An Efficient Solution to Traffic Characterization of VBR Video in Quality-of-Service Networks*

Jörg Liebeherr　　*Dallas E. Wrege*

Technical Report CS-96-10
Department of Computer Science
University of Virginia
Charlottesville, VA 22903
Email: {jorg|dallas}@cs.virginia.edu

Abstract

A network that offers deterministic, i.e., worst-case, quality-of-service (QoS) guarantees to variable-bit-rate (VBR) video must provide a resource reservation mechanism that allocates bandwidth, buffer space, and other resources for each video connection. Such a resource reservation scheme must be carefully designed, otherwise network resources are wasted. A key issue for the design of a resource reservation scheme is the selection of a *traffic characterization* that specifies the traffic arrivals on a video connection. The traffic characterization should accurately describe the actual arrivals so that a large number of connections can be supported, but it must also map directly into efficient traffic policing mechanisms that monitor arrivals on each connection. In this study, we present a fast and accurate traffic characterization method for stored VBR video in networks with a deterministic service. Our characterization approach approximates the so-called *empirical envelope* of a video sequence. We use this approximation to obtain a traffic characterization that can be efficiently policed by a small number of leaky buckets. We present a case study where we apply our characterization method to networks that employ a dynamic resource reservation scheme with renegotiation. We use traces from a set of 25-30 minute MPEG sequences to evaluate our method against other characterization schemes from the literature.

Key Words: Multimedia Networks, Traffic Characterization, Bounded Delay Service, Quality-of-Service, Deterministic Service, Renegotiation, Resource Reservation, VBR Video.

*This work is supported in part by the National Science Foundation under Grant No. NCR-9309224. Some results from this paper were presented at the IEEE Infocom '96 Conference.

1 Introduction

With the advent of high-speed packet-switching networks, it has become feasible to design multimedia networks that can transmit compressed variable-bit-rate (VBR) video in real time. However, video traffic is distinct from traditional data traffic in that video connections require guarantees on the *quality-of-service* (QoS) they receive, for example, bounded delay, minimum throughput, and bounded jitter. One approach to provide QoS guarantees is for the network to maintain a *resource reservation scheme* that allocates resources such as bandwidth and buffer space for each video connection. Since overallocating resources to a connection results in a low network utilization, a resource reservation scheme must be carefully designed. In this study we consider resource reservation schemes for QoS networks with a *deterministic service*, that is, a service that can provide worst-case delay guarantees to all packets on a video connection.

One key component of a resource reservation scheme is the *traffic characterization* used to specify the traffic arrivals on a video connection. A deterministic service requires a *deterministic* traffic characterization that provides an upper bound on traffic arrivals. Deterministic traffic characterizations have been used in previous work to support both deterministic and probabilistic QoS guarantees [6, 8, 16, 25, 33]. It is important to specify the traffic on a connection as accurately as possible since the traffic characterization will be used in *admission control tests* that verify if sufficient resources are available within the network to support the traffic on the connection at the desired QoS. If the traffic characterization is too pessimistic in describing the traffic, the admission control tests will overestimate the resource requirements of a connection, resulting in poor network utilization. Due to the complex timely correlations of VBR video sequences, elaborate traffic characterizations have been devised which achieve a high degree of accuracy [9, 11, 15, 27, 28].

While admission control functions require that traffic characterizations are accurate in describing the worst-case traffic, *traffic policing mechanisms* that monitor in real time if the traffic submitted to the network conforms to its traffic characterization require a simple traffic characterization [26]. Therefore, the choice of traffic characterization method is a tradeoff between the high accuracy preferred by admission control tests and the simplicity required for implementing traffic policing mechanisms.

An important class of traffic characterizations describes the worst-case traffic on a connection in terms of a piecewise-linear function [15]. If the resulting function is concave, the traffic characterization can be implemented with a set of *leaky buckets*[1] [6], one for each linear segment of the function. Since a leaky bucket can be implemented with a single counter and a single timer [26], leaky buckets seem to satisfy the need for simple traffic characterizations. In a previous study [33], we showed that concave piecewise-linear functions ("leaky buckets") are capable of accurately characterizing VBR video traffic. However, the number of leaky buckets needed for an accurate characterization was shown to be large. For example, up to a dozen leaky buckets are needed to accurately characterize MPEG-I video streams [1]. Since practical considerations limit the number of available leaky

[1] In this paper, a *leaky bucket* [32] is equivalent to the *Generic Cell Rate Algorithm (GCRA)* as specified in [1].

buckets to a small value, e.g., the limit is two for ATM connections [33], methods are needed that yield an accurate VBR traffic characterization, yet, with only a small number of leaky buckets.

In this study, we present a solution to the problem of constructing an accurate traffic characterization for stored VBR video with few leaky buckets. Our solution approach is based on the so-called *empirical envelope* of a video sequence [33]. As we will discuss in Section 3, the empirical envelope for a connection is the most accurate deterministic traffic characterization. However, the empirical envelope itself is not practical for use in QoS networks for two reasons: (1) the empirical envelope requires a large number of parameters which are computationally expensive to produce, and (2) the traffic specified by the empirical envelope cannot be policed using simple traffic policing mechanisms. The characterization method presented in this paper addresses both of these problems. First, we determine an approximation of the empirical envelope based on a subset of its parameters that can be computed quickly. We then use this approximation to determine a traffic characterization that can be policed by a small number of leaky buckets.

We demonstrate the effectiveness of our method in networks with a deterministic service using traffic traces of two 25-30 minute MPEG encoded video segments [10]. Our examples illustrate the minimum number of empirical envelope parameters and leaky bucket mechanisms needed to obtain an accurate traffic characterization. We show that only 200 out of a total 40,000 envelope parameters and three leaky bucket mechanisms are sufficient to produce traffic characterizations leading to utilizations within 91% of the results achievable with the empirical envelope. In a case study, we show how our methods can be employed in networks with dynamic resource reservation schemes, i.e., where the traffic characterization can be renegotiated after the connection is established. We demonstrate that a renegotiation scheme can yield increases in network utilization of 20-35%. The fast characterization method developed in this paper is well-suited to dynamic reservation schemes since renegotiation requires the calculation of multiple traffic characterizations.

The remainder of this paper is structured as follows. In Section 2 we review deterministic traffic characterizations of VBR video traffic and discuss previous work on selecting traffic parameters. We present our traffic characterization method in Sections 3 and 4; In Section 3 we describe a method for approximating the empirical envelope using only a small number of envelope parameters, and in Section 4 we describe an algorithm for selecting leaky bucket parameters. In Section 5 we present a case study where we apply our method to construct a renegotiation scheme for a deterministic service.

2 Deterministic Characterization of VBR Video Traffic

In this section, we present a framework for traffic characterization in QoS networks with a deterministic service. We first review a general approach to traffic characterization presented in [6, 21] that can be used to describe the traffic for most deterministic traffic models. We then discuss previous work on selecting parameters for particular traffic models to specify VBR video traffic.

2.1 Traffic Constraint Functions A^*

Let A denote the actual traffic on a connection, where $A[\tau, \tau+t]$ denotes the traffic arrivals in time interval $[\tau, \tau+t]$. Then, a worst-case characterization of the traffic A is given by a *traffic constraint function* A^* which provides an upper bound on A. A traffic constraint function A^* should satisfy two important properties, namely *time-invariance* and *subadditivity* [6, 21]. A function A^* provides a time-invariant bound for A if for all times $\tau \geq 0$ and $t \geq 0$ the following holds [6]:

$$A[\tau, \tau+t] \leq A^*(t) \tag{1}$$

Since a time-invariant traffic constraint function A^* bounds the maximum traffic over any time interval of length t, the admission control tests can be made independent of the starting time of a connection. A traffic constraint function A^* is subadditive if it satisfies the following inequality:

$$A^*(t_1) + A^*(t_2) \geq A^*(t_1 + t_2) \qquad \forall t_1, t_2 \geq 0 \tag{2}$$

A subadditive traffic constraint function allows the arrivals on a connection to attain the bound given by A^*. In other words, it is feasible that $A[\tau, \tau+t] = A^*(t)$ for any $t \geq 0$. Even though traffic constraint functions that are time-invariant but not subadditive have been proposed, e.g., [15], we point out that any such traffic constraint function A_1^* can be replaced by a subadditive function A_2^* such that $A_2^*(t) \leq A_1^*(t)$ for all $t \geq 0$. Finally, we wish to add that admission control tests for QoS networks generally assume that traffic constraint functions are both time-invariant and subadditive [6, 21, 25]. In the following, we call a traffic constraint function A^* for A *viable* if it satisfies both equations (1) and (2).

Practical traffic characterizations are obtained from a parameterized *traffic model* which expresses the arrivals admitted on a connection by some policing mechanism. For example, the (σ, ρ) traffic model [6] describes the worst-case traffic admitted by a leaky bucket mechanism with a burstiness parameter σ and a rate parameter ρ. We denote the traffic constraint function that provides a bound on the maximum traffic conforming to the (σ, ρ) model by B^*, where B^* is given by the following linear constraint [6]:

$$B^*(t) = \sigma + \rho t \qquad \text{for all } t \geq 0 \tag{3}$$

A generalization of the (σ, ρ) model is the $(\vec{\sigma}, \vec{\rho})$ traffic model [7, 33] which corresponds to a traffic policing mechanism where multiple leaky buckets are connected in series. For a connection that conforms to the $(\vec{\sigma}, \vec{\rho})$ traffic model with a set of m pairs $\{(\sigma_i, \rho_i)\}_{1 \leq i \leq m}$, the amount of traffic admitted to the network is limited by each of the (σ_i, ρ_i) pairs. The resulting traffic constraint function, denoted as B^*_m, is a concave function consisting of m piecewise-linear segments [7, 33]:

$$B^*_m(t) = \min_{1 \le i \le m} \{\sigma_i + \rho_i t\} \tag{4}$$

Note that B^* in equation (3) is identical to B^*_1 in equation (4). All traffic characterization methods considered in this paper determine traffic constraint functions that conform to the $(\vec{\sigma}, \vec{\rho})$ traffic model.

2.2 Previous Work

Several studies have considered deterministic traffic characterizations for VBR video traffic using the (σ, ρ) traffic model that corresponds to the leaky bucket policing mechanism. Most studies use only a single (σ, ρ) pair and explore the dependencies between the burstiness parameter σ and the rate parameter ρ [22, 24, 27, 28, 30]. In particular, for any fixed choice of rate ρ, the burst parameter σ should be selected as small as possible, that is [22]:

$$\sigma = \inf\{\hat{\sigma} \mid \hat{\sigma} + \rho t \ge A[\tau, \tau + t],\ \forall t,\ \tau \ge 0\} \tag{5}$$

Equation (5) illustrates a tradeoff between buffer space (i.e., burst) and bandwidth (i.e., rate) when selecting leaky bucket parameters. By combining the dependency in equation (5) with all rates ρ between the average and peak rate of a connection, one obtains an infinite candidate set of (σ, ρ) pairs from which all leaky bucket parameters should be selected. Note that it is computationally demanding to determine this candidate set of (σ, ρ) pairs.

Many schemes select parameters σ and ρ according to either network resource availability or the relative importance of bandwidth and buffer space. Pancha and El Zarki [24] choose parameters by fixing the burstiness parameter σ according to available buffer space, while the choice of (σ, ρ) in [3] depends on the relative availability of unallocated bandwidth and buffer space. An approach discussed by Guillemin et. al. in [14] assigns relative importance parameters α and β to buffer space and bandwidth, respectively; the pair (σ, ρ) is selected to minimize the quantity $\sigma^\alpha \cdot \rho^\beta$. The authors note that a "natural" choice is the case where both resources have the same cost, that is, $\alpha = \beta = 1$. A drawback of all of these methods is that they do not strive for high network utilization as a design goal. Also, all of these approaches consider the selection of parameters for only a single leaky bucket mechanism.

Guillemin et. al. present two heuristic algorithms in [14] that select a leaky bucket pair (σ, ρ) to approximate an "ideal" probabilistic traffic characterization, the so-called *time-ϵ quantile function $M_\epsilon(t)$* associated with a source. The heuristic algorithms are similar to the characterization method proposed in this paper in that they first determine a function that describes the traffic on a connection and then determine parameters based on this function. Assuming that N_t is a random variable specifying the number of packets generated over any interval of length t, a function $M_\epsilon(t)$ is used to specify with probability $1 - \epsilon$ the maximum traffic arrivals n in any interval of length t [14, 30]:

$$M_\epsilon(t) = \inf\{n,\ Pr\{N_t \geq n\} \leq \epsilon\} \tag{6}$$

The quantity $M_\epsilon(t)/t$ specifies the rate of the video sequence over multiple time scales t. The first heuristic in [14] selects a leaky bucket parameter $(\sigma + \rho t)$ such that the maximum difference between $M_\epsilon(t)/t$ and the "normalized" leaky bucket curve $(\sigma + \rho t)/t$ is minimized. The second heuristic minimizes the area y *between* the normalized curves $M_\epsilon(t)/t$ and $(\sigma + \rho t)/t$ over an interval $[0, T_0]$. As noted in [14], the selection of parameters for the second heuristic is heavily dependent on the choice of T_0 which is not set explicitly in the paper.

While the focus of our paper is on finding a traffic characterization for VBR video traffic, other studies exist that explore the benefits of reducing the burstiness of VBR traffic through either (1) *shaping* the traffic by spacing packets before submitting them to the network [12, 17, 18] or (2) sending packets early with respect to their playback time at the receiving application via *workahead smoothing* [23, 28, 31]. These techniques involve modification of the traffic A that is submitted to the network on a connection by buffering at either the sender, receiver, or a combination of both. While shaping and smoothing techniques have been shown to increase the achievable network utilization, these methods are orthogonal to the traffic characterization problem studied in this paper. Note that even after traffic is shaped or smoothed, a characterization method such as the one developed in this paper must be available to determine an accurate and policable characterization for the traffic submitted to the network.

3 A Fast Characterization Method for VBR Video

None of the the characterization approaches for VBR video with leaky buckets described above attempt to maximize the number of admissible connections in a QoS network. In previous work with Knightly and Zhang [33], we presented a traffic characterization, referred to as the "empirical envelope", that maximized the resource utilization. We showed how to approximate the empirical envelope with leaky buckets, however, the number of parameters of the resulting traffic constraint function was considerable: up to 12 leaky buckets were needed for an accurate characterization of an MPEG video sequence [33]. Also, the computational complexity of the characterization algorithms was substantial.

Here we present a method to obtain VBR video traffic characterizations that can be policed by a small, fixed number of leaky buckets. The computational complexity our our method is low and efficient as compared to the methods in [14, 33].

In Subsection 3.1 we discuss the tradeoffs of traffic characterization methods that are based on the empirical envelope. Following, in Subsections 3.2 and 3.3, we present and evaluate the new solution approach to VBR video characterization.

3.1 The Empirical Envelope E^*

The tightest traffic constraint function for a given traffic source is its *empirical envelope*, denoted by E^* [2, 33]. The empirical envelope E^* of a video sequence is optimal in the sense that, for any subadditive traffic constraint function A^* that satisfies equation (1), $A^*(t) \geq E^*(t)$ for all t. The empirical envelope E^* is given by the following equation [2, 33]:

$$E^*(t) = \sup_{\tau \geq 0} A[\tau, \tau + t] \qquad \forall t \geq 0 \tag{7}$$

Note from equation (7) that E^* is subadditive.

The following method presented in [33] obtains the empirical envelope of a given video sequence consisting of N frames with fixed inter-frame time r. We assume that frames are fragmented into 53-byte ATM cells with a payload of 48 bytes each, and these cells are transmitted at equally-spaced intervals over the frame time r. If the sequence of frame sizes of a video sequence is given by $\{f_1, f_2, \ldots, f_N\}$, then the empirical envelope E^* can be constructed by calculating [33]:

$$E^*(ir) = \max_{0<k<N-i+1} \sum_{j=k}^{k+i-1} f_j \qquad \text{for } i = 1, 2, \ldots N \tag{8}$$

Note that equation (8) defines N parameters $\{E^*(ir) \mid 1 \leq i \leq N\}$ for the empirical envelope, where $E^*(r)$ is equal to the largest frame in the video sequence, $E^*(2r)$ is equal to the largest two-frame sequence, etc. The values of the empirical envelope at times that are not multiples of the frame time are obtained by spacing the cells in $E^*(ir) - E^*((i-1)r)$ evenly over the frame time $[(i-1)r, ir]$.

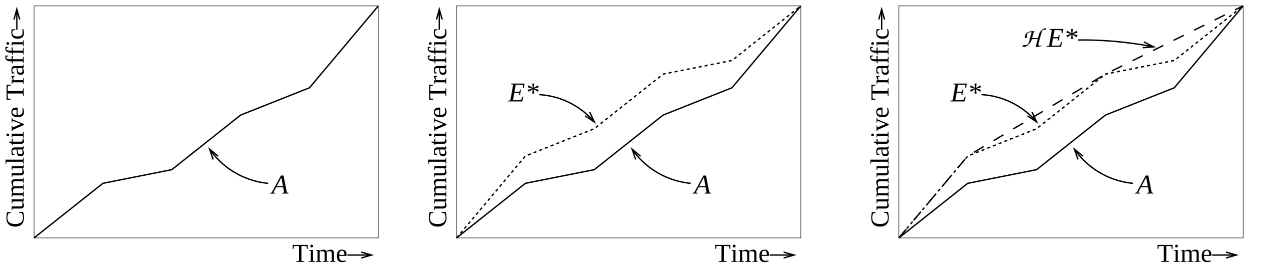
Since the empirical envelope E^* does not conform to a parameterized traffic model, it is difficult to police. In previous work, we showed how to determine a $(\vec{\sigma}, \vec{\rho})$-model traffic characterization based on the *concave hull of* E^*, which we denote by $\mathcal{H}E^*$ [33].[2] Since the function $\mathcal{H}E^*$ is the smallest piecewise-linear concave function larger than E^* [5], $\mathcal{H}E^*$ is most accurate traffic characterization that can be policed by leaky buckets.

In Figure 1 we illustrate the traffic characterization method from [33] with an example. The cumulative traffic arrivals A for a traffic source are depicted in Figure 1(a). Figures 1(b) and 1(c) show the empirical envelope E^* and the concave hull $\mathcal{H}E^*$, respectively, for this traffic source. The relationship between A, E^*, and $\mathcal{H}E^*$ for an actual MPEG-encoded video sequence is illustrated in Figure 2. Figure 2(a) shows a trace of 250 frames of an MPEG movie. The traffic is packetized into ATM cells with 48-byte payloads, and we plot the number of cells as a function of the frame sequence number. In Figure 2(b), we illustrate the cumulative cells A for the trace in Figure 2(a), and we also plot the empirical envelope E^* and its concave hull $\mathcal{H}E^*$.

3.2 Approximating the Envelope with Extrapolations

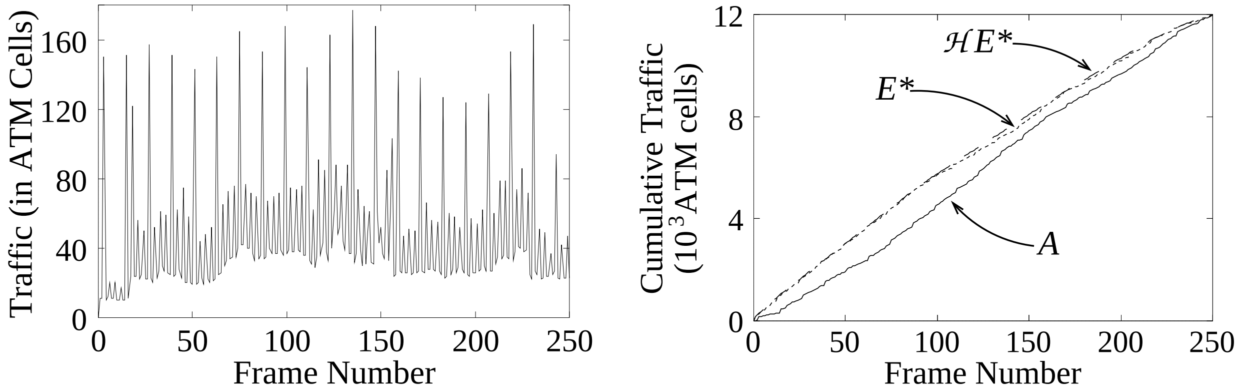
The traffic characterization method outlined in the previous subsection was shown [33] to produce very accurate traffic characterizations based on the empirical envelope. However, the empirical

[2] In this document, we use $\mathcal{H}$ to denote the concave hull operator, that is, $\mathcal{H}f$ is the concave hull of the function f.



(a) Cumulative traffic A. (b) Empirical envelope E^*. (c) Concave approximation $\mathcal{H}E^*$.

Figure 1: Characterization approach using the empirical envelope [33].



(a) MPEG traffic trace. (b) Cumulative traffic and constraint functions.

Figure 2: Functions A, E^* and $\mathcal{H}E^*$ for an actual MPEG trace.

envelope requires a large number of parameters, that is, one parameter per frame in the sequence. The number of operations required to compute all N parameters of the empirical envelope E^* for a video sequence with N frames is $O(N^2)$. Since N is generally large, e.g., it exceeds 200,000 for most feature-length motion pictures, it may not be possible to calculate the empirical envelope in real-time. Note that while the characterization $\mathcal{H}E^*$ uses fewer parameters, determining $\mathcal{H}E^*$ requires knowledge of the entire empirical envelope, and so its production is also computationally expensive. We therefore seek other traffic constraint functions that closely approximate the envelope but can be calculated with fewer parameters.

Here we present two methods for obtaining viable traffic constraint functions defined for all times t that are derived only from the first k parameters of the empirical envelope, i.e., $E^*(r)$, $E^*(2r)$, $\ldots$, $E^*(kr)$. Both methods construct a traffic constraint function through extrapolation of these k parameters. We first discuss the best-possible extrapolation based on the first k parameters of E^* and then present a simple characterization that can obtained with a fast extrapolation technique.

Any viable traffic constraint function obtained from the first k parameters of the envelope must be at least as large as the empirical envelope E^* for all times t. Since we know that E^* is a subadditive function, the best extrapolation is given by the *largest subadditive extrapolation of* $\{E^*(ir)\}_{1\leq i\leq k}$. We denote this largest subadditive extrapolation by E^*_k, where E^*_k is obtained by calculating:

$$E_k^*(ir) = \begin{cases} E^*(ir) & \text{for } i \leq k \\ \min\limits_{1 \leq j < i} \{E_k^*(jr) + E_k^*((i-j)r)\} & \text{for } i > k \end{cases} \tag{9}$$

E_k^* is equal to the empirical envelope for the first k frame times, and E_k^* is defined for subsequent times by exploiting the requirement for subadditivity of E_k^*.[3]

Although the function E_k^* is the tightest traffic constraint function that can be obtained directly from the first k parameters of the envelope, the production of E_k^* requires a large number of computations. Specifically, we see from equation (9) that the number of computations required to construct E_k^* is $O(N^2)$, the same number required for computing the empirical envelope itself. Since we seek an approximation that can be computed efficiently, we turn to other approximation schemes, and we will use E_k^* as a benchmark for other approximations.

As a more efficient extrapolation, we next consider a function that is obtained by simply repeating the first k parameters $\{E^*(ir)\}_{1 \leq i \leq k}$ for all times t. We call such a function the *repetition extrapolation*, which we denote by R_k^*. R_k^* is given as follows:

$$R_k^*(t) = \lfloor \frac{t}{kr} \rfloor E^*(kr) + E^*(t - \lfloor \frac{t}{kr} \rfloor (kr)) \quad \text{for } t \geq 0 \tag{10}$$

Observe that R_k^* can be immediately obtained from the first k parameters of the envelope, and so the computational complexity of computing R_k^* is $O(kN)$. For small values of k, R_k^* can be computed much more efficiently than the entire empirical envelope E^*.

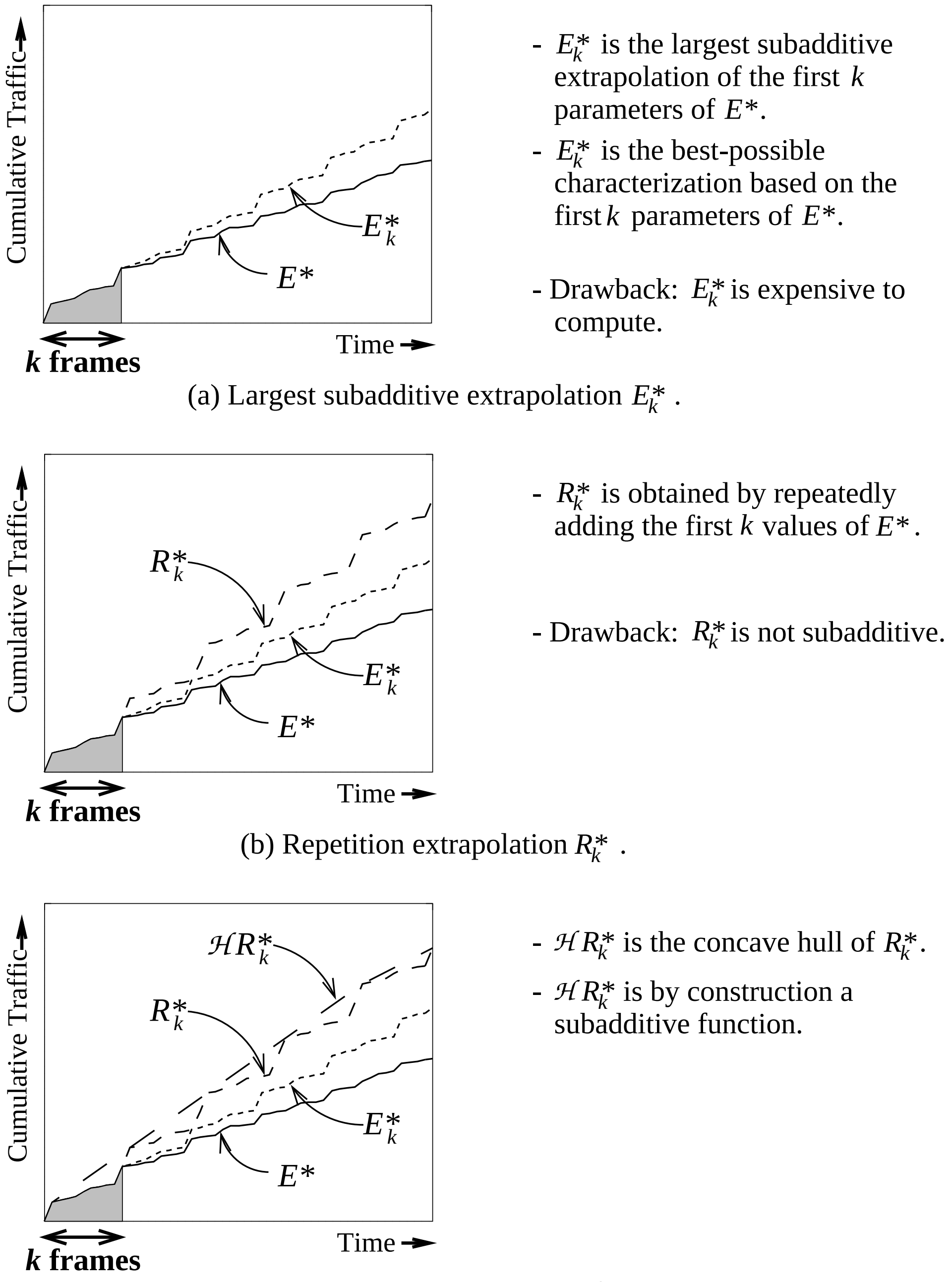
Although R_k^* provides a time-invariant bound on the traffic arrivals A in terms of equation (1), it is not necessarily subadditive and hence does not satisfy our requirement for a traffic constraint function. To remedy this problem we consider yet another function $\mathcal{H}R_k^*$, the *concave hull of* R_k. The concave hull $\mathcal{H}R_k^*$ is by construction a viable traffic constraint function since its subadditivity follows from its concavity. Note that $\mathcal{H}R_k^*$ can be expressed in terms of the $(\vec{\sigma}, \vec{\rho})$ model as follows:

$$\mathcal{H}R_k^*(t) \equiv B_n^* = \min_{1 \leq i \leq n} \{\overline{\sigma}_i + \overline{\rho}_i t\}, \tag{11}$$

where parameters $\overline{\sigma}_i$ and $\overline{\rho}_i$ are determined by some appropriate algorithm to compute the concave hull of a function, e.g. [33].

We review the extrapolation methods in Figure 3. Figure 3(a) illustrates the relationship between the empirical envelope E^* and its approximation E_k^*, the largest subadditive extrapolation of the first k parameters of E^*. E_k^* is the most accurate traffic characterization that can be obtained from the first k values of the empirical envelope. The repetition extrapolation R_k^*, depicted in Figure 3(b), can be efficiently computed by repeating the first k parameters of the empirical envelope. However, R_k^* is not subadditive and therefore is not a viable traffic constraint function. The concave hull $\mathcal{H}R_k^*$, shown in Figure 3(c), is by construction subadditive and can be used as a deterministic traffic constraint function.

[3] Note that equation (9) only defines E_k^* for times that are multiples of the frame time r. Similar to the production of the empirical envelope in equation (8), the values for intermediate values of E_k^* are determined by spacing cells evenly over each frame.



- E_k^* is the largest subadditive extrapolation of the first k parameters of E^*.
- E_k^* is the best-possible characterization based on the first k parameters of E^*.
- Drawback: E_k^* is expensive to compute.

(a) Largest subadditive extrapolation E_k^* .

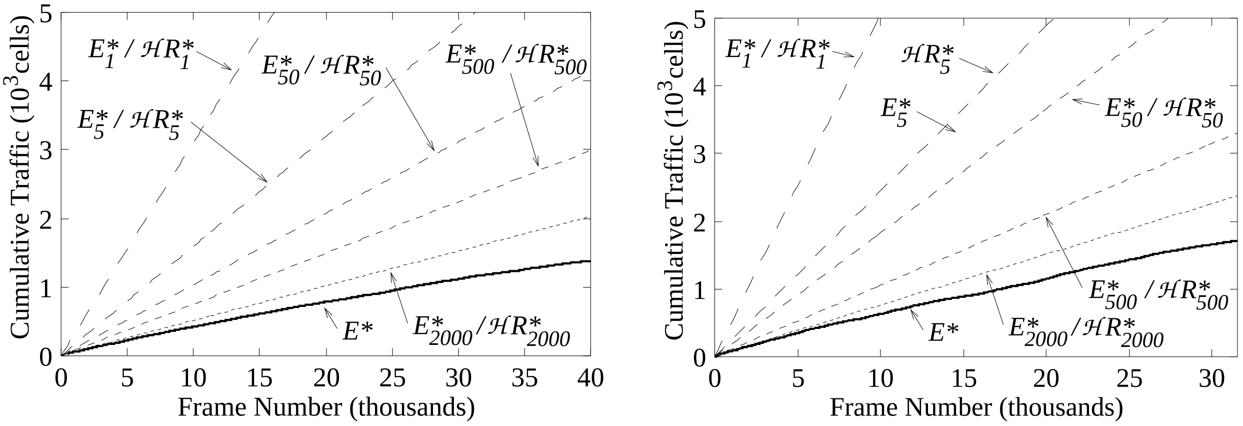
- R_k^* is obtained by repeatedly adding the first k values of E^*.
- Drawback: R_k^* is not subadditive.

(b) Repetition extrapolation R_k^* .

- $\mathcal{H}R_k^*$ is the concave hull of R_k^*.
- $\mathcal{H}R_k^*$ is by construction a subadditive function.

(c) Concave approximation $\mathcal{H}R_k^*$ of R_k^*.

Figure 3: Approximations of the empirical envelope.



(a) *Park*

(b) *News*

Figure 4: Traffic constraint functions.

A problem that remains to be solved is the potentially large number of $(\overline{\sigma}_i, \overline{\rho}_i)$ pairs needed for the concave hull $\mathcal{H}R_k^*$, mandating a large number of leaky bucket mechanisms for a single connection. We will address this problem in Section 4, where we present an algorithm that approximates $\mathcal{H}R_k^*$ with a traffic characterization that can be policed by a fixed (and small) number of leaky buckets.

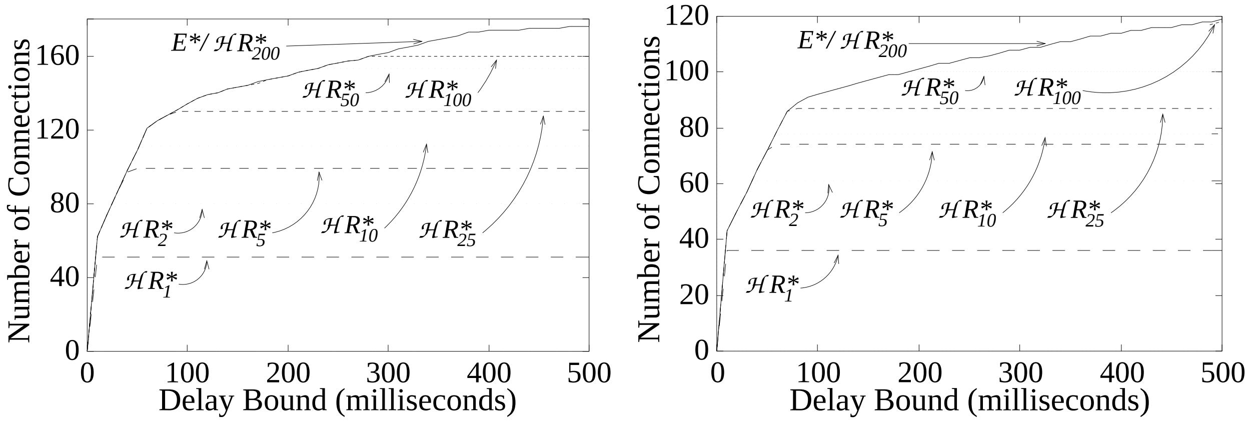
3.3 Evaluation

Here we evaluate the accuracy of traffic characterizations E_k^* and $\mathcal{H}R_k^*$ as approximations of the empirical envelope using actual traces of MPEG-compressed video. We are are interested in determining the size of k needed to generate an accurate characterization for a VBR video source.

We use two MPEG traces in the evaluation: one from the entertainment film *Jurassic Park* ("*Park*"), and the second from a news broadcast ("*News*"). These sequences were encoded in software with the Berkeley MPEG-encoder [29]. Both *Park* and *News* are thirty-minute video sequences with a frame size of 384x288 and frame pattern IBBPBBPBBPBB. We note that *News* generates burstier traffic than *Park*; the ratio of the peak rate to the average rate for *News* and *Park* are 6 and 4, respectively.

Figure 4(a) and 4(b) illustrate traffic constraint functions for the *News* and *Park* traces, respectively. We show the empirical envelope E^* as well as E_k^* and $\mathcal{H}R_k^*$ for $k \in \{1, 5, 50, 500, 2000\}$. For each traffic constraint function, we plot the cumulative number of cells as a function of the frame sequence number. In both graphs, the empirical envelope E^* is shown as a bold solid curve, while the functions E_k^* and $\mathcal{H}R_k^*$ are depicted by dotted and dashed curves, respectively. As expected, the approximation functions estimate the empirical envelope E^* more accurately for larger values of k.

A key observation from Figure 4 is that $\mathcal{H}R_k^* \approx E_k^*$ for most values of k; only $\mathcal{H}R_5^*$ and E_5^* for the *News* sequence in Figure 4(b) differ considerably. Since E_k^* is the tightest traffic characterization that can be produced from k frames of the empirical envelope, we note that the concave hull of the



(a) *Park*

(b) *News*

Figure 5: Utilization comparison.

repetition extrapolation $\mathcal{H}R^*_k$ is an accurate approximation of E^*_k, validating our selection of $\mathcal{H}R^*_k$ for the characterization.

We next consider the utilizations that can be achieved at a network switch using traffic constraint functions $\mathcal{H}R^*_k$. We assume a single multiplexer that operates at 155 Mbps, a data rate that corresponds to OC-3, and we further assume that the switch transmits its packets with a First-Come-First-Served (FCFS) discipline.[4] Figure 5 illustrates the network utilization obtained at a multiplexer using E^* as well as $\mathcal{H}R^*_k$ for various values of k. All connections at a multiplexer are assumed to be of the same type (either *Park* or *News*) and have identical delay bounds (in the range $0 \leq d \leq 500$ msec). For each characterization, we plot the maximum number of connections that can be admitted as a function of the delay bounds of those connections. For example, Figure 5(b) shows that the traffic constraint function $\mathcal{H}R^*_2$ can be used to support 61 *News* connections for delay bounds larger than 35 ms.

The general trend in both graphs is that the number of connections accepted using $\mathcal{H}R^*_k$ as the traffic constraint function increases with k. An important observation is that the function $\mathcal{H}R^*_{200}$ admits the same number of connections as the empirical envelope E^* for delay bounds up to 500 milliseconds. Thus, we can use our approximation function $\mathcal{H}R^*_k$ based on the first 200 parameters of the envelope (i.e., $\mathcal{H}R^*_{200}$) to characterize both video sequences for the delay bound range considered; we achieve the same utilization using $\mathcal{H}R^*_{200}$ as we would using the empirical envelope with 40,000 parameters.

However, while the function $\mathcal{H}R^*_k$ provides an accurate traffic characterization for VBR video, the number of leaky buckets required to enforce $\mathcal{H}R^*_k$ may be too large. For example, 12 leaky bucket mechanisms are needed to police $\mathcal{H}R^*_{200}$ for the *News* sequence (i.e., $\mathcal{H}R^*_{200} \equiv B^*_{12}$). Unfortunately, the number of leaky buckets available to monitor a connection in a real network is typically limited

[4] The exact admission control test for FCFS multiplexers is given by $d \geq \sum_{j \in \mathcal{N}} A^*_j(t) - t + s$ for all $t \geq 0$ [6]. In this admission control test, $\mathcal{N}$ denotes the set of all connections at a multiplexer, d denotes the maximum delay at the multiplexer, and s denotes the transmission time of a cell.

to only two or three. This problem is addressed in the next section where we present an algorithm that reduces the number of leaky buckets used to characterize a VBR video source.

4 Leaky Bucket Parameter Selection

At this point, we have obtained an accurate traffic characterization $\mathcal{H}R_k^*$ that conforms to the $(\vec{\sigma}, \vec{\rho})$ traffic model. We write $\mathcal{H}R_k^* \equiv B_n^*$ to indicate that n is the number of $(\overline{\sigma}_j, \overline{\rho}_j)$ pairs for the traffic characterization. Since n can be large, and since the number m of leaky buckets available to police a video connection is small, we will use a curve-fitting method that reduces B_n^* to B_m^* with $m < n$.

We formulate the problem as follows. Given a function $B_n^* = \min_{1 \leq i \leq n}\{\overline{\sigma}_j + \overline{\rho}_j t\}$, we want to find a set of $m < n$ (σ_i, ρ_i) pairs that determine a traffic constraint function B_m^*:

$$B_m^*(t) = \min_{1 \leq i \leq m} \{\sigma_i + \rho_i t\}, \tag{12}$$

such that $B_m^*(t) \geq B_n^*(t)$ for all t and B_m^* is a tight approximation of B_n^*. We use a cost function $C(B_m^*, B_n^*)$ to express the closeness of B_m^* to B_n^*. Assuming that we have such a cost function available, we select parameters (σ_i, ρ_i) for B_m^* as solutions to the following optimization problem:

Minimize $C(B_m^*, B_n^*)$
Subject to $B_m^*(t) \geq B_n^*(t)$ $\forall t \geq 0$.

In the remainder of this section we describe the cost function $C(B_m^*, B_n^*)$ and present a heuristic algorithm to solve the optimization problem.

4.1 Cost Function $C(B_m^*, B_n^*)$

The cost function $C(B_m^*, B_n^*)$ is introduced to express the difference between the two functions B_m^* and B_n^*. While the function B_m^* should approximate B_n^* as tightly as possible, it is not clear that the best cost function C is a simple or obvious choice such as the absolute distance between B_m^* and B_n^*. For example, since the burstiness of VBR video limits the number of admitted connections at small delay bounds, it is important that the function B_m^* approximates B_n^* closely for small values of t.

We have evaluated a number of candidate cost functions of the following general form:

$$C(B_m^*, B_n^*) = \int_0^{T_0} \frac{(B_m^*(t) - B_n^*(t))^\alpha}{(t+1)^\beta B_n^*(t)^\gamma}\, dt, \tag{13}$$

where T_0 and the exponents α, β, and γ in equation (13) determine the shape of the cost function. For example, a selection of $(2, 0, 0)$ for the (α, β, γ)-tuple results in an approximation where the square of the difference between B_m^* and B_n^* is minimized. However, a least-squares model may not be appropriate since the function B_m^* is required to be larger than B_n^*. We found the following cost function to result in accurate characterizations for the class of small delay bounds ($d \leq 500$ ms):

$$C(B_m^*, B_n^*) = \int_0^{k\,r} \frac{B_m^*(t) - B_n^*(t)}{B_n^*(t)}\, dt \tag{14}$$

This cost function measures the amount that B_m^* overestimates the function B_n^* relative to the size of B_n^*.

```
Input:   A set of $n$ pairs $\{(\overline{\sigma}_j, \overline{\rho}_j) \mid j = 1, \ldots, n\}$ that define the function $B_n^*$, the
         number $m$ of available $(\sigma_i, \rho_i)$ pairs, a cost function $C(B_m^*, B_n^*)$, and a
         sensitivity parameter $\epsilon$.
Output:  A set of $m$ pairs $\{(\sigma_i, \rho_i) \mid i = 1, \ldots, m\}$ that define the traffic constraint
         function $B_m^*$.

 1. Procedure Parameterize ($B_n^*$, $m$, $C(B_m^*, B_n^*)$, $\epsilon$)
 2.     For $i = 1$ To $m$                    /* Initialize $(\sigma_i, \rho_i)$ */
 3.         $\sigma_i \leftarrow \overline{\sigma}_{\lfloor \frac{in}{m} \rfloor}$
 4.         $\rho_i \leftarrow \overline{\rho}_{\lfloor \frac{in}{m} \rfloor}$
 5.     End For
 6.     Do                                    /* Greedy modifications */
 7.         Cost $\leftarrow C(B_m^*, B_n^*)$
 8.         For $i = m$ Down To 1
 9.             Select $(\sigma_i, \rho_i)$ to minimize $C(B_m^*, B_n^*)$, where $\sigma_{i-1} \leq \sigma_i \leq \sigma_{i+1}$
10.         End For
11.     While ( Cost $- C(B_m^*, B_n^*) > \epsilon$ )
12.     Output $B_m^* \leftarrow \min_{1 \leq i \leq m} \{\sigma_i + \rho_i t\}$
13. End Procedure
```

Figure 6: Parameterization algorithm.

4.2 A Heuristic Algorithm

As we mentioned in Section 2, the number of possible (σ, ρ) pairs is infinite, and the selection of a set of pairs that minimizes $C(B_m^*, B_n^*)$ is a combinatorial problem. For this reason, we turn to heuristic approximations for the optimization problem. Here we present a heuristic algorithm that determines m (σ_i, ρ_i) pairs to produce a traffic constraint function B_m^* with low cost $C(B_m^*, B_n^*)$. The algorithm takes as input the function B_n^*, the number of available (σ_i, ρ_i) pairs, the cost function $C(B_m^*, B_n^*)$, and a sensitivity parameter $\epsilon > 0$. The approach of the algorithm is to select initial values for all pairs (σ_i, ρ_i) and then iteratively modify these values to reduce the cost $C(B_m^*, B_n^*)$.

The algorithm is presented in Figure 6. The initialization of the pairs (σ_i, ρ_i) is shown in steps 2 through 5 of Figure 6. Observe that the initial values are a subset of the pairs $\{(\overline{\sigma}_j, \overline{\rho}_j) \mid j = 1, \ldots, n\}$ that determine B_n^*.

The heuristic improves the initial selection by altering the (σ_i, ρ_i) pairs using the iteration shown in steps 6 through 11 of the figure. In each iteration step, the (σ_i, ρ_i) pairs are modified to reduce the cost function C. The iteration terminates when the cost cannot be significantly reduced. The crucial step of the algorithm is step 9, where a single pair (σ_l, ρ_l) is modified to minimize the cost function. During this step, the values of all pairs $\{(\sigma_i, \rho_i) \mid i \neq l\}$ are kept constant, and the pair (σ_l, ρ_l) is selected subject to the constraint that $\sigma_{l-1} < \sigma_l < \sigma_{l+1}$ (with boundary conditions

Scheme	Parameters	Traffic Constraint Function A^*
Peak-rate	ρ_{peak}	$A^*_{peak}(t) = \rho_{peak} t$
Dual bucket	$\rho_{peak}, (\sigma_{avg}, \rho_{avg})$	$B^*_{db}(t) = \min\{\rho_{peak}\, t\ , \sigma_{avg} + \rho_{avg}\ t\}$
Fixed burst	$\rho_{peak}, (\sigma_{fixed}, \rho_{fixed})$	$B^*_{fixed}(t) = \min\{\rho_{peak}\, t\ , \sigma_{fixed} + \rho_{fixed}\ t\}$
Concave hull	$\{(\hat{\sigma}_j, \hat{\rho}_j) \mid j = 1, \ldots, m\}$	$B^*_{hull}(t) = \min_{1 \le j \le m}\{\hat{\sigma}_j + \hat{\rho}_j\ t\}$
Product	$\rho_{peak}, (\sigma_{product}, \rho_{product})$	$B^*_{product}(t) = \min\{\rho_{peak}\, t\ , \sigma_{product} + \rho_{product}\ t\}$
Distance	$\rho_{peak}, (\sigma_{distance}, \rho_{distance})$	$B^*_{distance}(t) = \min\{\rho_{peak}\, t\ , \sigma_{distance} + \rho_{distance}\ t\}$

Table 1: Traffic parameterization schemes with their parameters and traffic constraint functions.

for this selection given by $\sigma_1 \geq 0$ and $\sigma_m \leq \overline{\sigma}_n$). Note that the choice of ρ_l is dependent on σ_l according to the relationship described in equation (5).

REMARKS: In the empirical evaluation presented in Section 4.3, we select the (σ_i, ρ_i) pair of minimum cost in step 9 through an exhaustive search through all possible values of σ_i. However, with ρ_i expressed in terms of σ_i, it is possible to write $C(B^*_m, B^*_n)$ with σ_i as the only independent variable, and the selection can be determined analytically by setting $\frac{\partial C}{\partial \sigma_i} = 0$. Also, while we do not make guarantees on the running time of the algorithm, the examples that we ran converged rapidly. In all examples using a sensitivity parameter $\epsilon = 0$, no more than six iterations were required.

4.3 Empirical Evaluation

We are now ready to evaluate our fast traffic characterization method for VBR video sources by comparing it with other traffic characterization schemes from the literature. With the results from Sections 3 and 4, our characterization method computes a function B^*_m based on the function $\mathcal{HR}^*_{200} \equiv B^*_n$ which in turn is obtained from the first 200 frames of the empirical envelope E^*. We evaluate the characterization method using the MPEG video traces *Park* and *News* described in Section 3.3 and a single FCFS multiplexer at a switch that operates at 155 Mbps.

We compare the traffic characterizations obtained with our method to other schemes that have been considered in the literature. These benchmarks are shown in Figure 1, and their parameters are described in the following:

(a) *Peak-rate*: A peak-rate characterization is determined by a single rate parameter ρ_{peak} which is assumed to be the ratio of the size of the largest video frame f_j and the inter-frame time r, i.e., $\rho_{peak} = \dfrac{\max_{0<j \leq N} f_j}{r}$.

(b) *Dual bucket*: In addition to ρ_{peak} described above, the dual bucket scheme employs a pair $(\sigma_{avg}, \rho_{avg})$ where ρ_{avg} is the average traffic rate over the length of the video sequence, i.e., $\rho_{avg} = \frac{\sum_{j=1}^{N} f_j}{Nr}$. The value of σ_{avg} is dependent on ρ_{avg} according to the relationship in equation (5).

(c) *Fixed burst*: The scheme outlined in [24] uses a single pair $(\sigma_{fixed}, \rho_{fixed})$ with the burst parameter σ_{fixed} set equal to a "reasonable" buffer size suggested to be either 1000 or 2000 cells, where the parameter ρ_{fixed} is obtained from σ_{fixed} using equation (5). We set $\sigma_{fixed} =$ 1000 cells since this choice yields better empirical performance. We also add a cell-spacer to enforce the peak rate ρ_{peak} of the connection.

(d) *Concave hull*: The concave hull approach in [33] selects m $(\hat{\sigma}, \hat{\rho})$ pairs for traffic characterization that are taken directly from the concave hull of the empirical envelope $\mathcal{H}E^*$. Consider the n pairs $\{(\hat{\sigma}_j, \hat{\rho}_j) \mid j = 1, \ldots, n\}$ of $\mathcal{H}E^*$, where $\hat{\sigma}_i < \hat{\sigma}_j$ for $i < j$. The parameters selected by the concave hull approach are the m pairs from $\mathcal{H}E^*$ that have the smallest bursts, that is, the pairs $\{(\hat{\sigma}_j, \hat{\rho}_j) \mid j = 1, \ldots, m\}$.

(e) *Product*: In [14] a scheme is proposed that uses the peak rate ρ_{peak} and a pair $(\sigma_{product}, \rho_{product})$, where $\sigma_{product}$ and $\rho_{product}$ are chosen from the candidate set of leaky buckets determined by equation (5) such that the product $\sigma_{product} \cdot \rho_{product}$ is minimized.

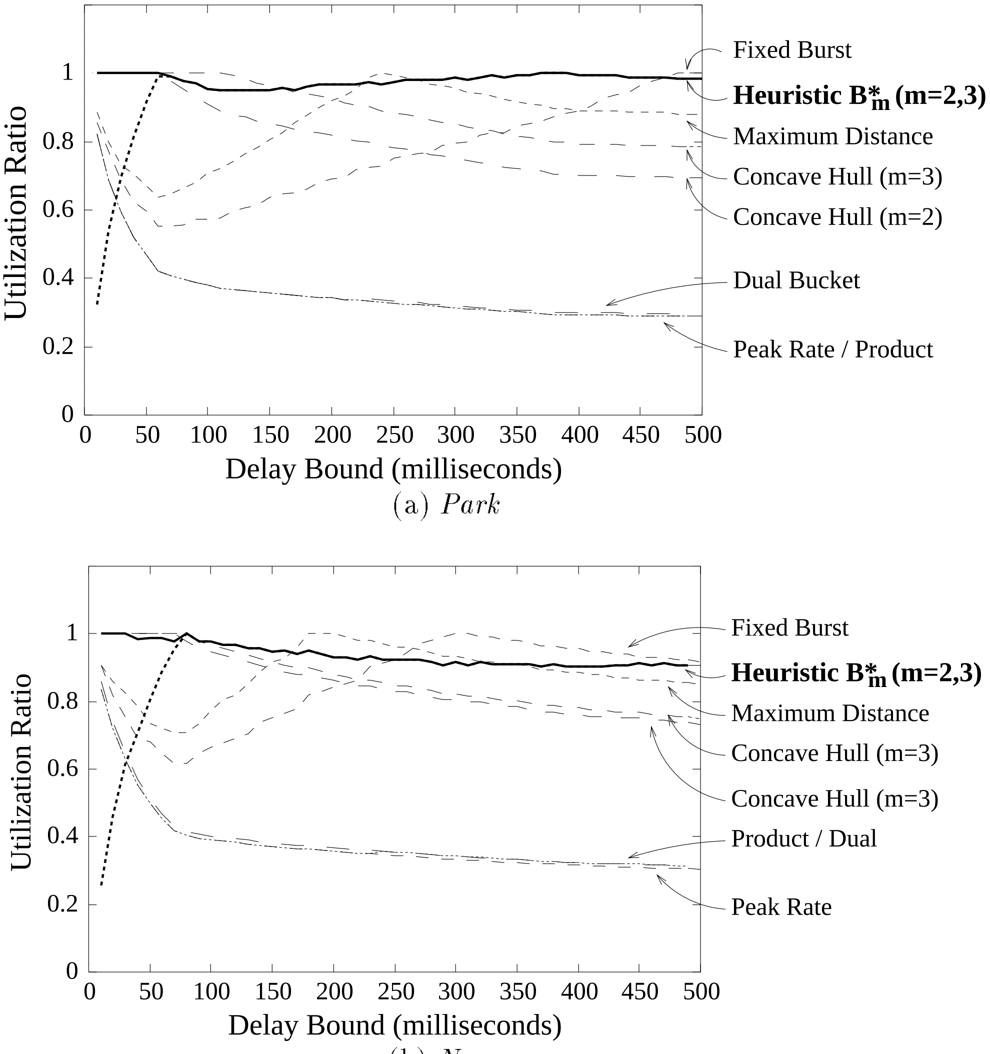
(f) *Distance*: This scheme from [14] uses the peak rate ρ_{peak} and a pair $(\sigma_{distance}, \rho_{distance})$ where $\rho_{distance}$ is selected such that $\delta = \sup_t \{ \frac{\sigma_{distance} + \rho_{distance} t}{t} - \frac{M_\epsilon^*(t)}{t} \}$ is minimized, $M_\epsilon(t) = \inf\{n,\ Pr\{N_t \geq n\} \leq \epsilon\}$ as discussed in Section 2.2, and $\epsilon = 0$ since we seek a worst-case bound.

We evaluate the accuracy of an arbitrary traffic constraint function A^* as follows. We assume that all traffic has the same traffic characterization A^* and identical delay bounds, and we compute the maximum number of admissible connections for all delay bounds as before. Since we wish to evaluate the ability of a particular traffic constraint function to approximate the empirical envelope, we plot the ratio of the number of admissible connections using A^* to the number obtained using the empirical envelope E^*, all as a function of the delay bound. In particular, for a given function A^* we plot:

$$\text{Utilization Ratio}(A^*, d) = \frac{\#\ \text{admissible connections with } A^* \text{ at delay bound } d}{\#\ \text{admissible connections with } E^* \text{ at delay bound } d} \tag{15}$$

Since all characterizations A^* considered will necessarily admit fewer connections than the empirical envelope, the metric allows us to determine how closely a particular characterization approximates the empirical envelope. For example, a traffic characterization A^* that admits the same number of connections as the empirical envelope would result in a constant curve Utilization Ratio$(A^*, d) = 1$.

Figures 7(a) and 7(b) show the utilization ratios of *Park* and *News* connections, respectively, for the entire suite of traffic characterizations described previously, namely, B_m^*, A_{peak}^*, B_{db}^*, B_{fixed}^*,



(a) *Park*

(b) *News*

Figure 7: Evaluation of characterization schemes.

B^*_{hull}, $B^*_{product}$, and $B^*_{distance}$. We depict characterizations B^*_m and B^*_{hull} for both $m = 2$ and $m = 3$ (σ, ρ) pairs.

The results for our heuristic characterization B^*_m are shown in Figure 7 as thick dashed and solid lines for $m = 2$ and and $m = 3$ (σ, ρ) pairs, respectively. For two (σ, ρ) pairs, note that our heuristic achieves a poor utilization for small delay bounds, while it is superior to other characterizations for most delay bounds greater than 50ms. This poor utilization at smaller delay bounds is due to the fact that our heuristic does not select the (σ, ρ) pair with $\rho = \rho_{peak}$. With three (σ, ρ) pairs, our heuristic B^*_3 achieves a utilization ratio of over 95% and 91% for all delay bounds in the *Park* and *News* sequences, respectively. The characterization B^*_3 produced by our heuristic method that employs three pairs is clearly the best characterization under consideration.

Notice the poor performance of the three characterizations A^*_{peak}, B^*_{db}, and $B^*_{product}$ in both graphs. While a peak-rate characterization yields relatively high utilizations for small delay bounds, the function A^*_{peak} achieves a utilization ratio of less than 40% for delay bounds greater than 60 ms for these video sequences. The additional leaky bucket employed in B^*_{db} and $B^*_{product}$ does not yield significant utilization gains. These three traffic characterizations are notably inferior to the other schemes.

5 Case Study: VBR Service with Deterministic Renegotiation

In this section we present a case study that applies our fast traffic characterization method to networks that renegotiate traffic parameters. In a network that employs renegotiation, traffic characterizations are occasionally modified to exploit long-term traffic variations of the VBR video traffic source, possibly leading to increased network utilization [4, 13, 34]. Since a renegotiation scheme requires multiple traffic characterizations for a single connection, a fast traffic characterization scheme such as the one described in this paper can be used to renegotiate traffic parameters. Here, we first discuss existing renegotiation strategies and point out modifications necessary to use renegotiation with a deterministic service. We next show how to apply our traffic characterization scheme to networks that employ renegotiation in a deterministic setting.

5.1 Renegotiation of Traffic Characterizations

Dynamic resource allocation schemes are motivated by studies showing that correlations of VBR video traffic occur over long time scales due to the extended duration of scenes [11, 19, 20]. By renegotiating the traffic characterization, for example, after each scene change, one can more accurately specify the traffic on a connection, resulting in a tighter characterization and hence higher network utilization.

Most renegotiation schemes that have been proposed attempt to renegotiate the traffic characterization of a connection whenever its long-term rate changes significantly [4, 13, 34]. Chong et. al. address the problem of predicting the rate changes of a live video source [4]. They consider both a recursive least-square method and an artificial neural network approach for the prediction.

In [13], Grossglauer et. al. propose a *Renegotiated Constant Bit Rate* (RCBR) scheme for both stored and live video which adds renegotiation and buffer monitoring to a static CBR service. They present algorithms for partitioning a video sequence into segments based on cost functions for both bandwidth allocation and number of renegotiations. Zhang and Knightly study a renegotiated VBR service for both stored and live video in [34]. Their algorithm for stored video proceeds by identifying the worst-case segment of the video sequence, characterizing this worst-case segment, and then iteratively repeating the procedure on the remaining video sequence after this worst-case segment is removed.

Although the above renegotiation schemes were shown to increase network utilization significantly, they cannot be used in a deterministic service. Since these schemes partition a video sequence into a number of segments and calculate a traffic characterization independently for each segment, it is possible that a situation occurs where several connections need to increase their resource allocation even if sufficient resources are not available. In such a scenario, the renegotiation requests cannot be accommodated, and either the video quality or the QoS must be compromised, resulting in a violation of the worst-case guarantees in a deterministic service. In the remainder of this section, we present a renegotiation scheme that does not incur the risk of compromising QoS guarantees, Note that this is the first renegotiation scheme proposed so far that is applicable to connections with a deterministic QoS. We use the discussion to demonstrate the effectiveness of our characterization method in such a renegotiation scheme.

5.2 Deterministic Renegotiation

A key requirement for a renegotiation scheme in a deterministic service is to ensure that the traffic characterization for a connection does not increase, i.e., connections only release resources and do not request additional resources. If the traffic characterizations do not increase, then all renegotiation requests can be satisfied and deterministic QoS guarantees are maintained.

Let the traffic on a video connection be given by A. We assume that the traffic characterization is negotiated at $u+1$ distinct times $\tau_0, \tau_1, \ldots, \tau_u$, where $\tau_i < \tau_j$ if $i < j$ and that the traffic characterization negotiated at time τ_i is given by $A^*_{\tau_i}$. Now, any traffic characterization $A^*_{\tau_i}$ must provide a bound on the worst-case traffic for the remainder of the video sequence, that is, for all i:

$$A^*_{\tau_i}(t) \geq A[\tau_i + \tau, \tau_i + \tau + t] \qquad \forall \tau, t \geq 0 \tag{16}$$

Further, to ensure that all renegotiation requests are satisfied, a newly-computed traffic characterization may not request additional resources, that is, we enforce that for all $\tau_i < \tau_j$:

$$A^*_{\tau_i}(t) \geq A^*_{\tau_j}(t) \qquad \text{for all } t \geq 0 \tag{17}$$

The condition in equation (16) ensures that any function $A^*_{\tau_i}$ is a viable traffic constraint function, while equation (17) guarantees that the renegotiation requests can be satisfied. To show that a set of traffic characterizations $\{A^*_{\tau_i}\}$ can be used in a renegotiation scheme with a deterministic service, it is sufficient to show that equations (16) and (17) are satisfied.

$$\{E^*_{\tau_i}(jr) \mid j = 1, \ldots, k\} \xrightarrow{\substack{\text{Repetition}\\ \text{Extension}}} R^*_{\tau_i,k} \xrightarrow{\substack{\text{Concave}\\ \text{Hull}}} \mathcal{H}R^*_{\tau_i,k} = B^*_{\tau_i,n} \xrightarrow{\substack{\text{Heuristic}\\ \text{Algorithm}}} B^*_{\tau_i,m}$$

Figure 8: Overview of traffic characterization method.

We construct a class of traffic constraint functions $\{E^*_{\tau_i}\}$ that satisfies both equations (16) and (17) by defining the function $E^*_{\tau_i}$ to be the empirical envelope of the sequence A for all times $t \geq \tau_i$, that is:

$$E^*_{\tau_i}(t) = \sup_{\tau \geq 0} A[\tau_i + \tau, \tau_i + \tau + t] \qquad \forall t \geq \tau_i \tag{18}$$

$E^*_{\tau_i}$ is the tightest characterization for the video sequence A for $t \geq \tau_i$. $E^*_{\tau_i}$ satisfies equation (16) by definition. To show that equation (17) is also satisfied, we note that for two traffic constraint functions $E^*_{\tau_i}$ and $E^*_{\tau_j}$ of the same video sequence A with $\tau_i < \tau_j$, the following holds:

$$E^*_{\tau_j}(t) = \sup_{\tau \geq 0} A[\tau_j + \tau, \tau_j + \tau + t] \leq \sup_{\tau \geq 0} A[\tau_i + \tau, \tau_i + \tau + t] \leq E^*_{\tau_i}(t), \tag{19}$$

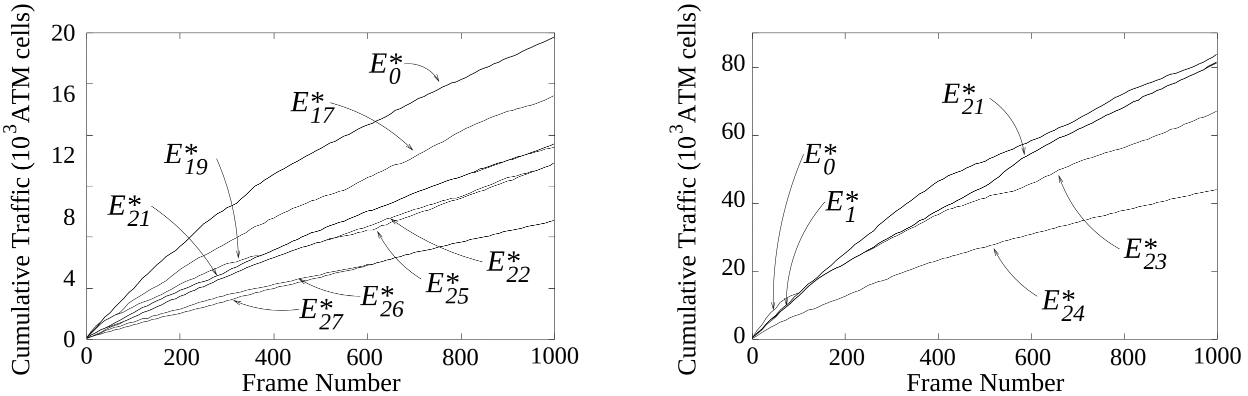
We have shown that $\{E^*_{\tau_i}\}$ are a class of valid traffic constraint functions that can be used in a deterministic service with renegotiation. If a renegotiation occurs τ_i time units into a video sequence, the resource allocation can be calculated according to $E^*_{\tau_i}$. However, the functions $E^*_{\tau_i}$ are similar to the empirical envelope E^* in that they employ a large number of parameters that are expensive to compute. In the next section we show how to apply our fast traffic characterization method to approximate these functions $\{E^*_{\tau_i}\}$.

5.3 Application of the Fast Video Characterization Method

Recall that the characterization method presented in Sections 3 and 4 proceeds in two steps, and its application to $\{E^*_{\tau_i}\}$ is shown in Figure 8. In the first step we calculate R^*_{k,τ_i}, the repetition extrapolation of the first k parameters of $E^*_{\tau_i}$, where R^*_{k,τ_i} has the same form as R^*_k given in equation (10). However, since R^*_{k,τ_i} is not a viable traffic constraint function, we calculate a $(\vec{\sigma}, \vec{\rho})$-model traffic characterization $\mathcal{H}R^*_{k,\tau_i}$ by computing the concave hull of R^*_{k,τ_i}. In the second step, to reduce the number of (σ, ρ) pairs required, we apply the heuristic algorithm from Section 4 to $\mathcal{H}R^*_{k,\tau_i} \equiv B^*_{\tau_i,n}$, yielding $B^*_{\tau_i,m}$.

We first consider the class of functions $\{\mathcal{H}R^*_{k,\tau_i}\}$. To show that $\{\mathcal{H}R^*_{k,\tau_i}\}$ can be used in renegotiation, we require that both equations (16) and (17) are satisfied. Equation (16) is satisfied by construction. To show that equation (17) is satisfied, we consider two functions R^*_{k,τ_i} and R^*_{k,τ_j}, where $\tau_i < \tau_j$. From equation (19), we obtain directly that $R^*_{k,\tau_i}(t) \geq R^*_{k,\tau_j}(t)$ for all t. We can then conclude that $\mathcal{H}R^*_{k,\tau_i}(t) \geq \mathcal{H}R^*_{k,\tau_j}(t)$ for all t, and thus $\{\mathcal{H}R^*_{k,\tau_i}\}$ satisfies equation (17).

Although the class of functions $\{\mathcal{H}R^*_{k,\tau_i}\}$ are appropriate for use in deterministic renegotiation, we cannot make the same claim about the functions $\{B^*_{\tau_i,m}\}$. Since the heuristic algorithm as presented determines a function $B^*_{\tau_i,m}$ based only on $\mathcal{H}R^*_{k,\tau_i}$, independent of the previous approximation $B^*_{\tau_{i-1},m}$, it is possible for the algorithm to select an approximation $B^*_{\tau_i,m}$ that is larger



(a) *Park* (b) *News*

Figure 9: Traffic constraint functions $E^*_{\tau_i}$.

than $B^*_{\tau_{i-1},m}$ for some values of t. Thus, the condition in equation (17) does not necessarily hold. To apply our characterization method to deterministic renegotiation requires a modification of the heuristic where either (1) renegotiation attempts are suppressed at times τ_i whenever $B^*_{\tau_{i-1},m}(t) < B^*_{\tau_i,m}(t)$ for some t, or (2) the search space of the heuristic is modified so that the only pairs (σ, ρ) considered are those that will yield $B^*_{\tau_{i-1},m}(t) \geq B^*_{\tau_i,m}(t)$ for all t.

5.4 Empirical Examples

We present examples based on MPEG video sequences to demonstrate the impact of renegotiation on network utilization. For the evaluation, we again use the *Park* and *News* traces described earlier in the paper.

In the first example we show how the traffic characterization changes as it is renegotiated throughout transmission of the sequence. We consider the class of characterizations $\{E^*_{\tau_i}\}$, where $\tau_i = i$ minutes for $i = 0, \ldots T-1$ and T is the length of the movie in minutes. Since the *Park* and *News* sequences are 28 and 25 minutes long, respectively, we consider 28 different traffic characterizations for *Park* and 25 for *News*. We plot these traffic characterizations in Figure 9, where we write $E^*_{\tau_i} = E^*_i$ since $\tau_i = i$. In the figure, we only depict the traffic characterizations E^*_i that are visibly smaller than all traffic characterizations E^*_j with $j < i$. For example, we see in Figure 9(a) that all traffic characterizations E^*_j with $0 < j < 17$ appear identical to E^*_0 when plotted.

In the next experiment, we illustrate the average network utilization gain with a deterministic renegotiation scheme using our traffic characterization method. Similar to the experiments in previous sections, we assume that a number of video connections are transmitted on a single 155 Mbps FCFS multiplexer, and we assume that all traffic has the same delay bound d and is of a single traffic type, namely either *Park* or *News*. To evaluate the average performance gain, we assume that the connections are at different transmission points of the stream, resulting in different traffic characterizations due to renegotiation. In particular, we call the frame that is transmitted by a connection at time t the *current frame* at time t, and we assume that the current frames for all connections are uniformly distributed over the entire set of frames $1, \ldots, N$. We consider a scenario in which renegotiation occurs periodically at multiples of a *renegotiation period*. For example, if the renegotiation period is 100 frames, then a connection with current frame 213 uses the traffic characterization computed 200 frames into the sequence based on only frames $200, \ldots, N$.

The experiment is as follows. Starting with an empty multiplexer, we add connections to the multiplexer, selecting a random current frame for each connection that (together with the renegotiation period) determines its traffic characterization. We continue adding connections as long as the admission control tests are satisfied, that is, as long as all connections are guaranteed a worst-case delay bound of d. We record the maximum number of admissible connections. This process is repeated 1000 times for each delay bound, and we plot the average number of admissible connections as a function of the delay bound. We obtained similar results for 10 runs of the above experiment.

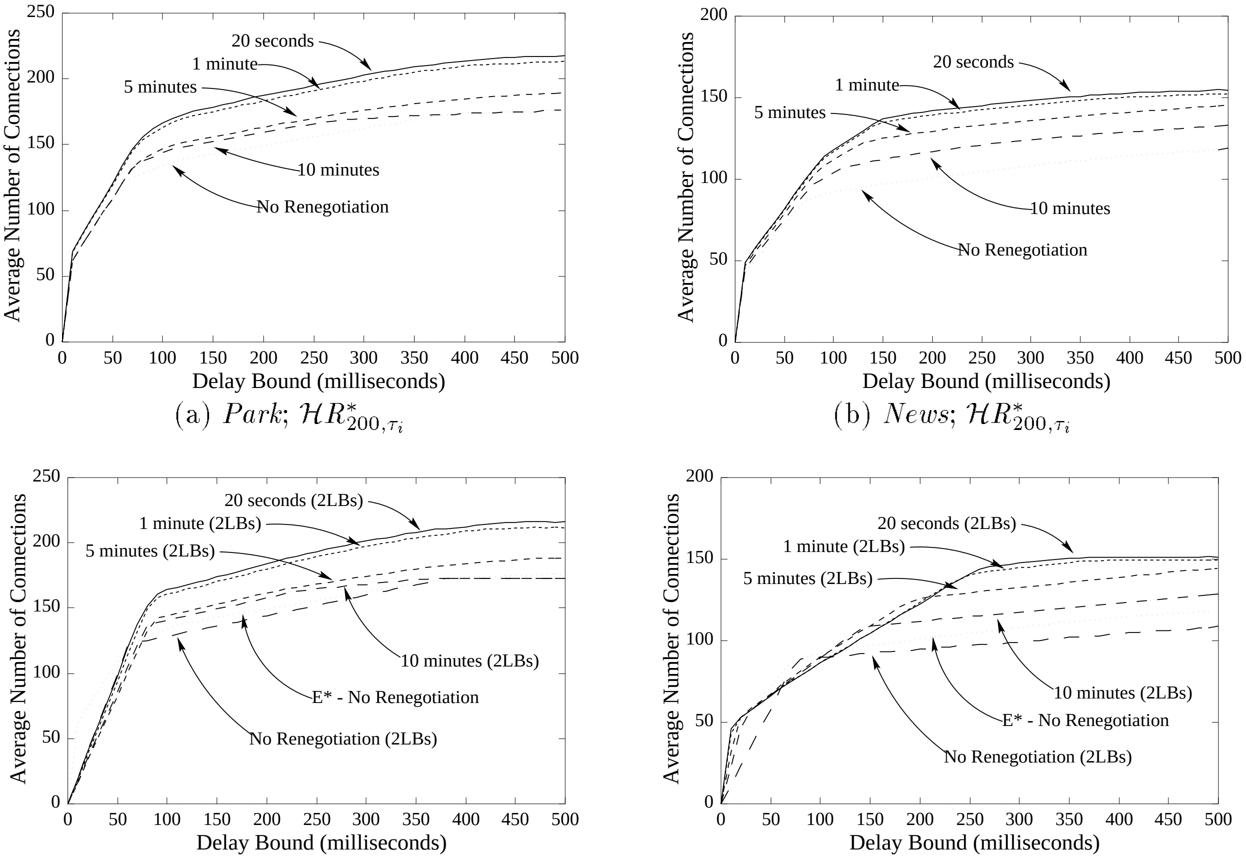
Figure 10 depicts the number of admissible connections for both the *Park* and *News* sequences for several renegotiation periods. We plot the maximum number of admissible connections as a function of delay bound. Figures 10(a) and (b) show results obtained using $\mathcal{HR}^*_{200,\tau_i}$ for the traffic characterization, while Figures 10(c) and (d) use the functions B^*_{m,τ_i} with two (σ, ρ) pairs. In all graphs, the dotted curves show the utilization obtained when the characterization $\mathcal{HR}^*_{200}$ is employed without any renegotiation. We plot curves corresponding to renegotiation periods of 20 seconds as well as 1, 5, and 10 minutes. For Figures 10(c) and (d), we also show the utilization obtained using B^*_2 without negotiation.

We see in Figure 10 that the renegotiation period significantly impacts the number of admissible connections. In Figure 10(a), note that the number of admissible *Park* connections increases by 20-30% for delay bounds larger than 50 ms if the renegotiation period is less than 1 minute. For the longer renegotiation periods, i.e., 5 minutes and 10 minutes, renegotiation provides gains of about 10%. For the *News* sequence, we see in Figure 10(b) that even infrequent renegotiation results in considerable utilization gains.

The plots in Figure 10(c) and (d) demonstrate the effectiveness of the heuristic in approximating the functions $\mathcal{HR}^*_{200,\tau_i}$ with B^*_{2,τ_i}. However, since the class of functions B^*_{m,τ_i} are not appropriate for renegotiation without modification, a smaller renegotiation period does not necessarily lead to an increase in network utilization.

6 Concluding Remarks

The traffic characterization used for VBR video connections has a significant impact on the number of admissible connections in a network with a deterministic service. We presented a method for traffic characterization based on the empirical envelope of a video sequence that uses a two-step process. We first approximate the empirical envelope with a characterization that can be policed by some number of leaky buckets, and we then determine the final characterization that can be policed by a small, fixed number of leaky buckets. Using two MPEG-compressed video sequences, we



(a) *Park*; $\mathcal{HR}^*_{200,\tau_i}$

(b) *News*; $\mathcal{HR}^*_{200,\tau_i}$

(c) *Park*; B^*_{2,τ_i}.

(d) *News*; B^*_{2,τ_i}.

Figure 10: Utilization comparison for different renegotiation periods.

demonstrated that our characterization method determines accurate characterizations that admit a large number of connections. With the caveat that our experimental evaluation is based on only a small number of video traces, the experiments in this paper gave the following insights:

- For the MPEG video sequences considered in the paper, we saw that as few as 200 parameters of the empirical envelope out of a total of 40,000 are sufficient to yield a characterization that admits the same number of connections as the empirical envelope. This observation suggests that the relevant information of an MPEG sequence is contained in a small segment of the envelope.

- Using our heuristic algorithm, three leaky buckets were shown to be sufficient to admit nearly the same number of connections as the empirical envelope. Based on the good performance of our heuristic algorithm, it may not be worthwhile to investigate more complex algorithms for video characterization.

- The dual leaky bucket scheme was shown to yield poor network utilization in all experiments from Section 4.3. However, the poor performance is not due to the fact that only two leaky buckets are used, but rather to a poor selection of leaky bucket parameters. Using a better heuristic algorithm such as the one developed in this paper, one can achieve markedly higher performance.

- In [33], the numerical examples indicated that a large number of leaky bucket pairs were needed to approximate the empirical envelope. Using our method, the number of leaky buckets needed to achieve performance similar to the envelope is small. This discrepancy can be explained as follows. First, the heuristic algorithm presented here is superior to the concave hull approach from [33]. Second, we note that the examples in this paper only consider delay bounds up to 500ms, while the examples in [33] consider delay bounds up to 2000ms. For a larger delay bound range, additional leaky buckets are necessary to closely approximate the empirical envelope.

- The deterministic resource renegotiation scheme that we describe is distinguished from previous approaches in that it is appropriate for services that provide constant video quality and deterministic QoS guarantees. The experimental data suggests that the expected utilization gain from this deterministic renegotiation is 20-35%.

References

[1] ATM Forum, ATM Forum Traffic Management Specification Version 4.0, October, 1995.

[2] C.-S. Chang. Stability, Queue Length, and Delay of Deterministic and Stochastic Queueing Networks. *IEEE Transactions on Automatic Control*, 39(5):913–931, May 1994.

[3] S. Chong and S. Li. (σ,ρ)-Characterization Based Connection Control for Guaranteed Services in High Speed Networks. In *Proc. IEEE Infocom '95*, pages 835–844, 1995.

[4] S. Chong, S. Li, and J. Ghosh. Predictive Dynamic Bandwidth Allocation for Efficient Transport of Real-Time VBR Video over ATM. *IEEE Journal on Selected Areas in Communication*, 13(1):12–23, January 1995.

[5] T. H. Cormen, C. E. Leiserson, and R. L. Rivest. *Introduction to Algorithms.* MIT Press, 1992.

[6] R. L. Cruz. A Calculus for Network Delay, Part I: Network Elements in Isolation. *IEEE Transactions on Information Theory*, 37(1):114–131, January 1991.

[7] R. L. Cruz. A Calculus for Network Delay, Part II: Network Analysis. *IEEE Transactions on Information Theory*, 37(1):132–141, January 1991.

[8] A. Elwalid, D. Mitra, and R. Wentworth. A New Approach for Allocating Buffers and Bandwidth to Heterogeneous, Regulated Traffic in an ATM Node. *IEEE Journal on Selected Areas in Communications*, 13(6):1115–1127, August 1995.

[9] V. Frost and B. Melamed. Traffic Modelling for Telecommunications Networks. *IEEE Communications Magazine*, 32(3):70–81, March 1994.

[10] D. Le Gall. MPEG: A Video Compression Standard for Multimedia Applications. *Communications of the ACM*, 34(4):46–58, April 1991.

[11] M. W. Garrett and W. Willinger. Analysis, Modeling and Generation of Self-Similar VBR Video Traffic. In *Proc. ACM Sigcomm '94*, pages 269–280, August 1994.

[12] L. Georgiadis, R. Guerin, V. Peris, and K. N. Sivarajan. Efficient Network QoS Provisioning Based on per Node Traffic Shaping. In *Proc. IEEE Infocom '96*, pages 102–110, March 1996.

[13] M. Grossglauer, S. Keshav, and D. Tse. RCBR: A Simple and Efficient Service for Multiple Time-Scale Traffic. In *Proc. ACM Sigcomm '95*, pages 219–230, August 1995.

[14] F. Guillemin, C. Rosenberg, and J. Mignault. On Characterizing an ATM Source via the Sustainable Cell Rate Traffic Descriptor. In *Proc. IEEE Infocom '95*, pages 1129–1136, April 1995.

[15] E. Knightly and H. Zhang. Traffic characterization and switch utilization using deterministic bounding interval dependent traffic models. In *Proc. IEEE Infocom '95*, pages 1137–1145, April 1995.

[16] E. W. Knightly. H-BIND: A New Approach to Providing Statistical Performance Guarantees to VBR Traffic. In *Proc. IEEE Infocom '96*, pages 1091–1099, March 1996.

[17] E. W. Knightly and P. Rossaro. Effects of Smoothing on End-to-End Performance Guarantees for VBR Video. In *Proc. International Symposium on Multimedia Communications and Video Coding*, October 1995.

[18] S. S. Lam, S. Chow, and D. K. Y. Yau. An Algorithm for Lossless Smoothing of MPEG Video. In *Proc. ACM Sigcomm '94*, pages 281–293, August 1994.

[19] A. Lazar, G. Pacifici, and D. Pendarakis. Modeling Video Sources for Real-Time Scheduling. In *Proc. IEEE Globecom '93*, pages 835–839, December 1993.

[20] S. Q. Li, S. Chong, C. Hwang, and X. Zhao. Link Capacity Allocation and Network Control by Filtered Input Rate in High Speed Networks. In *Proc. IEEE Globecom '93*, pages 744–750, December 1993.

[21] J. Liebeherr, D. E. Wrege, and Domenico Ferrari. Exact Admission Control in Networks with Bounded Delay Services. To appear: *IEEE/ACM Transactions on Networking.*

[22] S. Low and P. Varaiya. A Simple Theory of Traffic and Resource Allocation in ATM. In *Proc. IEEE Globecom '91*, pages 1633–1637, 1991.

[23] J. M. McManus and K. W. Ross. Video on Demand over ATM: Constant-Rate Transmission and Transport. To appear: *IEEE Journal on Selected Areas in Communications.*

[24] P. Pancha and M. El Zarki. Leaky Bucket Access Control for VBR MPEG Video. In *Proc. IEEE Infocom*, pages 796–803, April 1995.

[25] A. K. Parekh and R. G. Gallager. A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks: The Single-Node Case. *IEEE/ACM Transactions on Networking*, 1(3):344–357, June 1993.

[26] E. P. Rathgeb. Modeling and Performance Comparison of Policing Mechanisms for ATM Networks. *IEEE Journal on Selected Areas in Communications*, 9(4):325–334, April 1991.

[27] E. P. Rathgeb. Policing of Realistic VBR Video Traffic in an ATM Network. *International Journal of Digital and Analog Communications Systems*, 6:213–226, 1993.

[28] A. R. Reibman and A. W. Berger. Traffic Descriptors for VBR Video Teleconferencing Over ATM Networks. *IEEE/ACM Transactions on Networking*, 3(3):329–339, June 1995.

[29] O. Rose. Statistical properties of MPEG video traffic and their impact on traffic modeling in ATM systems. Technical Report 101, Institute of Computer Science, University of Wurzburg, February 1995. The traces used in this paper are available via anonymous ftp from the site ftp-info3.informatik.uni-wuerzburg.de in the directory /pub/MPEG/.

[30] C. Rosenberg and B. Lague. A Heuristic Framework for Source Policing in ATM Networks. *IEEE/ACM Transactions on Networking*, 2(4):387–397, August 1994.

[31] J. Salehi, Z. Zhang, J. Kurose, and D. Towsley. Supporting Stored Video: Reducing Rate Variability and End-to-End Resource Requirements through Optimal Smoothing. To appear: *Proc. ACM Sigmetrics '96*, May 1996.

[32] J. S. Turner. New Directions in Communications (or Which Way to the Information Age?). *IEEE Communications Magazine*, 25(8):8–15, October 1986.

[33] D. E. Wrege, E. W. Knightly, H. Zhang, and J. Liebeherr. Deterministic Delay Bounds for VBR Video in Packet-Switching Networks: Fundamental Limits and Practical Tradeoffs. To appear: *IEEE/ACM Transactions on Networking*, June 1996.

[34] H. Zhang and E. Knightly. A New Approach to Support Delay-Sensitive VBR Video in Packet-Switched Networks. In *Proc. 5th Intl. Workshop on Network Operating System Support for Digital Audio and Video (NOSSDAV)*, pages 275–286, April 1995.