Exploiting the Temporal Structure of MPEG Video for the Reduction of Bandwidth Requirements *

Marwan Krunz and Satish Tripathi
Institute for Advanced Computer Studies
Department of Computer Science
University of Maryland
College Park, MD 20742
Email: krunz@cs.umd.edu

Abstract

We propose a new bandwidth allocation scheme for VBR video traffic in ATM networks. The scheme is tailored to MPEG-coded video sources that require stringent and deterministic quality-of-service guarantees. By exploiting the temporal structure of MPEG sources, we show that our scheme results in an *effective bandwidth* which, in most cases, is less than the source peak rate. The reduction in the bandwidth requirement is achieved without sacrificing any perceived QoS. Efficient procedures are provided for the computation of the effective bandwidth under heterogeneous MPEG sources. The effective bandwidth strongly depends on the *arrangement* of the multiplexed streams which is a measure of the degree of synchronization between the GOP patterns of different streams. Assuming that all possible arrangements are equi-probable, we derive an expression for the asymptotic tail distribution of the effective bandwidth. From the tail distribution, we compute several performance measures for the call blocking probability when the allocation is made based on the effective bandwidth. In the case of homogeneous sources, we give a closed-form expression for the 'best' arrangement that results in the 'optimal' effective bandwidth. Numerical examples based on real MPEG traces are used to demonstrate the advantages of our scheme.

Keywords: bandwidth allocation, MPEG, statistical multiplexing, CAC.

* This research was partially supported by the NSF grant # CCR 9318933.

1 Introduction

One of the major challenges in designing a BISDN/ATM network is to guarantee the quality-of-service (QoS) requirements for all transported streams without underutilizing the available bandwidth capacity. The QoS requirements can be easily satisfied by allocating bandwidth based on the peak rates of the individual sources. However, due to the burstiness of many sources (i.e., large peak rate to mean rate ratio), source-peak-rate allocation results in low utilization. To increase the utilization, statistical multiplexing can be used, which allows the available bandwidth to be shared among various streams on a need basis. By means of statistical multiplexing, it is possible to allocate an aggregate amount of bandwidth that is less than the sum of peak rates of the individual streams. This conventional use of statistical multiplexing results in possible cell queueing and buffer overflow. The amounts of cell delay and cell loss depend on the traffic model. Because of the statistical nature of commonly used traffic models, the use of statistical multiplexing is often limited to sources with *statistical* QoS requirements. Typically, a stream with deterministic QoS requirements (e.g., no cell losses) is not statistically multiplexed with other streams. Depending on its delay requirement, such a stream is either allocated its peak rate, or (if the some buffering delay can be tolerated) its peak rate over a finite interval [6].

In this paper, we investigate the bandwidth requirements of video streams that are generated by MPEG encoders. We only consider the video compression part of the MPEG standard. Although our numerical examples are based on traces produced by MPEG-I encoders, the proposed allocation scheme and the associated analysis are applicable for both MPEG-I and II. We show that, contrary to the general belief, statistical multiplexing can be used to an advantage with MPEG video traffic while providing stringent and deterministic QoS guarantees. By exploiting the deterministic and periodic manner in which frame types are generated, we show that MPEG streams can be statistically multiplexed (with an effective bandwidth per source that is less than the source peak rate) with no cell losses and very minor queueing delay. The effective bandwidth depends on the relative degree of synchronization among the multiplexed streams. We provide a simple algorithm for computing the effective bandwidth for an arbitrary synchronization structure. This algorithm can be used as part of call admission control at a switching/multiplexing network node. In situations where it is possible to have some control on the starting times of MPEG streams (e.g., in a video server), we give the form of the best synchronization structure for the multiplexed MPEG streams that has the optimal (minimum) effective bandwidth.

The rest of the paper is structured as follows. Section 2 describes the traffic model that is used to characterize an MPEG stream. Based on this model, the effective bandwidth for multiplexed MPEG streams is defined in Section 3. Efficient procedures for computing the effective bandwidth are given in Section 4. These procedures are suitable for implementation at intermediate ATM switches. In Section 5, we investigate the design of the call admission control (CAC) algorithm when resources are allocated based on the effective bandwidth. The tail distribution of the effective bandwidth for randomly 'arranged' MPEG streams is derived and used to obtain the blocking probabilities for the CAC algorithm. The 'optimal' effective bandwidth that results from the 'best' arrangement of multiplexed MPEG streams is investigated in Section 6. The derived expressions for the best synchronization arrangement and the associated optimal effective bandwidth are particularly significant at a video server where there is more flexibility to control the starting instants of video sources (compared to an intermediate node). Numerical results based on actual video streams are given in Section 7. The paper is concluded in Section 8.

2 The MPEG Source Model

A standard MPEG encoder employs several modes of compression resulting in the generation of three types of compressed frames: Intra-coded (I), Predictive (P), and Bidirectional (B) frames. In general, I frames are larger than P frames which, in turn, are larger than B frames (the frame size refers to the number of bits used to encode the frame). When compressing a video sequence, typical MPEG encoders use a pre-defined GOP pattern to determine the types of the compressed frames. Although the MPEG standards do not restrict the manner in which frame types are determined, specifying a single pattern before the start of the encoding process reduces the complexity of the encoder (in contrast to adaptively changing the GOP pattern). In addition, the use of a pre-defined GOP pattern results in more deterministic (and periodic) traffic behavior which, as we show in this paper, can be exploited to reduce the bandwidth requirements of MPEG streams. Hence, we assume throughout this paper that each MPEG stream is compressed using one GOP pattern. Different streams are allowed to have different GOP patterns. The GOP pattern defines the number and temporal order of P and B frames to be generated between two successive I frames. It is used repeatedly to compress the whole video sequence. An example of a video sequence that uses the GOP pattern *'IBBPBB'* is shown in Figure 1. The sizes of compressed frames depend on the frame types (as well as the scene dynamics). Therefore, one should expect significant impact of the periodicity of the GOP pattern on the characteristics of the traffic and, consequently, the bandwidth allocation strategies. To maintain constant-quality video, compressed frames are generated at a fixed frame rate (e.g., 30 frames/sec), resulting in VBR traffic.

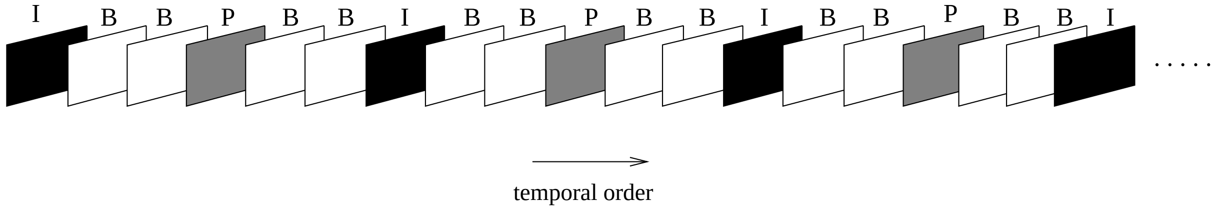

Figure 1: Example of an MPEG stream with a fixed GOP pattern.

Since B frames are coded using future I/P frames, the order in which frames are sent over the network (i.e., the transmission order) is slightly different from their encoding order. However, starting from the second I frame, the transmission and encoding orders look similar with respect to frame types. Therefore, we will ignore the first few frames in a stream, and assume, for simplicity, that frame types in an MPEG stream are represented by exact replications of the GOP pattern.

In the following development, we will be primarily concerned with a special class of GOP patterns that satisfies the following 'regularity' requirement: the number of *successive* B frames in a given MPEG stream is constant. This restriction is needed to obtain tractable analytical results, and is not required for the validity of the effective bandwidth algorithms (so long as a single GOP pattern is used repeatedly in a given stream). In fact, 'regular' GOP patterns that satisfies the above requirement are often used in practice to simplify the codec design. Regular GOP patterns can be specified by two parameters:

L : number of frames between two consecutive I frames in an MPEG stream.

Q : number of frames between an I frame and the subsequent I/P frame (whichever comes first) in an MPEG stream.

The regularity of the GOP pattern implies that L is a multiple of Q. Notice that it is possible to have $L = Q = 1$, in which case only I frames are generated (i.e., JPEG-like stream).

To provide deterministic QoS guarantees for video traffic, the analytical framework used to dimension the network resources must be based on deterministic characterization of the traffic. For our purposes, we use a deterministic traffic model which is similar, to some extent[1], to the D-BIND model that was suggested in [6]. In our model, an MPEG stream s_i is represented by a traffic envelop, $\overline{b}_i(t)$, that provides a time-varying upper bound on the actual bit rate of s_i. This envelop is completely specified by five parameters: $I^{(i)}_{max}$, $P^{(i)}_{max}$, $B^{(i)}_{max}$, $L^{(i)}$, and $Q^{(i)}$, where the first three parameters denote, respectively, the maximum sizes of I, P, and B frames in s_i. The parameters $L^{(i)}$ and $Q^{(i)}$ specify the GOP pattern of s_i. The traffic envelop is a piecewise constant function which alternates between the levels $I^{(i)}_{max}$, $P^{(i)}_{max}$, and $B^{(i)}_{max}$ according to the GOP pattern that is specified by $L^{(i)}$ and $Q^{(i)}$. Frame sizes are given in ATM cells. Cells are evenly distributed over a frame period (e.g., after pre-buffering at the source side).

3 Effective Bandwidth for Multiplexed MPEG Streams

Consider N MPEG streams, $s_1, \ldots, s_N$, with very stringent and deterministic QoS requirements (say, no losses and small queueing delays), to be transported over an ATM network. Typically, such requirements are met by allocating bandwidth based on the peak bit rate of each source.

[1]The D-BIND model provides a time-invariant bound on the cumulative arrivals. We use a more restrictive model that provides a time-varying bound on the *rate* of arrivals.

Under source-peak-rate allocation, bandwidth resources are not shared among different streams. Common belief is that statistical multiplexing should only be used for streams that content with statistical QoS guarantees. However, it is the nature of the traffic model that determines whether the offered guarantees are deterministic or statistical. Using the traffic envelop that was described in Section 2, we show that statistical multiplexing can be used advantageously with MPEG sources while supporting stringent, deterministic QoS requirements.

Let $\overline{b}_i(t)$ be the traffic envelop for s_i. For now, assume that $L^{(i)} = L$ for all i. Denote the arrival time of the first frame of s_i at the multiplexing node by t_i. We let $t_1 \triangleq 0$, so that the first stream can be used as a reference point. The lag in frame durations between a GOP period of s_1 and the following closest GOP period of s_i is given by $u_i = t_i \bmod L$. The vector $u = (u_2, u_3, \ldots, u_N)$, which we refer to as an *arrangement*, completely specifies the synchronization structure for the N streams with regard to their GOP periods (note that $u_1 \triangleq 0$). Clearly, $\overline{b}_i(t)$ is periodic in time with period L, so is the traffic envelop for the *aggregate* traffic resulting from the superposition of the N streams, $\overline{b}_{tot}(t) = \sum_i \overline{b}_i(t - u_i)$. We define the *effective bandwidth* for N streams with an arrangement u as:

$$C(u, N) \triangleq \frac{1}{N} \max_{t \geq 0} \overline{b}_{tot}(t) = \frac{1}{N} \max_{t \geq 0} \left(\sum_{i=1}^{N} \overline{b}_i(t - u_i) \right) \tag{1}$$

Because of the periodicity of $\overline{b}_{tot}(t)$, it is sufficient to take the maximum over an interval of length L. Equation (1) can also be written as:

$$C(u, N) = \frac{\sum_{j \in \Lambda_I} I_{max}^{(j)} + \sum_{j \in \Lambda_P} P_{max}^{(j)} + \sum_{j \in \Lambda_B} B_{max}^{(j)}}{N} \tag{2}$$

where $\Lambda_I, \Lambda_P, \Lambda_B$ are pairwise mutually disjoint sets with $\Lambda_I \bigcup \Lambda_P \bigcup \Lambda_B = \{s_1, \ldots, s_N\}$.

When $L^{(i)}$ varies with i, (1) and (2) are still applicable, except that $\overline{b}_{tot}(t)$ is now periodic with period $\widetilde{L}$, where

$$\widetilde{L} = \textit{least common multiple of} \{L^{(1)}, L^{(2)}, \ldots, L^{(N)}\} \tag{3}$$

and the maximization in (1) must be taken over a time interval of length $\widetilde{L}$ (also in the definition of u_i, L should be replaced by $\widetilde{L}$).

The concept of *effective bandwidth* (also known as *equivalent capacity*) was investigated in several previous studies within a stochastic framework (for example, see [4] and [1]). In this paper, the effective bandwidth is defined within a deterministic framework to guarantee zero cell loss rate and negligible queueing delays. The following simple example demonstrates the bandwidth gains that can be achieved by multiplexing MPEG streams while simultaneously supporting deterministic QoS guarantees. Let $N = 2$, $L = 6$, $Q = 3$, and $u_2 = 1$. Assume that both streams are characterized by the same traffic envelop, $\bar{b}(t)$, with $I_{max} > P_{max} > B_{max}$ (see Figure 2). Then,

$$C(u,2) \triangleq \frac{1}{N} \max_{t \geq 0} \bar{b}_{tot}(t) = \frac{1}{N} \max_{t \geq 0} \left(\bar{b}(t) + \bar{b}(t-1) \right) = \frac{I_{max} + B_{max}}{2} < I_{max} \tag{4}$$

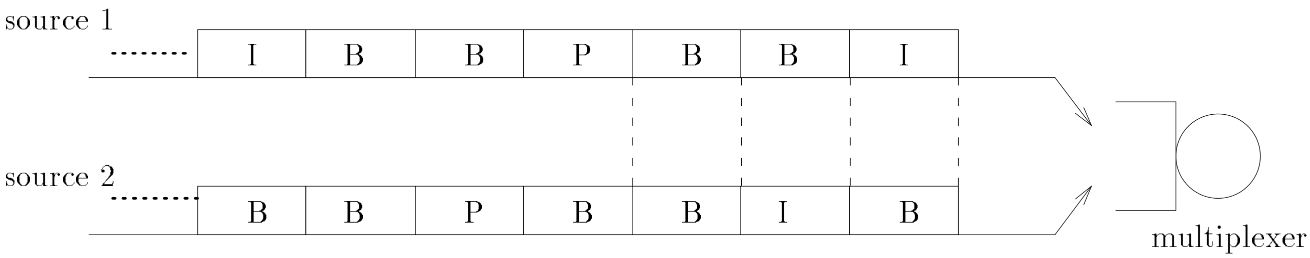

Figure 2: Example of the effective bandwidth for two multiplexed streams.

By superposing the two streams and allocating bandwidth for the aggregate traffic, the required amount of bandwidth per source decreased from I_{max} (source-peak-rate allocation) to $(I_{max} + B_{max})/2$. A very small buffer of N cells is needed at the input to the multiplexer in case cells from several sources arrive simultaneously. Notice that bandwidth gains from statistical multiplexing are obtained via spatial averaging, and *not* temporal averaging.

It is clear that the effective bandwidth depends on the degree of synchronization among the multiplexed MPEG streams. If the two streams in the above example send I frames simultaneously (i.e., $u_2 = 0$), then $C(u,2) = I_{max}$, and statistical multiplexing introduces no advantages over source-peak-rate allocation. Fortunately, the chance that both streams are in the same phase (i.e., sending I frames simultaneously) is small.

4 Computing the Effective Bandwidth

In this section, we give efficient procedures for the computation of $C(u,N)$ or, in some cases, an upper bound on it. In practice, such computation must be done on-line so that the allocated bandwidth for a group of ongoing video connections at a node can be updated dynamically upon the admittance of a new connection or the termination of an ongoing one. The procedures to be described are valid for heterogeneous MPEG sources (i.e., sources with different traffic envelops). Two separate cases are considered in the computation of $C(u,N)$ (or its upper bound), depending on the range of values that can be assumed by the elements of u.

4.1 Aligned Boundaries Case

Suppose that the elements of u can take only integer values in $\{0, 1, \ldots, \widetilde{L}-1\}$. Consequently, frame boundaries of various sources are exactly aligned in time, and $\overline{b}_{tot}(t)$ is a fixed-interval piecewise-constant periodic function with period $\widetilde{L}$. For any integer k, $\overline{b}_{tot}(t)$ is constant for all $t \in (k, k+1)$. Figure 2 depicts an example of an aligned boundaries case. In practice, frame boundaries are not necessarily aligned, but can be made so: if multiplexing takes place at a video server, alignment of frame boundaries can be imposed by delaying the starting time of a stream by no more than a frame period (which will not be noticed by video clients). If multiplexing takes place at the output port of a switch, then the alignment is done by introducing some delay in the path of each stream before entering the multiplexer.

The advantage of having frame boundaries aligned is that $\overline{b}_{tot}(t)$ in this case is equivalent to a discrete-time function for which $C(u, N)$ can be exactly computed on-line. The time axis is slotted where a slot is one frame period. Because of the periodicity of $\overline{b}_{tot}(t)$, computation of $C(u, N)$ requires only maintaining the values of the traffic envelops for the first $\widetilde{L}$ slots (from 0 to $\widetilde{L} - 1$). Such slots are referred to as *phases*. We denote the value of $\overline{b}_i(t - u_i)$ during phase j by $\overline{b}_{i,j}$. Thus,

$$\overline{b}_{i,j} \triangleq \overline{b}_i(\tau - u_i) \;\text{ for any } \tau \in (j, j+1) \tag{5}$$

(to accommodate negative t, $\overline{b}_i(t)$ is extended in the negative time axis). To compute $C(u, N)$, the multiplexing node maintains a matrix $M = [m_{ij}]$ of size $N \times \widetilde{L}$. Each multiplexed MPEG stream is associated with one row in the table. For $i = 1, \ldots, N$, and $j = 1, \ldots, \widetilde{L}$, $m_{ij} = \overline{b}_{i,j-1}$. In addition, the node maintains a row vector $V = \left[v_1, \ldots, v_{\widetilde{L}}\right]$, where

$$v_j = \sum_{i=1}^{N} m_{ij} \;\; \forall \;\; j \tag{6}$$

which gives the peak bit rate for the aggregate traffic during phase $j - 1$. Now, $C(u, N)$ is simply given by:

$$C(u, N) = \frac{1}{N} \max_{0 \le j \le \widetilde{L}-1} v_j \tag{7}$$

When the $(N+1)$th stream arrives at the node, a row is added to M based on $\overline{b}_{N+1}(t)$ and u_{N+1}. To prevent updating the number of columns in M, it is best to choose $\widetilde{L}$ in advance to accommodate anticipated values of $L^{(i)}$ (which are quite few in practice). The effective bandwidth is re-computed by updating V (using $v_j := v_j + m_{N+1,j}$), and then applying (7) with $N+1$ replacing N. Similar procedure is required when an ongoing connection is dropped. Clearly, very few steps are needed to re-compute the effective bandwidth upon the adding/dropping of a stream.

4.2 Non-Aligned Boundaries Case

The elements of u in the non-aligned boundaries case can take any real value in $[0, \widetilde{L})$. Unless some extra work is done to enforce their alignment, frame boundaries in real systems are generally non-aligned. Exact computation of $C(u, N)$ for the non-aligned boundaries case is computationally expensive. Instead, we provide an upper bound on $C(u, N)$ which can be computed on-line.

Consider N heterogeneous streams with $u_i \in [0, \widetilde{L})$ for all i. Since u_i does not necessarily take integer values, using a table of size $N \times \widetilde{L}$, as in the previous case, is not sufficient for computing $C(u, N)$. The reason is that the constant-valued segments of $\overline{b}_{tot}(t)$ can be of variable lengths. Thus, $\overline{b}_{tot}(t)$ could vary at most $N\widetilde{L}$ times within a period of $\widetilde{L}$ (compared to $\widetilde{L}$ times in the aligned boundaries case). Since N changes dynamically, the size of the table and the cost of updating it can be computationally prohibitive, if $C(u, N)$ is to be obtained on-line. Our solution is to provide an upper bound on $C(u, N)$. As before, we use a slotted time system where each slot is a frame period. Slots are synchronized locally at the node using a counter (from 0 to $\widetilde{L} - 1$) and a timer. Since $\overline{b}_{tot}(t)$ is periodic in $\widetilde{L}$, we only record the peak bit rates in the first $\widetilde{L}$ slots (i.e., phases 0 to $\widetilde{L} - 1$). A matrix $\widehat{M} = [\widehat{m}_{ij}]$ of dimensions $2N \times \widetilde{L}$ is maintained. Each ongoing stream, s_i, is associated with two rows of $\widehat{M}$, where the $\widetilde{L}$ sampled values of of $\overline{b}_i(t)$ are recorded in one row assuming that s_i is exactly aligned with phase $\lfloor u_i \rfloor$, and in the second row assuming that s_i is exactly aligned with phase $\lceil u_i \rceil \bmod \widetilde{L}$. By definition, the two rows representing s_i are adjacent, so that the ith stream is associated with the $(2i-1)$th and the $(2i)$th rows of $\widehat{M}$. Hence,

$$\widehat{m}_{ij} = \begin{cases} \overline{b}_{(i+1)/2, j-1} & \text{if } i \text{ is odd} \\ \overline{b}_{i/2, j-2} & \text{if } i \text{ is even} \end{cases} \tag{8}$$

where $\overline{b}_{i,j}$ is now defined as:

$$\overline{b}_{i,j} \triangleq \overline{b}_i\left(\tau - \lfloor u_i \rfloor\right) \quad \text{for any } \tau \in (j, j+1) \tag{9}$$

In addition to $\widehat{M}$, the node maintains a row vector $\widetilde{V} = \left[\widetilde{v}_1, \ldots, \widetilde{v}_{\widetilde{L}}\right]$, where

$$\widetilde{v}_j = \sum_{i=1}^{N} \max\left\{m_{2i-1,j},\ m_{2i,j}\right\} \quad \forall\ j \tag{10}$$

which gives the peak bit rate for the aggregate traffic during phase $j - 1$. An upper bound on $C(u, N)$ is given by:

$$\overline{C}(u, N) = \frac{1}{N} \max_{1 \le j \le \widetilde{L}} \widetilde{v}_j \tag{11}$$

To see why (11) is an upper bound, substitute (8) in (10) to obtain

$$\begin{aligned}
\widetilde{v}_j &= \sum_{i=1}^{N} \max\left\{\overline{b}_{i,j-1},\ \overline{b}_{i,j-2}\right\} &(12)\\
&= \sum_{i=1}^{N} \max_{j-2\le\tau\le j}\left\{\overline{b}_i\left(\tau-\lfloor u_i\rfloor\right)\right\} = \sum_{i=1}^{N} \max_{j-2-\lfloor u_i\rfloor\le\tau\le j-\lfloor u_i\rfloor}\left\{\overline{b}_i\left(\tau\right)\right\} &(13)
\end{aligned}$$

But $j-2-\lfloor u_i\rfloor \le j-1-u_i < j-u_i \le j-2-\lfloor u_i\rfloor$. Thus,

$$\widetilde{v}_j \ge \sum_{i=1}^{N} \max_{j-1-u_i\le\tau\le j-u_i}\left\{\overline{b}_i\left(\tau\right)\right\} = \sum_{i=1}^{N} \max_{j-1\le\tau\le j}\left\{\overline{b}_i\left(\tau-u_i\right)\right\} \tag{14}$$

From (11) and (14), we have

$$\begin{aligned}
\overline{C}(u,N) &\ge \frac{1}{N}\max_{1\le j\le\widetilde{L}}\left\{\sum_{i=1}^{N}\max_{j-1\le\tau\le j}\left\{\overline{b}_i\left(\tau-u_i\right)\right\}\right\}\\
&\ge \frac{1}{N}\max_{1\le j\le\widetilde{L}}\left\{\max_{j-1\le\tau\le j}\left\{\sum_{i=1}^{N}\overline{b}_i\left(\tau-u_i\right)\right\}\right\}\\
&= \frac{1}{N}\max_{0\le\tau\le\widetilde{L}}\left\{\sum_{i=1}^{N}\overline{b}_i\left(\tau-u_i\right)\right\} = C(u,N) &(15)
\end{aligned}$$

(in the above equations, the maximization over j is taken on integer values while the maximization over τ is taken on real values).

Upon the arrival of the $(N+1)$th connection to a node with N ongoing connections, two rows are added to $\widehat{M}$ based on $\overline{b}_{N+1}(t)$ and u_{N+1}, and $\widetilde{v}_j$ is updated using:

$$\widetilde{v}_j := \widetilde{v}_j + \max\left\{m_{2(N+1)-1,j},\ m_{2(N+1),j}\right\}\quad \forall\, j \tag{16}$$

Then, the new bound on the effective bandwidth is obtained using (11) (with $N+1$ replacing N). When the ith ongoing connection terminates, $\widetilde{v}_j$ is updated using:

$$\widetilde{v}_j := \widetilde{v}_j - \min\left\{m_{2i-1,j},\ m_{2i,j}\right\}\quad \forall\, j \tag{17}$$

It is obvious that $C(u,N)$-based reservation has a better chance to achieve bandwidth gains than $\overline{C}(u,N)$. For example, in the homogeneous case, if u has a multivariate uniform distribution with state space $\{0,1,\ldots,L-1\}$ in the aligned boundaries case, and $[0,L)$ in the non-aligned boundaries case, then $\Pr\{\overline{C}(u,N)=I_{max}\} = 1/L^{N-1}$ and $2/L^{N-1}$ for the aligned and non-aligned cases, respectively.

5 CAC and Blocking Probability under $C(u, N)$ Reservation

From the previous section, it is clear that there is a small non-zero probability that $C(u,N)$ is equal or very close to the source peak rate. In this section, we investigate the impact of $C(u,N)$ reservation on CAC and the call blocking probability. To obtain tractable results, our treatment is limited to the homogeneous traffic case, in which all MPEG sources have a common traffic envelop, $\overline{b}(t)$, that is characterized by the 5-tuple (I_{max}, P_{max}, B_{max}, L, Q) with $I_{max} > P_{max} > B_{max}$. For heterogeneous streams with relatively close, but different, maximum frame sizes, and similar L and Q, a common traffic envelop can be obtained by taking I_{max} as the largest I frame in all the streams (similarly, for P_{max} and B_{max}). We only consider the case of aligned boundaries, although extension to the case of non-aligned boundaries (using $\overline{C}(u,N)$ reservation) is straightforward.

Let W be the total bandwidth capacity (in cells/frame period) of the multiplexing node. Suppose that N streams are already admitted and are being allocated their effective bandwidth. Hence, the available free capacity is $W - NC(u,N)$, where u is the arrangement of the N streams. Our goal is to compute the blocking probability for a a new connection request that arrives at the node. Clearly, the the blocking probability depends on W, N, $\overline{b}(t)$, and $C(u,N)$. We fix the first three factors, and assume a multivariate discrete uniform distribution for u. Since the node cannot anticipate in advance the arrival instant of the first frame of the $(N+1)$th stream, the CAC algorithm must be designed assuming a worst-case scenario. Hence, the blocking probability can be generally defined as:

$$\text{Blocking Probability} \stackrel{\triangle}{=} \Pr\{I_{max}^{(N+1)} > W - NC(u,N)\} \tag{18}$$

where $C(u,N)$ is the only random quantity. For homogeneous traffic, (18) can be written as:

$$\text{Blocking Probability} = \Pr\left\{C(u,N) > \frac{W - I_{max}}{N}\right\} \stackrel{\triangle}{=} G_N\left(\frac{W - I_{max}}{N}\right) \tag{19}$$

where $G_N(x) \stackrel{\triangle}{=}$ the complementary distribution function for $C(u,N)$. Thus, to obtain the blocking probability, we need to compute $G_N(x)$.

5.1 Asymptotic Tail Distribution of $C(u, N)$

For the present case of aligned boundaries and homogeneous streams, $C(u,N)$ can be written as:

$$C(u,N) = \frac{1}{N} \max_{0 \le j \le L-1} \left(\sum_{i=1}^{N} \overline{b}_{i,j} \right) \tag{20}$$

where $\overline{b}_{i,j}$ was defined in (5). It is also possible to write $C(u,N)$ as:

$$C(u,N) = \frac{n_I I_{max} + n_P P_{max} + (N - n_I - n_P) B_{max}}{N} \tag{21}$$

where n_I and n_P are random variables with probability space $\{0, 1, \ldots, N\}$. Therefore, $G_N(x)$ can be obtained from the joint distribution of (n_I, n_P). Unfortunately, complete specification of this distribution depends on the relative values of I_{max}, P_{max}, and B_{max}. It is, however, possible to obtain the asymptotics for the tail of this distribution under the assumption that $I_{max} > P_{max} > B_{max}$. First, we need to introduce some elementary results. Notice that obtaining $C(u, N)$ in the aligned boundaries case requires only computing $\overline{b}_{tot}(t)$ in the first L slots (i.e., phases 0 to $L-1$). To simplify the notation, we use $\overline{b}_{tot}(i)$ to mean $\overline{b}_{tot}(\tau)$ for any $\tau \in (i, i+1)$. A stream s_i is said to be *in* phase k if $u_i = k$, i.e., s_i sends an I frame during phase k. For all $i \in \{0, \ldots, L-1\}$, let

$$r_i \triangleq \text{number of streams in phase } i$$
$$z_i \triangleq \text{number of streams in phases that are multiples of } Q \text{ from phase } i$$

Thus, r_i and z_i give the numbers of streams sending I and P frames, respectively, during phase i. The following proposition follows directly from the periodicity of the GOP patterns.

Proposition 1 *Consider any two streams i and j with $u_i = k_1$ and $u_j = k_2$, $k_1 \neq k_2$. If during phase k_1 stream j sends a B frame, then during phase k_2 stream i sends a B frame. Similarly, if during phase k_1 stream j sends a P frame, then during phase k_2 stream i sends a P frame.* □

From Proposition 1, it is easy to see that for any two phases, i and j, with $| i - j |=$ a multiple of Q, we have $r_i + z_i = r_j + z_j$. Based on this result, we introduce the following proposition.

Proposition 2 *Let phase k be such that $r_k = \max_i r_i$. If $r_k > N/2$, then $n_I = r_k$. Moreover, phase k is the only phase for which $C(u, N) = \overline{b}_{tot}(k)/N$.*

Proof: First, suppose that j is a phase such that $|k - j| =$ a multiple of Q, then $r_k + z_k = r_j + z_j$. In addition, both phases will have the same number of sources that send B frames. Since $r_k > r_j$ (strictly since only one phase can exist with $r_k > N/2$), $\overline{b}_{tot}(k) > \overline{b}_{tot}(j)$, and the assertion is true.

Next, suppose that $|k - j| \neq$ a multiple of Q, then all r_k streams that are in phase k will send B frames during phase j (by Proposition 1). Thus, $N - r_j - z_j \geq r_k$, which leads to $r_j + z_j \leq N - r_k < N/2 < r_k$. Consequently, $\overline{b}_{tot}(j) < \overline{b}_{tot}(k)$ and the assertion is true. □

The implication of Proposition 2 is that when $r_k > N/2$, the peak bit rate for the superposed streams during phase k "majorizes" the peak bit rate during all other phases, regardless of the exact values of I_{max}, P_{max}, and B_{max}. Therefore, we can compute the joint probability for (n_I, n_P)

by simply computing the joint probability for (r_k, z_k), where $r_k = \max_i r_i$. Based on the above, for $i > N/2$ we have:

$$p_{ij} \stackrel{\triangle}{=} \Pr\{n_I = i, n_P = j\} = \sum_{l=0}^{L-1} \Pr\{r_l = i, z_l = j\}, \quad \text{for any } j \in \{0, \ldots, N-i\} \tag{22}$$

Note that when $i > N/2$, the events $[r_l = i]$ and $[r_m = i]$, $l \neq m$, are mutually exclusive. Since we defined $u_1 = 0$, the first term in the above sum is given by:

$$\Pr\{r_0 = i, z_0 = j\} = \binom{N-1}{i-1} \left(\frac{1}{L}\right)^{i-1} \binom{N-i}{j} \left(\frac{L/Q-1}{L}\right)^j \left(\frac{L-L/Q}{L}\right)^{N-i-j} \tag{23}$$

Observe that there are $\binom{N-1}{i-1}$ possibilities for the $N-1$ streams (excluding the first stream) to send $i-1$ I frames, each possibility with probability $(1/L)^{i-1}$. Among the remaining $N-i$ streams, there are $\binom{N-i}{j}$ possibilities to send j P frames, each possibility with probability $((L/Q-1)/L)^j$ (since the number of P frames in a GOP period is $L/Q-1$). Similar argument justifies the last term in (23), which is related to the probability of sending B frames.

In a similar manner, it is easy to show that for $l \in \{Q, 2Q, 3Q, \ldots, (L/Q-1)Q\}$, we have:

$$\Pr\{r_l = i, z_l = j\} = \binom{N-1}{i} \left(\frac{1}{L}\right)^{i} \binom{N-1-i}{j-1} \left(\frac{L/Q-1}{L}\right)^{j-1} \left(\frac{L-L/Q}{L}\right)^{N-i-j} \tag{24}$$

Finally, for $l \in \{0, 1, 2, \ldots, Q-1, Q+1, Q+2, \ldots, 2Q-1, 2Q+1, \ldots, L-1\}$, we have:

$$\Pr\{r_l = i, z_l = j\} = \binom{N-1}{i} \left(\frac{1}{L}\right)^{i} \binom{N-1-i}{j} \left(\frac{L/Q-1}{L}\right)^{j} \left(\frac{L-L/Q}{L}\right)^{N-1-i-j} \tag{25}$$

From (23), (24), and (25), and after some manipulations, (22) can be written as:

$$p_{ij} = \frac{\binom{N}{i}\binom{N-i}{j}(L/Q-1)^j (L-L/Q)^{N-i-j}}{L^{N-1}} \tag{26}$$

Since (26) is valid only for $i > N/2$, we must choose x sufficiently large such that the event $[C(u,N) > x]$ necessarily implies $[n_I > N/2]$. Let $x^* \stackrel{\triangle}{=} \inf\{x \;:\; [C(u,N) > x] \Rightarrow [n_I > N/2]\}$.

Then, for $x > x^*$

$$G_N(x) = \sum_{\substack{i,j \text{ such that} \\ f_{ij} > x}} p_{ij} \tag{27}$$

where

$$f_{ij} \triangleq \frac{iI_{max} + jP_{max} + (N - i - j)B_{max}}{N} \tag{28}$$

It is easy to see that

$$x^* = \frac{(N/2)I_{max} + (N/2)P_{max}}{N} = \frac{I_{max} + P_{max}}{2} \tag{29}$$

since any value of $C(u, N)$ that is greater than the RHS of (29) implies necessarily that $n_I > N/2$. For $x > x^*$, $G_N(x)$ is obtained from (26), (27), and (28).

5.2 Blocking Probability Performance Measures

In this section, we compute several performance measures for the blocking probability of MPEG connection requests. Our results are related to the homogeneous aligned-boundaries case. Throughout this section, N denotes the number of ongoing connections. To provide a comparison between source-peak-rate allocation and effective bandwidth allocation, we let the total node capacity, $W = (N - K)I_{max}$, where K is a fixed known integer.

5.2.1 Zero-Order Blocking Probability

The first case of interest is when $K = 0$, i.e., W is equal to the sum of the ongoing sources' peak rates. Under source-peak-rate allocation, no more connections can be admitted. If allocation is made based on $C(u, N)$, and j new connection requests arrive simultaneously at the node, then the probability that these requests are rejected is given by:

$$\begin{aligned} P_0^{(N,j)} &\triangleq \Pr\{j \text{ new requests are rejected when } W = NI_{max}\} \\ &= \Pr\{jI_{max} > W - NC(u, N)\} \\ &= \Pr\{C(u, N) > \frac{N - j}{N} I_{max}\} = G_N\left(\frac{N - j}{N} I_{max}\right) \end{aligned} \tag{30}$$

Since $G_N(x)$ is known only for $x > x^*$, we must choose N sufficiently large so that the above results are valid. Let $P_{max} = \alpha I_{max}$ for some $0 < \alpha < 1$. Let $N^* \triangleq \min\{n : (n - j)I_{max}/n > x^*\}$. It is easy to show that $N^* = \left\lceil \frac{2j}{1-\alpha} \right\rceil$. Thus, for $N \geq N^*$ and a fixed j, the zero-order blocking probability is given by (30). Of particular interest is the case when $j = 1$ since connection requests seldom arrive in batches.

5.2.2 Higher-Order Blocking Probability

Suppose that $K > 0$. This means that $C(u, N)$-based allocation has already succeeded in admitting a number of connections that could not be admitted according to source-peak-rate allocation. To compute the blocking probability for additional j connection requests, we define the K-order blocking probability (where $K \geq 1$) as:

$$\begin{aligned} P_K^{(N,j)} &\triangleq \Pr\{j \text{ new requests are rejected when } W = (N-K)I_{max}\} \\ &= \Pr\{jI_{max} > W - NC(u,N) \,/\, W \geq NC(u,N)\} \\ &= \frac{\Pr\left\{C(u,N) > \frac{N-K-j}{N}I_{max},\; C(u,N) \leq \frac{N-K}{N}I_{max}\right\}}{\Pr\left\{C(u,N) \leq \frac{N-K}{N}I_{max}\right\}} \\ &= \frac{G_N\left(\frac{N-K-j}{N}I_{max}\right) - G_N\left(\frac{N-K}{N}I_{max}\right)}{1 - G_N\left(\frac{N-K}{N}I_{max}\right)} \end{aligned} \tag{31}$$

which can be obtained for $N \geq N^* = \left\lceil \frac{2(K+j)}{1-\alpha} \right\rceil$.

5.2.3 End-to-End Blocking Probability

The blocking probabilities given in the previous two sections are related to a single node. Network architects are often interested in the end-to-end blocking probability for a connection. Using the blocking probabilities at individual nodes, it is easy to derive the end-to-end blocking probability for MPEG connections that are allocated their effective bandwidth. Denote the K-order blocking probability for j simultaneous requests at the rth node by $P_K^{(N_r,j)}(r)$, where N_r is the number of ongoing connections at the rth node. We assume that the blocking probabilities at different nodes are independent. Let $P_{block}(n)$ be end-to-end blocking probability for a connection that traverses n nodes. Then,

$$P_{block}(n) = 1 - \prod_{i=1}^{n}\left(1 - P_K^{(N_i,j)}(i)\right) \tag{32}$$

If $\max_i\{P_K^{(N_i,j)}(i)\} \ll 1/n$, then $P_{block}(n)$ can be approximated by:

$$P_{block}(n) \approx \sum_{i=1}^{n} P_K^{(N_i,j)}(i) \tag{33}$$

6 "Optimal" Effective Bandwidth

Since $C(u, N)$ varies with u, it is natural to seek the 'best' arrangement that produces the "optimal" effective bandwidth, $C_{opt}(N)$, for N multiplexed MPEG streams. In this section, we obtain expressions for a 'best' arrangement and its associated $C_{opt}(N)$. Our treatment is limited to the homogeneous case (extension to the heterogeneous case is possible, but will be deferred to a future paper). Without loss of generality, we assume that frame boundaries are aligned. If frame boundaries are generally non-aligned, $C_{opt}(N)$ is the same as in the aligned boundaries case. The reason is that when frames boundaries are not aligned, the effective bandwidth is greater than or equal the effective bandwidth of some arrangement with aligned boundaries. Since aligned-boundaries is a special case of the non-aligned boundaries, a 'best' arrangement in the special case is also a 'best' arrangement in the general case.

The optimal effective bandwidth is defined as:

$$C_{opt}(N) \triangleq \min_{u \in \mathcal{U}} C(u, N) \tag{34}$$

where $\mathcal{U}$ is the set of all possible *distinct* arrangements of N streams. The size of $\mathcal{U}$ is given by:

$$|\mathcal{U}| = \sum_{i=1}^{m} \binom{L}{i} \binom{N-2}{i-1} \quad \text{where } m = \min\{N-1, L\} \tag{35}$$

The size of $\mathcal{U}$ increases rapidly with N. Therefore, obtaining $C_{opt}(N)$ on-line using (34) is computationally prohibitive for moderate and large N. Instead, we give the form of a 'best' arrangement, u^*, and the expression for the associated $C_{opt}(N)$. It turned out that the form of u^* is quite intuitive, although proving its optimality is not trivial. Note that it is possible to have several distinct 'best' arrangements that results in $C_{opt}(N)$. Table 1 depicts the form of u^* and the expression for $C_{opt}(N)$. The mathematical proof for these results is outlined in the appendix.

A best arrangement of N streams for $N = 1, 2, \ldots$, is given by:

$$u^* = (\underbrace{0, 1, 2, \ldots, L-1, 0, 1, 2, \ldots, L-1, \ldots,}_{w \text{ times}} 0, 1, 2, \ldots, N - wL - 1) \tag{36}$$

Optimal effective bandwidth is:

$$C_{opt}(N) = \frac{(w+1)I_{max} + (m-w)P_{max} + (N-1-m)B_{max}}{N} \tag{37}$$

where

$$\begin{aligned} w &\triangleq \text{largest nonnegative integer } k \text{ that satisfies } N > kL \\ m &\triangleq \text{largest nonnegative integer } k \text{ that satisfies } N > kQ \end{aligned}$$

Table 1: A best arrangement of N streams and the associated optimal effective bandwidth.

Suppose that N ongoing streams are arranged as in (36). A newly admitted stream can be added to the existing ones, resulting in a best arrangement of $(N+1)$ streams without disrupting the original structure of the N streams. In other words, u^* of $(N+1)$ streams can be obtained by simply concatenating a single number to u^* of N streams. When N streams are arranged according to u^* and $N \leq L$, the removal of *any* stream will still result in a best arrangement. When $N > L$, only the removal of certain streams preserves the optimality of the arrangement.

7 Numerical Results

In this section, we use real MPEG traces to provide numerical examples of the analytical results presented in previous sections. The traces were captured by several research groups [3, 6, 7, 10] for various types of video (action movies, advertisements, and a lecture). All the traces were generated using MPEG-I encoders (see the references for details on how these traces were obtained). The traces are listed in Table 2 along with the parameters of their traffic envelops. Frame sizes are converted to ATM cells.

Figure 3 gives the percentage of $C(u,N)/I_{max}$ as a function of u in the aligned boundaries case. Here, we use the *Wizard of Oz* parameters with $N = 3$. For simplicity, $u = (u_1, u_2, u_3)$ is varied by varying u_3 in $\{0, \ldots, L-1\}$ with $u_2 = 0$, 1, and 2 ($u_1 \stackrel{\triangle}{=} 0$). Using the same parameters, the percentage of $\overline{C}(u,N)/I_{max}$ in the non-aligned boundaries case is shown in Figure 4 as a function of u_3 (which is varied continuously in $[0, L)$). It is clear that except for one possible arrangement, $u = (0,0,0)$, statistical multiplexing can reduce the bandwidth requirements without sacrificing any performance guarantees. In fact, even when the number of sources is as small as 3, the bandwidth requirement for a stream can be reduced in some cases to less than 50% of the source peak rate.

Using the *Wizard of Oz* parameters, $G_N(x)$ (the tail distribution of $C(u,N)$) is plotted in Figure 5 with $N = 15$. In this case, the critical value of x above which $G_N(x)$ is defined is given by $x^* = 818$ cells. For $x = (N-1)I_{max}/N$, the zero-order blocking probability for a new stream is given by $P_0^{(N,1)} = G_N((N-1)I_{max}/N) \approx 1.866 \times 10^{-10}$. Zero-order blocking probabilities of one new request, $P_0^{(N,1)}$, are plotted in Figure 6 as a function of N, for three traces. Each plot in the figure is given for $N \geq N^*$. The zero, first, and second-order blocking probabilities of one new request are plotted against N, based on *Lecture* parameters. In Figure 8, the variation of $C_{opt}(N)$ (given as a percentage of the source peak rate, I_{max}) is shown as a function of N, using different L and Q values. Maximum frame sizes (I_{max}, P_{max}, and B_{max}) are taken from the *Wizard of Oz* trace which was compressed using $L = 15$ and $Q = 3$. For simplicity, the same maximum sizes are used in to obtain $C_{opt}(N)$ under other L and Q values. Although one might expect that for a given movie, the maximum sizes of compressed frames vary with L and Q, our experiments (discussed below) suggest that compressing a video segment using different (L,Q) pairs has little impact on I_{max}, P_{max}, and B_{max}.

Several noteworthy observations can be inferred from Figure 8. First, as N increases, $C_{opt}(N)$ decreases, but not monotonically, and converges slowly to some positive value. The limiting value

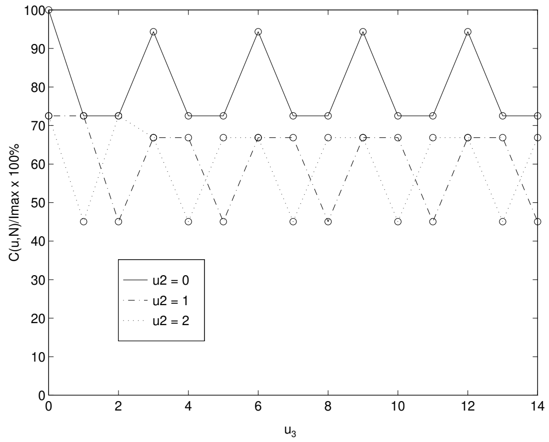

Figure 3: $C(u,N)$ – aligned boundaries.

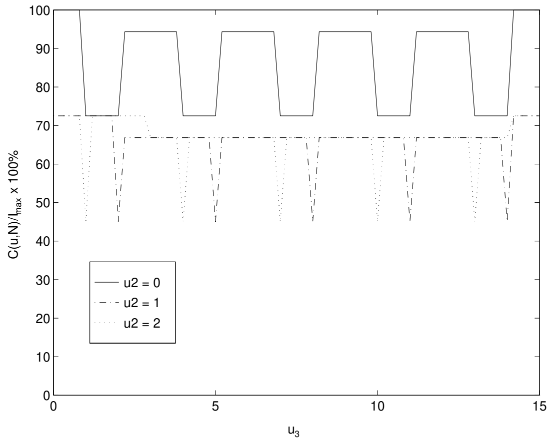

Figure 4: $\overline{C}(u,N)$ – non-aligned boundaries.

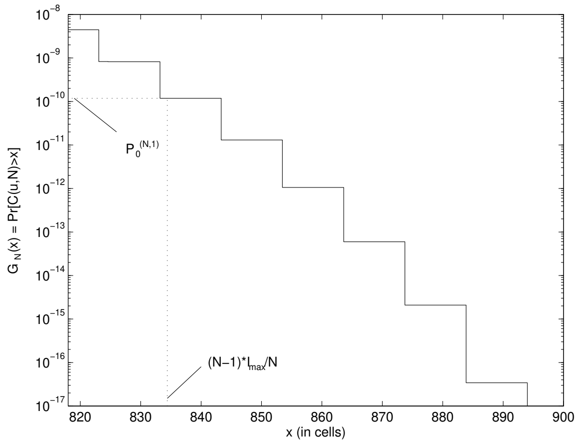

Figure 5: $G_N(x)$.

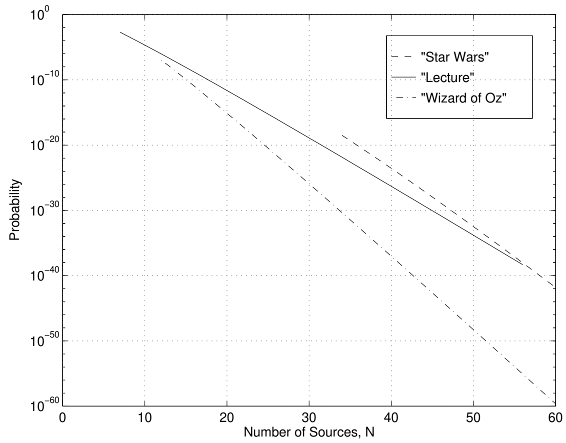

Figure 6: Zero-order blocking probability.

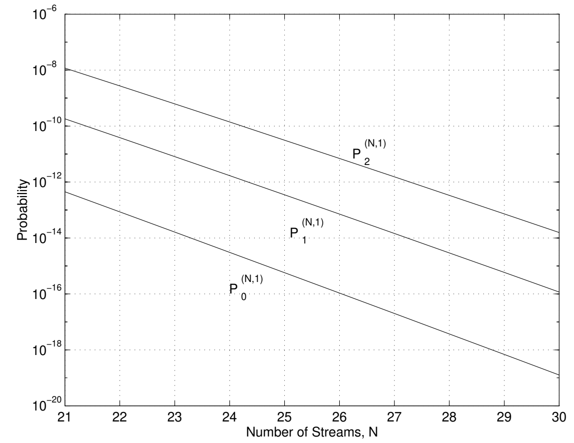

Figure 7: Higher-Order blocking probabilities.

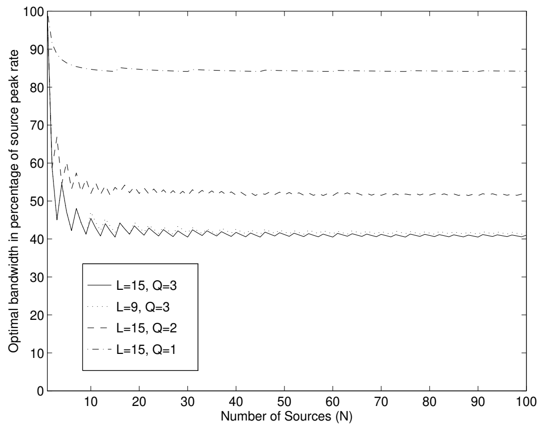

Figure 8: $C_{opt}(N)/I_{max} \times 100\%$ versus N.

of $C_{opt}(N)$ can be determined from (37). For large N, $w \rightarrow N/L$ and $m \rightarrow N/Q$. Thus,

$$C^*_{opt} \stackrel{\triangle}{=} \lim_{N\rightarrow\infty} C_{opt}(N) = (1/L)I_{max} + (1/Q - 1/L)P_{max} + (1 - 1/Q)B_{max} \quad (38)$$

The limiting value of $C_{opt}(N)$ is, in fact, achievable when $N = kL$ for $k = 1, 2, 3, \ldots$, implying that the highest possible multiplexing gains are obtained whenever the number of multiplexed streams is a multiple of L. For moderate and large N, $C_{opt}(N)$ is almost insensitive to N (compare the plots for $(L, Q) = (15, 3)$ and $(L, Q) = (9, 3)$. This is expected since P_{max} is close to (but smaller than) I_{max}. When $P_{max} \approx I_{max}$, C^*_{opt} in (38) reduces to $(1/Q)P_{max} + (1 - 1/Q)B_{max}$ which does not depend on L. On the other hand, the optimal effective bandwidth seems to depend heavily on Q. In the above example, when $L = 15$ and Q is varied from $Q = 1$ (only I and P frames) to $Q = 3$, C^*_{opt} decreased from $C^*_{opt} = 84\%\ I_{max}$ to $C^*_{opt} = 40.5\%\ I_{max}$. Clearly, the relative impact of L and Q depends on the relative values of I_{max}, P_{max}, and B_{max}. In most cases, P_{max} is closer to I_{max} than to B_{max}. We verified that by examining the traces of several MPEG-compressed movies. The traces are listed in Table 2. The last column in the table gives the limit on the statistical multiplexing gain (given as a percentage of the source peak rate).

Trace	Length (in frames)	I_{max}	P_{max}	B_{max}	L	Q	$(C^*_{opt}/I_{max}) \times 100\%$
Star Wars [3]	174136	483	454	169	12	3	55%
Wizard of Oz [7]	41760	894	742	157	15	3	41%
Advertisements [6]	16316	215	214	162	6	3	84%
Lecture [6]	16316	131	92	32	6	3	45%
Silence of the Lambs [10]	40000	350	231	144	12	3	53%

Table 2: Empirical MPEG traces for different video movies with various GOP patterns (frame sizes in cells). The last column shows C^*_{opt} as a percentage of source peak rate.

To study the impact of L and Q on the maximum sizes of I, P, and B frames, we chose a segment from *Wizard of Oz* movie, and compressed it several times using different L and Q values. The segment corresponds to 12600 frames (from frame No. 29191 to frame No. 41790 in the movie). Table 3 depicts the GOP patterns that were used and the measured I_{max}, P_{max}, and B_{max}. In addition, the table gives the limiting value for $C_{opt}(N)$, which is computed from (38). It is clear that the GOP pattern has a very insignificant impact on the maximum frames sizes (note, however, that the overall average of frames sizes can considerably vary from one GOP pattern to another). This can be intuitively justified by the fact that a movie consists of several 'scenes'. A scene can be loosely defined as a segment of the movie with relatively consistent level of activity. Sizes of I frames (similarly, P and B frames) within a scene are close is value. Since on the average a scene lasts for several seconds [2], changing the compression pattern (whose time scale is smaller than one second) will have little effect on the maximum sizes of I, P, and B frames within a scene.

From the last column of Table 3, it is obvious that L has a very negligible effect on C^*_{opt}, whereas

Compression Pattern	L	Q	I_{max}	P_{max}	B_{max}	$(C^*_{opt}/I_{max}) \times 100\%$
I	1	1	908	—	—	100%
IP	2	1	898	756	—	92.1%
IPP	3	1	898	756	—	89.5%
IPPP	4	1	896	756	—	88.3%
IPPPP	5	1	896	740	—	86.1%
IBPB	4	2	896	733	161	54.4%
IBPBPB	6	2	898	742	161	53.2%
IBPBPBPB	8	2	889	742	161	52.9%
IBPBPBPBPB	10	2	894	742	161	52.2%
IBBPBB	6	3	898	719	157	41.7%
IBBPBBPBB	9	3	896	742	157	41.2%
IBBPBBPBBPBB	12	3	896	742	157	40.7%
IBBPBBPBBPBBPBB	15	3	893	742	157	40.5%

Table 3: Encoding of a video segment using different GOP patterns.

increasing Q results in a significant reduction in C^*_{opt}. However, a large Q means more B frames between successive I/P frames, which is undesirable from the perspective of the decoder. Hence, Q should be chosen such that it provides a good compromise between the decoder complexity (and the associated decoding delay) and the multiplexing gain.

8 Summary

MPEG encoders often use a pre-specified GOP pattern to determine the types of compressed frames. The periodic and deterministic nature of this pattern can be used advantageously to reduce the bandwidth requirements of MPEG streams. By means of statistical multiplexing, we showed that the amount of bandwidth that must be allocated to a source while guaranteeing very stringent QoS requirements (i.e., no cell losses and negligible queueing delay) can be less than the source peak rate. Bandwidth gains are obtained by exploiting the structure of the GOP pattern of the multiplexed streams. The amount of bandwidth gain that can be achieved depends largely of the synchronization structure (i.e., the *arrangement*) of the multiplexed streams. We measure the bandwidth gain using the notion of effective bandwidth. We analyzed the impact of effective bandwidth reservation on call admission control, and derived several performance measures for the blocking probability of one of more new connection requests. Among all possible *arrangements*, we gave the form of a 'best' arrangement that has the optimal (i.e., minimum) effective bandwidth. An expression for the optimal effective bandwidth was also derived. Examples of actual MPEG streams from various compressed movies were presented and used to show the possible bandwidth gains that can be obtained from statistical multiplexing of MPEG streams. One aspect that was not addressed (mainly, due to space limitation) is the policing/shaping of streams that is required to achieve the bandwidth gains using our approach. This issue will be addressed in a future paper.

Appendix

Best Arrangement and the Optimal Effective Bandwidth

In this appendix, we prove that u^* in (36) is a best arrangement of N sources, and that $C(u^*, N) = C_{opt}(N)$ is given by (37). We first establish that for all N, $C(u^*, N)$ is given by the RHS of (37). Then, we show that $C_{opt}(N)$ is also given by the RHS of (37). The following result will be required in the proofs. Let u be an arbitrary arrangement of N streams. It is easy to show that in (21), $n_I \geq 1$ for any u and N. This fact follows from the definition of $C(u, N)$, Proposition 1, and that $I_{max} > P_{max} > B_{max}$.

An inspection of the structure of u^* reveals that there are exactly $m + 1$ streams whose phases differ, pairwise, by a nonnegative integer multiple of Q. Among those, there are a maximum of $w + 1$ streams that are in the same phase (m and w were defined in Table 1). Thus,

$$C(u^*, N) = \frac{(w+1)I_{max} + (m-w)P_{max} + (N-1-m)B_{max}}{N} \tag{39}$$

Consider an arbitrary arrangement $u = (u_1, \ldots, u_N)$. We will show that $C(u, N)$ satisfies:

$$C(u, N) \geq \frac{sI_{max} + lP_{max} + (N-s-l)B_{max}}{N} \tag{40}$$

with $s \geq w + 1$ and $s + l \geq m + 1$. If $C(u, N)$ satisfies (40), then $C(u, N)$ is greater than or equal the RHS of (39), which implies that u^* is a best arrangement and $C_{opt}(N)$ is given by (39).

First, suppose that the elements of u are distinct (i.e., $u_i \neq u_j$ for all $i \neq j$). This is possible when $N \leq L$ (thus, $w = 0$). There must be at least $m + 1$ streams whose phases differ pairwise by a nonnegative multiple of Q (in general, a set of distinct $kX + 1$ integers, where k and X are nonnegative integers and $X \neq 0$, must have at least $k + 1$ elements which differ, pairwise, by a multiple of X). Hence, $\overline{b}_{tot}(j) \geq I_{max} + mP_{max} + (N-1-m)B_{max}$ for some phase j. Therefore, $C(u, N)$ must satisfy (40) with $s = w + 1$ and $l = m - w$ ($w = 0$ in this case), and the assertion is true.

Next, suppose that the elements of u are not distinct. Let

$$\alpha \stackrel{\triangle}{=} \max_{0 \leq j \leq L-1} r_j \tag{41}$$

Clearly, $\alpha \geq \max\{2, w + 1\}$. We use the term *chain* to refer to a subset of the N streams whose phases differ pairwise by a multiple of Q (including those that have the same phase). It is easy to see that there can be no more than Q chains in a given arrangement. Let q be the number of chains ($q \leq Q$). Denote the chains by $W_1, W_2, \ldots, W_q$, with corresponding sizes $\eta_1, \eta_2, \ldots, \eta_q$ ($\sum_j \eta_j = N$). For each chain W_j, let $C_j(u, N)$ be the maximum aggregate peak rate divided by

N, where the maximization is taken only over the phases of the streams in W_j. For $j = 1, \ldots, q$, $C_j(u, N)$ can be given by:

$$C_j(u,N) = \frac{n_I^{(j)} I_{max} + n_P^{(j)} P_{max} + n_B^{(j)} B_{max}}{N} \tag{42}$$

where $n_I^{(j)} + n_P^{(j)} + n_B^{(j)} = N$. The total number of streams sending I or P frames during the phase of any stream in W_j is given by η_j. At least one of the chains, say W_1, contains α streams that are in the same phase, say phase i. Hence, it must be true that $C_1(u, N)$ results from the aggregate bit rate during phase i (recall that $r_i + z_i = r_j + z_j$ for phases i and j with $\mid i - j \mid =$ a multiple of Q). Therefore, $n_I^{(1)} = \alpha$. By definition, $C(u, N) = \max_j C_j(u, N)$, which implies

$$\begin{aligned} C(u,N) &\geq \frac{\sum_{j=1}^{q} C_j(u,N)}{q} \\ &= \frac{1}{q} \frac{I_{max} \sum_{j=1}^{q} n_I^{(j)} + P_{max} \sum_{j=1}^{q} n_P^{(j)} + B_{max} \sum_{j=1}^{q} (N - n_I^{(j)} - n_P^{(j)})}{N} \end{aligned} \tag{43}$$

Replacing $n_P^{(j)}$ by $\eta_j - n_I^{(j)}$, and with some rearrangements, (43) becomes:

$$C(u,N) \geq \frac{1}{q} \frac{(I_{max} - P_{max}) \sum_{j=1}^{q} n_I^{(j)} + P_{max} \sum_{j=1}^{q} \eta_j + B_{max} \sum_{j=1}^{q} (N - \eta_j)}{N} \tag{44}$$

Observe that $n_I^{(j)} \geq 1$ for $j = 2, \ldots, q$. Moreover, $n_I^{(1)} = \alpha$. Thus, $\sum_{j=1}^{q} n_I^{(j)} \geq \alpha + q - 1$ Since $\sum_{j=1}^{q} \eta_j = N$ and $I_{max} > P_{max}$, (44) reduces to

$$\begin{aligned} C(u,N) &\geq \frac{1}{q} \frac{(\alpha + q - 1)(I_{max} - P_{max}) + N P_{max} + (qN - N) B_{max}}{N} \\ &= \frac{\acute{s} I_{max} + \acute{l} P_{max} + (N - \acute{s} - \acute{l}) B_{max}}{N} \end{aligned} \tag{45}$$

where

$$\acute{s} \triangleq \frac{\alpha + q - 1}{q} \geq \max\left\{ \frac{1+q}{q}, \frac{w+q}{q} \right\} \tag{46}$$

$$\acute{l} \triangleq \frac{N - \alpha - q + 1}{q} \tag{47}$$

Therefore,

$$\acute{s} + \acute{l} = \frac{N}{q} \geq \frac{N}{Q} \geq \frac{mQ + 1}{Q} = m + \frac{1}{Q} \tag{48}$$

However, the expression for $C(u, N)$ must consist of integer numbers of I_{max} and P_{max}. Thus, $C(u, N)$ must satisfy (40) with $s + l \geq m + 1$ and $s \geq \max\{\lceil \frac{1+q}{q} \rceil, \lceil \frac{w+q}{q} \rceil\} = w + 1$. $\square$

References

[1] A. I. Elwalid and D. Mitra. Effective bandwidth for general Markovian traffic sources and admission control of high speed networks. *IEEE Journal on Selected Areas in Communications*, 1(3):329–343, June 1993.

[2] M. R. Frater, J. F. Arnold, and P. Tan. A new statistical model for traffic generated by VBR coders for television on the Broadband ISDN. *IEEE Trans. on Circuits and Systems for Video Technology*, 4(6):521–526, Dec. 1994.

[3] M. W. Garrett and M. Vetterli. Congestion control strategies for packet video. In *Proc. of Fourth Int. Workshop on Packet Video*, Aug. 1991. Kyoto, Japan.

[4] R. Guerin, H. Ahmadi, and M. Naghshineh. Equivalent capacity and its application to bandwidth allocation in high-speed networks. *IEEE Journal on Selected Areas in Communications*, 9(7):968–981, Sept. 1991.

[5] D. P. Heyman, A. Tabatabai, and T. V. Lakshman. Statistical analysis and simulation study of video teleconferencing traffic in ATM networks. *IEEE Trans. on Circuits and Systems for Video Technology*, 2(1):49–59, Mar. 1992.

[6] E. W. Knightly and H. Zhang. Traffic characterization and switch utilization using a deterministic bounding interval dependent traffic model. In *Proc. of IEEE INFOCOM '95*, pages 1137–1145, 1995.

[7] M. Krunz and H. Hughes. A traffic model for MPEG-coded VBR streams. In *Proc. of the ACM SIGMETRICS/PERFORMANCE '95 Conference*, pages 47–55, May 1995.

[8] P. Pancha and M. El Zarki. Bandwidth-allocation schemes for variable-bit-rate MPEG sources in ATM networks. *IEEE Trans. on Circuits and Systems for Video Technology*, 3(3):190–198, June 1993.

[9] D. Reininger et al. Statistical multiplexing of VBR MPEG compressed video on ATM networks. In *Proc. of IEEE INFOCOM '93*, volume 3, pages 919–926, 1993.

[10] O. Rose. Statistical properties of MPEG video traffic and their impact on traffic modeling in ATM systems. In *Proceedings of the 20th Annual Conference on Local Computer Networks*, Minneapolis, MN, 1995.