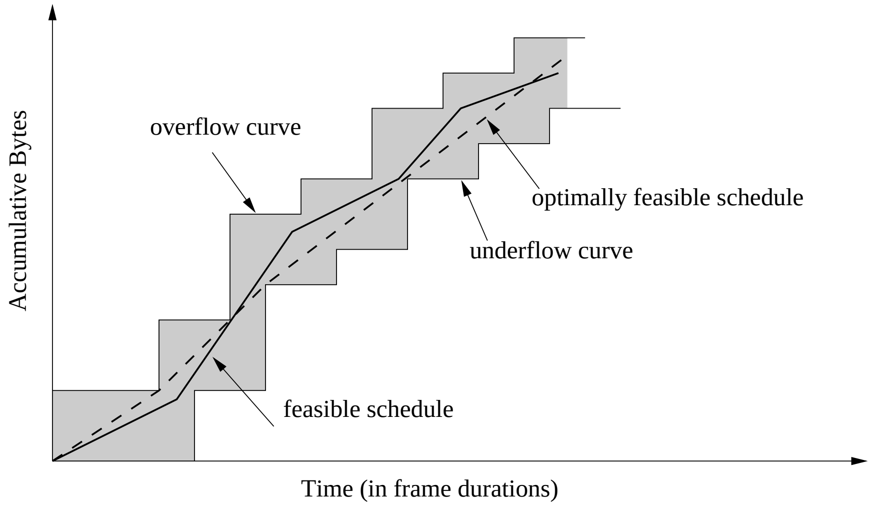

Figure 7: Feasible transmission schedule in video smoothing.

JSQ Prefetching

Join-the-Shortest-Queue (JSQ) is yet another interesting approach for prefetching archived, VBR-coded video [10]. It assumes the existence of a single shared link between the server and VOD clients. This link typically extends from the server to a HE switch. A small buffer is needed at the set-top box to temporarily store prefetched frames. The principal idea in JSQ prefetching is to keep track of the number of frames that has been transmitted to each client. Using this information, the server exploits the VBR nature of video by sending different numbers of frames to different clients. The server selects the prefetched frames based its knowledge of the queue sizes at various clients, favoring clients with short queues (hence, the name JSQ.) In some sense, the JSQ policy creates a pooling effect in which clients collaborate to maximize the utilization of the shared link without causing buffer overflow or underflow of their buffers. JSQ can be extended to support certain interactive functions, including pause/resume and fast jump. Under ideal conditions, the JSQ approach can achieve up to 100% utilization of the shared link.

5.3 Video Multicast

In broadcast-type video applications (e.g., HDTV, distance learning), the per-stream bandwidth requirement can be significantly reduced be means of multicasting. Multicasting is particularly efficient for video applications in which a large number of recipients request a small number of videos. To support VOD services, the server must simultaneously multicast multiple copies of each movie that are staggered in time (i.e., they have different logical playback times). A client that requests a movie has to wait until the start of one of these staggered copies.

Multicasting provides limited support for interactive VOD functions. Near-interactive jump and pause/resume operations can be supported by a combination of multicasting and prefetching [2]. When a jump operation is requested (say, in the forward direction), the client is moved to a different multicast group with a logical time that is closest to the logical time of the requested jump. Since the number of the different instances of the same movie is limited, this scheme can only support "discontinuous" jumps. When a pause is requested, the client continues to receive and buffer frames but without displaying them. If the pause period reaches the phase difference between two successive staggered copies of the movie, the buffer is flushed and the new instant is used to feed the buffer. Following a resume request, the client proceeds to read frames from the head of the buffer, while simultaneously receiving frames from the network and storing them at the back end of the buffer. Fast-forward with scanning is limited by the number of frames that are available in the set-top buffer.

6 Conclusions

Various factors impact the characteristics and requirements of video traffic, including the target quality, the compression scheme, client interactivity, and the adaptivity of the video application. These factors influence the choice of the network transport service. In ATM networks, both CBR and rt-VBR services can be readily used to support video traffic. The ABR service can also be used with adaptive video applications. In IP networks, the IETF has been working on a new set of Internet services, some of which can be used for video transport.

Efficient bandwidth allocation for video traffic is crucial to maintaining the cost effectiveness of a video service. In this paper, we gave an overview of three fundamental approaches for reducing the bandwidth requirement in packet video: statistical multiplexing, temporal smoothing, and multicasting. The appropriateness of each approach depends on the nature and requirements of the transported traffic. Statistical multiplexing is most suitable for real-time video in which multiple distinct streams are to be transported over the same path. This would occur, for example, in omni-directional video, where several co-located cameras generate distinct video streams, which are sent to the same destination. Temporal smoothing is the scheme of choice for archived video, whose traffic profile is known in advance. Multicasting is best suited when a single video stream is to be transmitted to multiple destinations that share some portions of the end-to-end path. Resource allocation for video traffic continues to be an exciting area of research. Some of the open issues relate to network support for interactive functions, dynamic (on-line) resource allocation for video traffic, and the provisioning of statistical end-to-end QoS for video sources.

Acknowledgments

The author wishes to express his appreciation to the guest editors and the anonymous referees for their helpful comments which considerably improved this article. This work was supported by the National Science Foundation under grant ANI-9733143.

References

[1] A. Adas. Traffic models in broadband networks. *IEEE Communications Magazine*, 35(7):82–89, July 1997.

[2] K. C. Almeroth and M. H. Ammar. The use of multicast delivery to provide a scalable and interactive video-on-demand service. *IEEE Journal on Selected Areas in Communications*, 14(6):1110–1122, Aug. 1996.

[3] W.-C. Feng and J. Rexford. A comparison of bandwidth smoothing techniques for the transmission of prerecorded compressed video. In *Proceedings of INFOCOM '97*, Apr. 1997.

[4] V. S. Frost and B. Melamed. Traffic modeling for telecommunications networks. *IEEE Communications Magazine*, 32(3):70–81, Mar. 1994.

[5] M. Grossglauser, S. Keshav, and D. Tse. RCBR: A simple and efficient service for multiple time-scale traffic. In *Proceedings of the ACM SIGCOMM '95 Conference*, pages 219–230, Aug. 1995.

[6] G. Karlsson. Asynchronous transfer of video. *IEEE Communications Magazine*, 24(8):118–126, Aug. 1996.

[7] E. W. Knightly. H-BIND: A new approach to providing statistical performance guarantees to VBR traffic. In *Proc. of the INFOCOM '96 Conference*, Mar. 1996.

[8] M. Krunz and S. K. Tripathi. Exploiting the temporal structure of MPEG video for the reduction of bandwidth requirements. In *Proceedings of the IEEE INFOCOM '97 Conference*, pages 67–74, Kobe, Japan, Apr. 1997.

[9] T. V. Lakshman, P. Mishra, and K. K. Ramakrishnan. Transporting compressed video over ATM networks with explicit rate feedback congestion. In *Proc. of the IEEE INFOCOM '97 Conference*, 1997.

[10] M. Reisslein and K. W. Ross. Join-the-Shortest-Queue prefetching for VBR video on demand. In *Proc. of International Conference on Networking Protocols*, 1997.

Biographies

Marwan M. Krunz [S'92 – M'95] received his BS in Electrical Engineering from Jordan University, Amman, Jordan, in 1990, and the MS and PhD degrees in Electrical Engineering from Michigan State University, Michigan, in 1992 and 1995, respectively. From 1995 to 1997, he was a postdoctoral research associate at the University of Maryland Institute for Advanced Computer Studies (UMIACS), College Park, Maryland. Since 1997, he has been an assistant professor in the Department of Electrical and Computer Engineering at the University of Arizona. His current research interests are in teletraffic modeling, traffic control and resource allocation in high-speed networks, video-on-demand, resource allocation in wireless networks, and QoS-based routing for ATM networks. He is a recipient of the National Science Foundation CAREER Award. He is a technical editor for the IEEE Communications Interactive Magazine, and serves on the Technical Program Committee of several conferences.