February 2003

On Evaluating Loss Performance Deviation: A Simple Tool and Its Practical Applications

Ying Xu
University of Pennsylvania

Roch A. Guérin
University of Pennsylvania, guerin@acm.org

Postprint version. Published in *Lecture Notes in Computer Science*, Volume 2601, Quality of Service in Multiservice IP Networks: Proceedings of the Second International Workshop 2003, (QoS-IP 2003), pages 1-18.
Publisher URL: http://www.springerlink.com/link.asp?id=8eqykx5h3anbxkcv

On Evaluating Loss Performance Deviation: A Simple Tool and Its Practical Applications

Abstract

The focus of this paper is on developing and evaluating a practical methodology for determining if and when different types of traffic can be safely multiplexed within the same service class. The use of class rather than individual service guarantees offers many advantages in terms of scalability, but raises the concern that not all users within a class see the same performance. Understanding when and why a user will experience performance that differs significantly from that of other users in its class is, therefore, of importance. Our approach relies on an analytical model developed under a number of simplifying assumptions, which we test using several real traffic traces corresponding to different types of users. This testing is carried out primarily by means of simulation, to allow a comprehensive coverage of different configurations. Our findings establish that although the simplistic model does not accurately predict the absolute performance that individual users experience, it is quite successful and robust when it comes to identifying situations that can give rise to substantial performance deviations within a service class. As a result, it provides a simple and practical tool for rapidly characterizing real traffic profiles that can be safely multiplexed.

Comments

Postprint version. Published in *Lecture Notes in Computer Science*, Volume 2601, Quality of Service in Multiservice IP Networks: Proceedings of the Second International Workshop 2003, (QoS-IP 2003), pages 1-18.

Publisher URL: http://www.springerlink.com/link.asp?id=8eqykx5h3anbxkcv

On Evaluating Loss Performance Deviation: A Simple Tool and Its Practical Implications

Ying Xu and Roch Guérin
Multimedia and Networking Lab
Department of Electrical Engineering
University of Pennsylvania
E-mail: (yingx,guerin)@ee.upenn.edu

Abstract

The focus of this paper is on developing and evaluating a practical methodology for determining if and when different types of traffic can be safely multiplexed within the same service class. The use of class rather than individual service guarantees offers many advantages in terms of scalability, but raises the concern that not all users within a class see the same performance. Understanding when and why a user will experience performance that differs significantly from that of other users in its class is, therefore, of importance. Our approach relies on an analytical model developed under a number of simplifying assumptions, which we test using several real traffic traces corresponding to different types of users. This testing is carried out primarily by means of simulation, to allow a comprehensive coverage of different configurations. Our findings establish that although the simplistic model does not accurately predict the absolute performance that individual users experience, it is quite successful and robust when it comes to identifying situations that can give rise to substantial performance deviations within a service class. As a result, it provides a simple and practical tool for rapidly characterizing real traffic profiles that can be safely multiplexed.

I. Introduction

To design a successful QoS service model, it is often necessary to trade-off *both* the performance guarantees that the service model can provide *and* the associated operational and management complexity so that the QoS guarantees needed by end users can be provided while the service model is kept simple enough to meet deployment constraints. Most recently, there has been a renewed interest in investigating aggregate QoS solutions, due to the fact that they can greatly improve scalability, and therefore reduce the associated operational and management complexity. In a typical aggregate QoS service model, individual flows are multiplexed into a common service class and treated as a single stream when it comes to QoS guarantees, which eliminates the need for per-flow based state and processing. Such an approach is reflected in recent proposals such as the IETF Diff-Serv model [4]. By coarsening the different levels of QoS that the network offers into a small number of service classes, the Diff-Serv model is expected to scale well. However, this gain in scalability through aggregation comes at the cost of losing awareness of the exact level of performance that an individual user experiences. More specifically, because performance is now monitored and enforced only at the aggregate class level, it is not clear whether performance guarantees extend to all individual users in the service class. In particular, it is possible that a given user experiences poor performance without this being noticed at the class level.

This work was supported in part through NSF grant ITR00-85930.

This issue was explored in an early work [13], which focused on the loss probability of an individual user as the relevant metric. [13] developed a number of *explicit* models to evaluate the individual loss probabilities and the overall loss probability, in which user traffic are represented by either Markov ON-OFF sources or periodic sources. Using these analytical models, the deviation between individual loss probabilities and the overall loss probability was evaluated for a broad range of configurations. The major conclusion was that there are indeed cases where significant performance deviations can exist, and that user traffic parameters can have a major influence on whether this is the case or not.

In this paper, we go beyond the work of [13] by developing a simple methodology capable of evaluating loss performance deviations in realistic settings. Specifically, we concentrate on the ability to predict performance deviations when aggregating *real* traffic sources such as voice and video, which may not be accurately captured by the simplified source models of [13]. We choose voice and video sources for our study, as they correspond to applications with a need for service guarantees, and therefore, most likely to benefit from QoS solutions. Our investigation starts with mapping real traffic sources onto "equivalent" Markovian sources. This mapping is, however, limited to matching first order statistics. The rationale for using such an approach is, as is discussed later in the paper, based on our expectation that performance deviations are less sensitive to source modeling inaccuracies than absolute performance measures. The efficacy of this simple approach is then evaluated by means of simulations for a broad range of scenarios where multiple real-time traffic sources, i.e., voice and/or video sources, are aggregated, which we believe can also help gain a better understanding of whether and when an aggregate service model is suitable for supporting real-time applications.

Our findings establish that while the simplistic source mapping we use indeed affects the accuracy of estimating absolute performance measures, it can accurately identify scenarios where significant performance deviations will occur. The reliability and robustness of the approach are further discussed in Section VI, where we use worst case analysis to explicitly evaluate the intrinsic limitations in predicting performance deviation using only first order statistics.

The rest of the paper is structured as follows: Section II describes our model and methodology for loss deviation prediction. Sections III and IV report experimental results for two different boundary configurations, while results for the intermediate configuration are reported in Section V. In Section VI, we conduct a worst case analysis of the performance of our solution. We concentrate on the main results, while most of the derivations are relegated to Appendix I. Finally, Section VII summarizes the findings and implications of the paper.

II. Model and Methodology

The system we consider consists of multiple users that belong to a common service class, each of which is connected through its access link to an aggregation point, which is a single server, finite buffer, FIFO queue. In this section, we review the different configurations that we consider and the performance measures that we use to evaluate loss performance deviation. For completeness, we also summarize the analytical models, developed in [13], that we use to evaluate the loss probability of an individual user. The voice and video traffic sources and their characteristics are then

described and the method to map them into the analytical model is also presented. At last, the simulation environment that we use is briefly discussed.

A. System Parameters and Loss Deviation Evaluation

1) System Parameters: The first parameter we vary is the number of users aggregated in the same service class, i.e., the number of traffic sources multiplexed in the FIFO queue. In particular, we focus on two cases: a two-source configuration and a "many-source" one, which correspond to two different boundary conditions, i.e., an environment where few large bandwidth connections share resources, and one where many "small" flows are multiplexed into the same queue. Another system parameter we consider is the size of the FIFO queue into which flows are multiplexed. In our simulations, we consider both a system with a relative small buffer size and another system with a relative large buffer size. In most cases, the qualitative behavior of the two systems is similar, and thus typically only results corresponding to one of them are reported.

B. Analytical model

As addressed before, we use Markov ON-OFF source [3] to model individual users aggregated in a service class. A Markov ON-OFF source can be represented by a 3-tuple $< R, b, \rho >$, where R is the transmission (peak) rate when the source is active, b is the average duration of an active or ON period, and ρ represents the fraction of time that the source is active, or its utilization. We chose the Markov ON-OFF source model not only because it is amenable to analysis, but also because it can capture the built-in ON-OFF patterns exhibited in the voice and video traffic. In the next sub-sections, two models, one for a system where only two sources are aggregated and the other for a system where many sources are aggregated, are described. We only present the main results, while all the relating derivations can be found in [13].

1) Analytical model when aggregating two sources: The individual loss probability in a two source model can be evaluated simply by specializing an N-source system to the case $N = 2$. In an N-source system, each source i is characterized by the 3-tuple described in the previous paragraph, denoted by (R_i, b_i, ρ_i). The input process to the buffer can then be described through a state vector: $\mathbf{S} = (s_1, s_2, \cdots, s_N)$, where s_i is 0 when source i is OFF and 1 when it is ON. For any state, the input rate γ_S to the system is given by $\gamma_{\mathbf{S}} \doteq \mathbf{S} \cdot \mathbf{R^T}$, where $\mathbf{R} = (R_1, R_2, \cdots, R_N)$ is the peak rate vector of the sources. Then, under the standard assumption that the system is ergodic and stationary, the loss probability experienced by source i in a finite buffer system of size x can then be approximated by:

$$P_L^i = \frac{\sum_{\substack{\mathbf{S}:(s_i=1,\\ \gamma_{\mathbf{S}}>C)}} (\pi_{\mathbf{S}} - F_{\mathbf{S}}(x))(R_i - C \cdot \frac{R_i}{\gamma_{\mathbf{S}}})}{\rho_i R_i} = \frac{r_L^i}{r_S^i}, \tag{1}$$

where $\pi_{\mathbf{S}}$ is the stationary probability that the input is in state $\mathbf{S}$ and r_L^i and r_S^i correspond to the long term loss rate and sending rate of source i. The quantity $F_{\mathbf{S}}(x)$ is the stationary probability that the queue length is smaller than

x and the system is in state $\mathbf{S}$. When aggregating *Markov ON-OFF sources*, $\mathbf{S}$ which can be readily obtained from existing body of work such as [8] or [12].

Similarly, the overall loss rate P_L can be expressed as:

$$P_L = \frac{\sum_{\mathbf{S}:\gamma_{\mathbf{S}}>C} (\pi_{\mathbf{S}} - F_{\mathbf{S}}(x))(\gamma_{\mathbf{S}} - C)}{\sum_{i}^{N} \rho_i R_i} = \frac{r_L}{r_S}, \tag{2}$$

where r_L and r_S correspond to the overall long term loss rate and sending rate, respectively.

2) Analytical model when aggregating many ON-OFF sources: When the system consists of many users, equations (1) and (2) are still applicable, but are computationally difficult to evaluate due to the well-known state explosion problem. Due to this reason, we rely on a bufferless model based on rate envelop multiplexing technique [9], [10] to conduct our study. In such a model, the total input traffic is divided into two parts: the background traffic and the traffic associated with a specific source (without loss of generality, we assume it is source N) we focus on. Denote λ_t, λ_t^b and λ_t^N as the random variables associated with the rate envelop at time t of the total traffic, the background traffic, and the rate of source N respectively, while m is the total input rate. Using these assumptions and notations, the overall loss probability P_L and the loss probability P_L^N of source N can be obtained from the following equations:

$$\begin{aligned} P_L^N &= \frac{E\left[(\lambda_t - C)^+ \cdot \frac{\lambda_t^N}{\lambda_t}\right]}{\rho_N * R_N} \\ &= E\left[\left(\frac{\lambda_t^b + R_N - C}{\lambda_t^b + R_N}\right)^+\right], \end{aligned} \tag{3}$$

and

$$\begin{aligned} P_L &= \frac{E[(\lambda_t - C)^+]}{m} \\ &= \frac{\rho_N E[(\lambda_t^b + R_N - C)^+] + (1-\rho_N)E[(\lambda_t^b - C)^+]}{m}, \end{aligned} \tag{4}$$

where both P_L^N and P_L can be numerically evaluated if the distribution of λ_t^b is explicitly specified.

3) Evaluation Methodology: Let P_L denote the overall loss probability, P_L^i the loss probability of user i, and P_{max}^N the maximum loss probability experienced by an individual user, say, user N, where N is the total number of users in the system. We first evaluate the *minimum* amount of bandwidth C needed to ensure an overall loss probability $P_L \leq P_{max}$, where P_{max} corresponds to the target loss probability of the service class. We then compute the *maximum* loss probability ratio $PLR = P_{max}^N / P_L$. In addition, we also evaluate the *minimum* percentage by which the bandwidth allocated to the service class needs to be increased to ensure that $P_{max}^N \leq P_{max}$. In other words, if C_N denotes the minimum amount of bandwidth needed so that $P_{max}^N \leq P_{max}$, we evaluate the quantity $(C_N - C)/C$.

The (maximum) loss probability ratio reflects the magnitude of the loss deviation when service guarantees are only provided at the aggregate level, and is therefore the metric we use to assess whether traffic aggregation is safe or

not. Conversely, the additional bandwidth needed measures the cost of eliminating deviations for a given traffic mix. In addition to these two metrics, we also introduce a "consistency test" that measures the ability of our model to successfully predict scenarios where severe performance deviations will arise. The test relies on a "threshold" region, $[PLR_{min}, PLR_{max}]$, that is used to separate risky traffic mixes, i.e., mixes that give rise to large deviations, from safe ones. Specifically, if the loss probability ratio is less than PLR_{min}, the traffic mix is considered "definitely safe", if it is greater than PLR_{max}, the traffic mix is deemed "definitely dangerous", and if it falls between those two values, i.e., inside the threshold region, the traffic mix is flagged as potentially risky and requiring additional investigations. Ideally, the threshold region should be as narrow as possible, but depending on the accuracy of the model, too narrow a region may either yield a large number of false alarms or fail to properly identify harmful configurations. A wider threshold region reduces those problems at the cost of some imprecision in terms of classifying certain scenarios.

We believe that introducing such a "grey" region provides greater flexibility, and can help minimize the impact of inaccuracies inherent in our evaluation methodology. Our hope is that a reasonably narrow threshold region is sufficient to flag the various traffic mixes for which the model might provide inaccurate answers.

C. Real-time Traffic Traces and Their Characteristics

In this section, we describe the traffic characteristics of the voice and video sources we use to test our model, and the method we use to map them onto Markov ON-OFF sources.

1) Source Traces: The video traces we use to conduct our study are obtained from an online video trace library [2]. We chose this public library because of the diversity of the traces it provides. The library contains traces of 26 different hour-long video sequences, each encoded using two different standards (the MPEG4 and the H.263 standard), and under three different quality indices (low, medium and high). Moreover, for each recorded trace, statistics reflecting its characteristics are also available. We refer to [5] for a detailed description of the trace collection procedure and the methodology used to extract the trace statistics.

Trace Name	Mean Frame Size (Bytes)	MeanBit Rate (Kb)	Frame Size (Peak/Mean)	Frame Size (Covariance)
Jurassic (low)	768.6	153.7	10.6	1.4
Jurassic (high)	3831.5	766.3	4.4	0.6
Lecture (low)	211.7	42.3	16.2	2.3

TABLE I: Video Trace Statistics

In this report, we mainly focus on 3 different sequences encoded using the MPEG4 codec with a frame rate of 25 frames/sec. These include the low quality and high quality traces of the movie Jurassic Park I and the low quality trace of a lecture video. The high quality trace of Jurassic Park I (Jurassic high) represents a video source that requires high transmission rate due to its quality requirement. The low quality trace of Jurassic Park I (Jurassic low) represents a

lower rate source encoded from the same movie scene. At last, due to its scene characteristics, the low quality lecture trace (Lecture low) is selected because it has the lowest mean rate among all the traces in the video library. Several important statistics of the three video sources are given in Table I.

We also use voice traces from [7] in our investigation. In particular, the voice source[1] that we choose corresponds to Fig. 4b in [7] and was generated using the NeVoT Silence Detector(SD) [11]. We refer to [7] for a complete description of the codec configuration and trace collection procedure. In the following table important statistics of the voice source are given.

Trace Name	Mean Spurt Length (ms)	Mean Gap Length (ms)	Peak Rate (kb)	Utilization
Voice	326	442	64	0.425

TABLE II: Voice Trace Statistics

2) Converting to Markov ON-OFF sources: As mentioned before, both the voice traffic and the video traffic inherit built-in ON-OFF patterns, which facilitates the usage of a Markov ON-OFF source to model them. Our task, is thus to develop a simple methodology that can efficiently map the real traffic sources onto the Markov ON-OFF source, based on their characteristics reflected by statistics given in Table I and II.

The raw video traffic consists of variable size frames that arrive at regular time intervals (40ms for the trace that we use). In our study, we assume that *whole* frames arrive instantaneously, so that the access link serves as a shaper, bounding the maximum rate of the output traffic by its own transmission speed. Assuming a fluid model, the output traffic can be thought of as a distorted version of the input traffic, where the frame transmission (ON period) extends over a time interval that is a function of both the frame size and the input link speed. If the link speed is too slow to transmit a frame in 40ms, consecutive frames will be concatenated and form an extended burst. In the experiments that we conduct, we deliberately eliminate such frame concatenation by setting the link speed larger than the peak frame-level bit rate [2]. The shaped traffic emanated by the access link can thus be modeled by an equivalent ON-OFF source, with:

$$R = R_{link}; \quad b = \frac{F_{avg}}{R_{link}}; \quad \rho = \frac{b}{40 \cdot 10^{-3}} \tag{5}$$

where R_{link} is the access link speed and F_{avg} is the average frame size of the video trace.

We do not expect that such a simple source model will fully capture all the complex traffic characteristics of a video source. For example, although the mean of the burst duration is accurately captured, its distribution does not have to be

[1]We use only one voice trace since the usage of voice sources in this paper is mainly to evaluate the performance deviation when it is mixed with video sources. Because of their statistical similarity, different voice sources will experience similar losses when they are aggregated. This was verified in simulations that are omitted here.

[2]We found that if the access link speed is less than the peak frame-level bit rate, then several extremely long bursts may occur and the delay at the access link could be quite large.

exponential. This may affect the accuracy of estimating the absolute loss probability. However, we believe that using merely first order statistics to predict loss deviation is a reasonable approach. This assumption is based in part on the experience obtained from [13], where we saw that deviations were to a large extent a function of the first order statistics of a source, e.g., the peak rate, the utilization, and the average burst duration, among which the first two parameters usually played a dominant role. Furthermore, as discussed in Section VI, there are actually many cases for which the differences between the model's prediction and the actual deviation can be shown to be small, and most important, independent of the burst duration distribution.

As with video sources, the traffic of voice sources can be divided into "ON" and "OFF" periods, corresponding to the talk spurt and silence period in human speech. Again, while those periods need not always be exponentially distributed, our expectation is that this will only minimally impact our ability to predict loss probability ratios when mixing voice and other traffic sources. Hence, the mapping of voice sources onto ON-OFF sources is carried out using the values shown in Table II, i.e., $R = 64$Kb$, b = 326$ms and $\rho = 0.425$. Note that since voice traffic is reasonably smooth, i.e., during a burst period the voice source generates multiple packets of 80 Bytes with an equal spacing of 10ms, the peak rate of the equivalent ON-OFF voice source is fixed at 64kb instead of equal to the link capacity[3]. As we shall see later, this smoothness enables the voice source to avoid major performance degradation when aggregated with video sources, in spite of its small traffic contribution.

Once video and voice sources have been mapped onto Markov ON-OFF sources, we can apply either Equations (1) and (2) (in the two-source case) or Equations (3) and (4) (in the many-source case) to predict the amount of deviation experienced by individual users. More specifically, we first use either Equations (1) (2) or Equations (3) (4) to estimate the amount of bandwidth needed to guarantee an overall loss target P_{max}, denoted as C_{est}. We then bring C_{est} back to the same set of equations so that predictions of both the individual and overall loss probabilities, and thus the loss probability ratio, denoted as PLR_{est}, are achieved. We use the analytical model, instead of relying on simulations, to evaluate the amount of bandwidth needed for an overall loss probability target, since the goal of our methodology is to provide an *off-line* solution for efficient deviation prediction with a minimally computational cost. Simulations, on the other hand, can give us more accurate estimations of actual capacity allocated, but will significantly increase the computational task and require the access to the source traces.

D. Simulation Environments and Experiment Configurations

In order to assess the accuracy of our simplified model, we rely on simulations to evaluate the true loss probabilities and loss probability ratios. We use the NS2 [1] simulator, as it accepts real traffic traces and can accurately model packet level behaviors. In all the simulations, the packet size of the voice traffic is fixed at 80Bytes while the packet size of the video traffic is fixed at 100Bytes, so that a video frame larger than 100Bytes is fragmented into multiple packets. Throughout the paper, the threshold region is set as $[3, 5]$ and the target loss probability P_{max}, when fixed,

[3]The peak rate will be equal to the access link speed only when it is less than 64Kb, which is lower than the values assumed in this paper.

is set to 10^{-4} [4]. To eliminate the impact of possible synchronization between different video streams, each source is assigned a random phase that is $i.i.d$ within the frame level. For simplicity, most of our experiments consider only two different types of sources, where one of the two types of sources is selected and has its parameters varied so that it experiences the larger loss probability.

III. Loss Deviation Prediction When Aggregating Two Sources

This section is devoted to scenarios with only two sources are aggregated. This basic configuration, despite or rather because of its simplicity, provides useful insight in capturing the complex behavior of loss deviation. From [13], we know that a user experiences losses that differ significantly from the overall losses, if it both contributes significantly to the onset of congestion when active, and does it in a way that is undetected by resource allocation mechanism. When applied to two (few) sources, this rule indicates that the utilization of an individual source plays a dominant role in determining any loss deviation it experiences. Specifically, when multiplexing two sources, if one of them has a much smaller utilization than the other but a comparable peak rate, its total traffic contribution will be minor and unlikely to trigger the allocation of significant resources, yet its impact on congestion, and therefore losses, when active can be substantial. This observation was quantified in [13] using analytical models, and as we will see in this paper, also apply to the real-life traffic sources we consider. More important, we will establish that the simple methodology we have developed, is indeed capable of accurately capturing this qualitative behavior.

A. Aggregating Two Video Sources

In this sub-section, we consider scenarios consisting of aggregating two video sources. The first case we consider involves multiplexing Jurassic low and Jurassic high, which corresponds to two video streams of similar characteristics but different quality and, therefore, bit rate. A second scenario considers the case of two video streams, Lecture low and Jurassic high, which differ in both characteristics and quality (bit rate). In each case, we select the lower bit rate sequence to be the one experiencing the larger loss probability. The buffer size is set to 5000 Bytes and to 2000 Bytes[5] in the first and second scenarios, respectively.

The loss probability ratio of each scenario is shown in Figure 1. From Figure 1(a), we see that when Jurassic low and Jurassic high are multiplexed, performance deviations can arise, as the loss probability ratio approaches a value of about 6. More important, the simple model we propose can be seen to accurately capture the trend and magnitude of this deviation. The relatively limited range of the loss probability ratio (it never exceeds 6) can be explained by the fact that although the two sources differ in utilization, the low bit rate source, Jurassic low, is not a negligible contributor to the overall traffic. This in turn limits the extent to which the performance experienced by the Jurassic low stream can deviate from the overall performance. In general, the impact of the relative traffic contribution of the two sources being

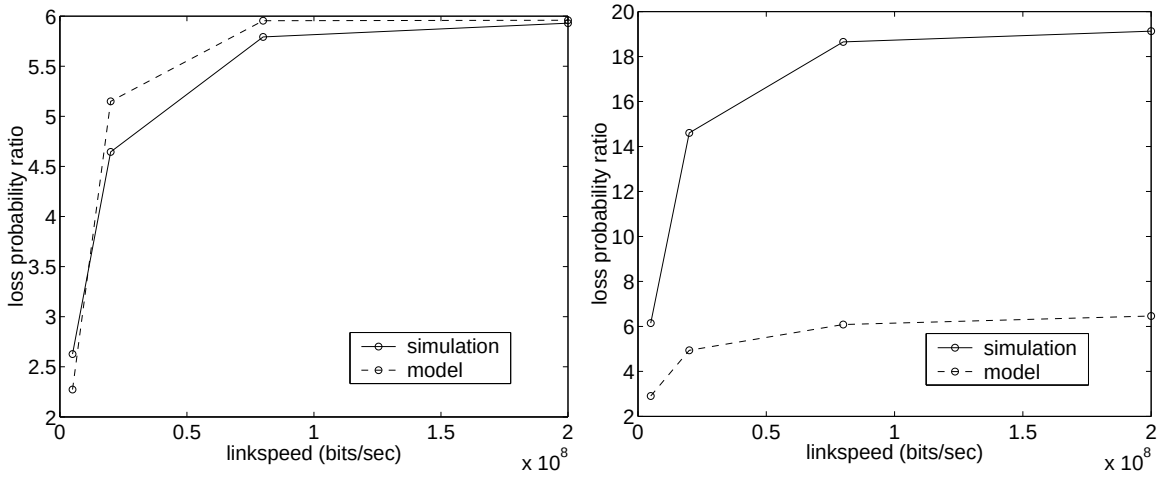
[4]In order to reduce the computational cost, we allow a 6% tolerance region when evaluating the required bandwidth C, i.e., the actual loss probability is allowed to be within the range $(9.97 * 10^{-5}, 10.03 * 10^{-5})$.

[5]We also explore different buffer size (200Bytes) but didn't find significantly different results from the data here.

multiplexed can be captured through an upper bound on the loss probability ratio that only involves the mean rate of the two sources. Specifically, in a configuration with N sources being multiplexed, where, without loss of generality, the N-th source is the one experiencing the larger loss probability, the loss probability ratio satisfies:

$$PLR = \frac{r_L^N / r_S^N}{r_L / r_S} \leq \frac{r_L^N / r_S^N}{r_L^N / r_S} = \frac{1}{r_S^N / r_S}, \tag{6}$$

Equation (6) states that the loss probability ratio PLR can never exceed a value that is inversely proportional to the fraction of traffic contributed by the worst performer in a service class. The smaller the fraction this user contributes, the larger the maximum loss probability ratio it may experience. In this example, Jurassic low is the source that loses more and from Table I, we see that $r_S^2 = 153.7$Kb, which is about 16.7% of the total traffic, resulting in a loss probability ratio satisfying $PLR \leq 6$. As can be seen from Figure 1(a), as the access link speed of Jurassic low increases, the loss probability ratio will indeed approach (and never exceed) this value. Equation 6 indicates that when multiplexing two (or a small number of) sources, the more likely candidate for experiencing significant loss deviation is the source with the smallest bit rate. Furthermore, Equation 6, also shows that in the case of two sources, the higher bit rate source can never experience a loss probability of more than twice the overall loss probability.

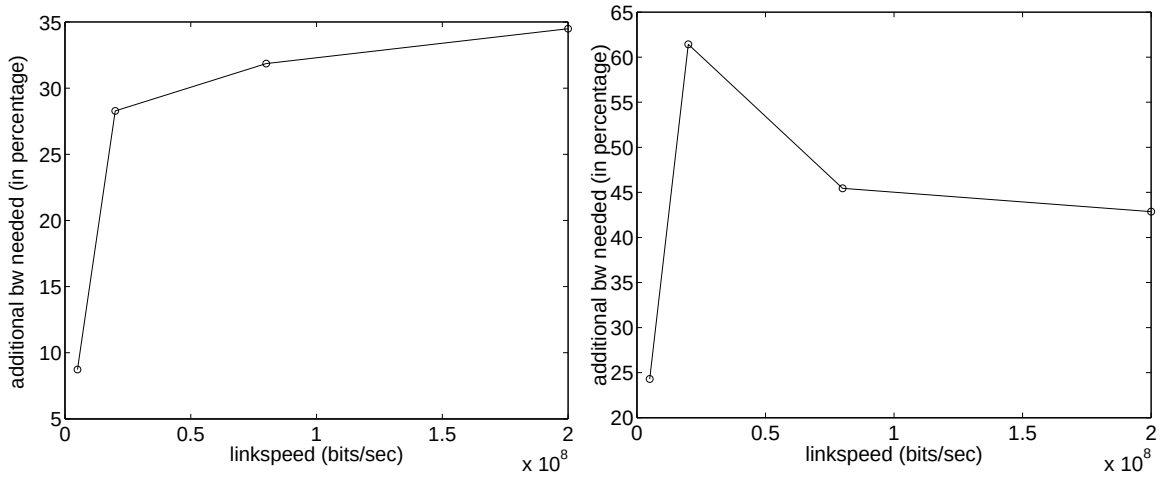


(a) Aggregating Jurassic Low and Jurassic High (b) Aggregating Lecture Low and Jurassic High

Fig. 1. Loss Probability Ratio When Aggregating Video Sources

Figure 1(b) displays the loss probability ratio when multiplexing Lecture low and Jurassic high, for which a maximum value even greater than that of the first scenario can be attained. In particular, we see that the Lecture low source can experience losses that are up to 20 times larger than the overall losses. This is primarily due to the lower (relative) rate of Lecture low, which only contributes about 5.2% of the total traffic. From equation (6), this implies that the loss probability ratio should satisfy $PLR \leq 19.1$. From Figure 1(b), we see that as the link speed increases, the loss probability ratio indeed approaches this upper bound. The loss probability ratio predicted by the model is, however, not as accurate as in the case of Figure 1(a), as it consistently underestimates the actual loss probability ratio. This is because high order statistics of the traffic generated by both sources still influence the magnitude of the loss probability

ratio. This impact notwithstanding, the model still generates a value that in most cases is sufficiently high to at least trigger a warning regarding to the potential danger of multiplexing these two sources. In particular, except for the lowest link speed, the loss probability ratio predicted by the model is always within or above the selected threshold region of $[3, 5]$.



(a) Aggregating Jurassic Low and Jurassic High (b) Aggregating Lecture Low and Jurassic High

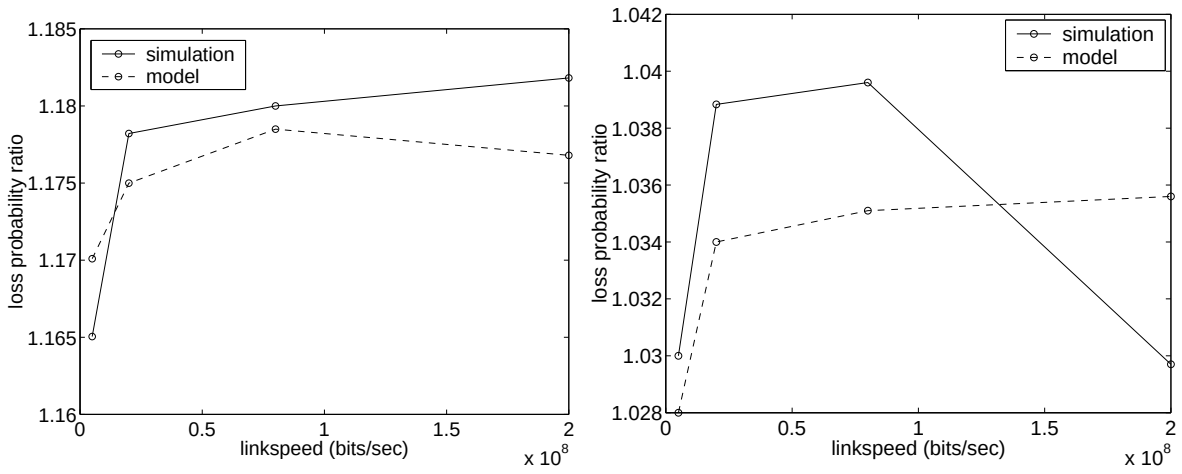
Fig. 2. Additional Bandwidth Needed When Aggregating Video Sources

Figure 2 provides a different perspective on the difference in performance between sources in the two previous scenarios. The figure shows the percentage of additional bandwidth needed in order to ensure the desired loss probability to even the poorer performer of the two sources. As can be seen, this amount can be quite large (30% for the more benign case of Jurassic low and Jurassic high, and up to 60% for the more severe case of Lecture low and Jurassic high). This means that the observed differences in performance cannot be easily fixed through standard over-provisioning, and identifying such potentially dangerous scenarios is, therefore, important.

B. Aggregating One Video Source and One Voice Source

In this section, we study scenarios where one video source and one voice source are multiplexed. From Table I and Table II, we can see that the voice source has a mean transmission rate that is much lower than that of the video source, for both the Jurassic low and Jurassic high sources. When used in equation (6) this translates into a relatively large upper bound for PLR, which could, therefore, indicate a potentially dangerous traffic mix. This turns out not to be the case because of the reasonably smooth nature of the voice source, which makes it mostly immune to performance deterioration when multiplexed with either the Jurassic low or the Jurassic high video sources. Instead, it is the video source that experiences somewhat poorer performance. The magnitude of this performance deviation is, however, limited, because the video source is the major contributor to the overall transmission rate (Equation (6) gives an upper bound of 1.04 and 1.18 for the PLR for Jurassic low and Jurassic high, respectively). As a result, our main concern in these scenarios is whether the model will generate a false alarm, i.e., falsely predict that mixing a voice and a video

sources is potentially dangerous.



(a) Aggregating Voice Source and Jurassic Low (b) Aggregating Voice Source and Jurassic High

Fig. 3. Loss Probability Ratio When Aggregating One Video Source and One Voice Source

The corresponding results are shown in Figure 3. From the graphs, we see that for both Jurassic low and Jurassic high, the model tracks the actual loss probability ratio reasonably accurately. More important, it will not generate any false alarm and will consistently identify the traffic mixes as safe. As an aside, the amount of additional bandwidth needed to ensure that both the voice and video sources experience the desired loss probability target can also be found to be small (less than 3%), which further confirms that mixing a voice and a video source should be fine in practice.

IV. Loss Deviation Prediction When Aggregating Many Sources

The previous section dealt with the special case of only two sources, which while interesting in its own right, is clearly not the only or even most common scenario one would expect to encounter in practice. In this section, we consider another set of possibly more realistic scenarios consisting of mixing many voice and video sources. As before, our goal is to determine whether the model is capable of accurately distinguishing between safe and unsafe traffic mixes. For sake of simplicity and in order to limit the number of combinations to consider, we limit ourselves to multiplexing two different types of sources.

From the discussion of Section III, we know that for one or more sources to experience substantially worse losses than other sources, the peak rate of the source should typically be high while its contribution to the total rate should be small enough to ensure that the losses experienced by itself only have a small impact on the overall loss probability. The latter typically implies a combination of a low average transmission rate and a small number of its peers. We therefore concentrate on such scenarios and evaluate the model's ability to properly identify cases when severe loss deviations can occur.

A. A First Many-Source Example

In this first example, the dominant traffic contributors consist of 500 voice sources that are multiplexed with a *single,* Jurassic low video source[6]. The loss performance deviation is examined for a system where the buffer size is set to 5000 Bytes and for configurations where the access link speed of the video source increases from 5Mbps to 200Mbps, while the access link speed of the voice sources is fixed at 5Mbps. Given that the Jurassic low source has a reasonably high peak rate but only contributes a small amount to the total bit rate, we expect this scenario to be a prime candidate for generating significant differences in the loss probabilities that the voice and video sources experience. The resulting loss probability ratio is shown in Figure 4.

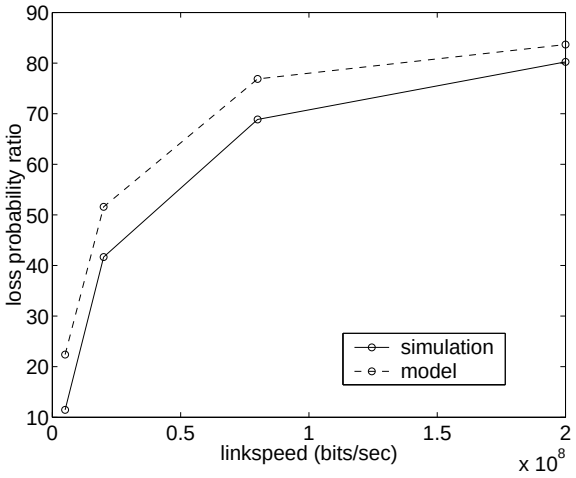


Fig. 4. Loss Probability Ratio When Aggregating 500 voice sources and one video source (Jurassic low)

As seen in Figure 4, as the video source's access link speed grows, so does the loss probability ratio, and this evolution is accurately captured by the model even if it slightly over-estimates the actual value. This over-estimation does, however, not affect the model's ability to correctly identify cases where significant deviations occur. The figure also shows that for the configurations considered, the range of possible deviations is substantially larger than that of the two-source scenarios of Section III. This is not unexpected, as the traffic contributed by the video source is much smaller (only 1.12%) in comparison to the total traffic volume than was the case in the two-source scenarios. This very small contribution to the overall traffic translates into an upper bound of 89.5, when using Equation 6, and we can again see that this value is essentially achieved at high link speeds.

As for the two-source case, we also investigate the amount of additional bandwidth needed to ensure that all sources experience at least the target performance of 10^{-4}. The results are shown in Figure 5, where we can again see that over-allocation is not an option in the cases where substantial performance deviations exist. In particular, when the link speed is high, more than twice the allocated bandwidth is required to guarantee adequate performance for the video sources.

[6]We have conducted another set of experiments in which 500 voice sources and a single Jurassic high video source are aggregated. Since the results are similar to those reported in this section, both qualitatively and quantitatively, we omit reporting it here

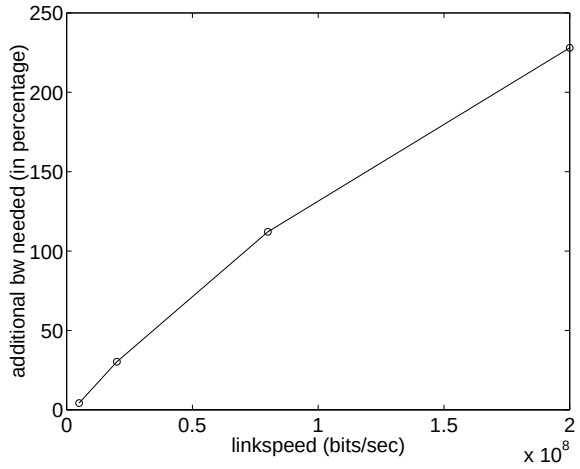


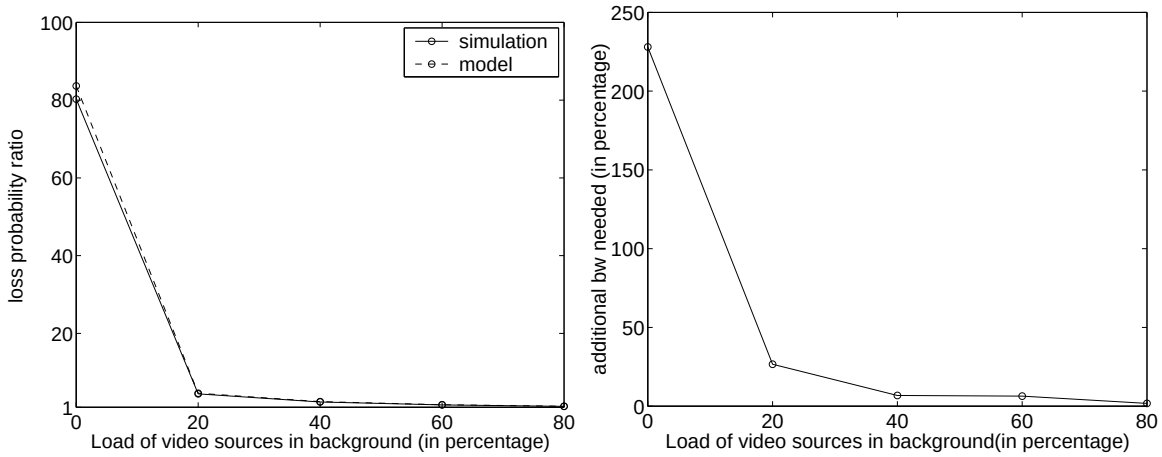
Fig. 5. Additional Bandwidth When Aggregating 500 voice sources and one video source (Jurassic low)

The example of this section is representative of most of the many-source cases we have investigated, even if it only involves a single video source and many voice sources. As discussed earlier, significant performance deviation arises in the many-source case only when sources have both a high peak rate (high enough to trigger the onset of congestion when the source becomes active) and represent a small contribution to the overall bit rate (their higher losses do not contribute much to the overall loss probability). The one video source, many voice sources scenario is a perfect illustration of such a configuration. In contrast, a scenario with a small number of voice sources and many video sources would not result in any significant deviations and would be very close to what was observed in the previous section for a one voice source and one video source configuration. In the next section, we consider a number of intermediate scenarios that help illustrate how fast one transitions from one behavior to the other.

B. Transiting From Unsafe To Safe Traffic Mixes

In this section, we consider a number of scenarios consisting of a mixture of voice and video sources, and where the fraction of traffic contributed by each type of sources varies. Specifically, we start with a configuration consisting of 500 voice sources and one video source (Jurassic low). The access link speed of voice sources is taken to be 5 Mbps, while it is set to200 Mbps for the video source. We then increase the number of video sources while keeping the total traffic rate fixed. As a result, the number of voice sources decreases and the total traffic contribution of video sources increases, which affects the possibility for performance deviations. In particular, as the traffic contribution from video sources becomes dominant, performance deviations are all but eliminated.

The results are shown in Figure 6(a), which illustrates that as the traffic contribution from the video sources first increases, the loss probability ratio experiences a sharp initial drop. This is followed by a slower decrease towards a level where performance deviations are essentially negligible. Specifically, when video sources contribute about 20% of the total traffic (this corresponds to about 20 video sources), the loss probability ratio has dropped to 4.6, down from 80 for a single video source. The loss probability ratio drops further to 2.3 for a video traffic contribution of 40%, and to 1.7 for a video traffic contribution of 60%. As illustrated in Figure 6(b), this general behavior is mirrored in the



(a) Loss Probability Ratio (b) Additional Bandwidth in Percentage

Fig. 6. A Variable Traffic Mix Scenario

amount of additional bandwidth needed to eliminate the performance penalty incurred by video sources.

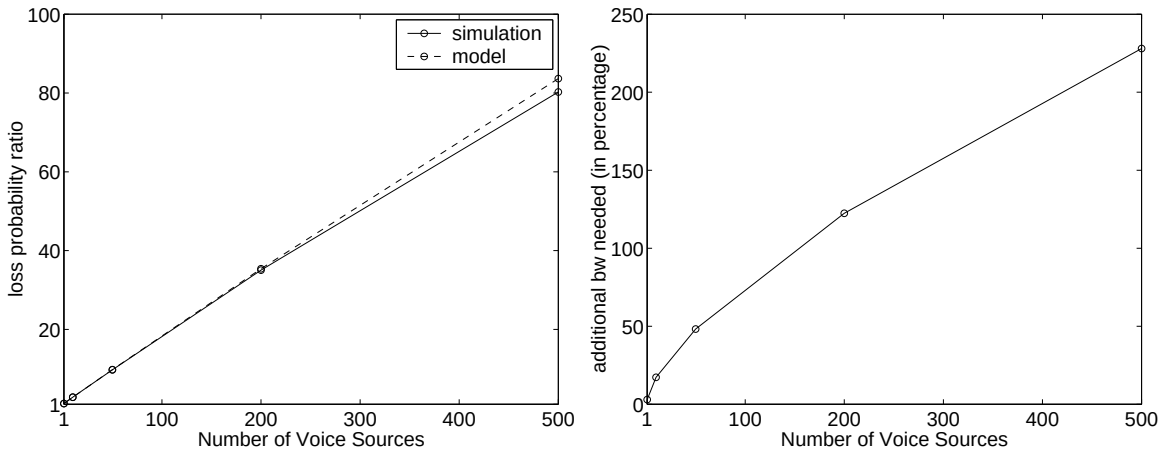
More important from the perspective of assessing the ability of our model to accurately predict whether it is safe or not to mix traffic, the model consistently tracks actual values across the range of scenarios. From a practical standpoint, this section provides an important guideline for determining when it is safe to mix traffic sources that differ significantly in both their peak and mean rates. Specifically, although there are a number of configurations for which it is safe to mix such sources, transiting from unsafe to safe configurations is not a well demarcated process. As a result, a generally safe practice is to altogether avoid multiplexing sources that differ too much in both their peak and mean rates.

V. Intermediate Configurations

The main purpose of this section is to investigate further the process of traversing from safe to unsafe traffic mixes. In particular, we have seen that the behavior of loss performance deviation differs significantly between the two-source and the many-source cases. In the two-source case, source utilization alone can result in significant performance deviations, while in the many-source case both peak rate and utilization are involved. In this section, we study a few intermediate configurations that can help shed some light on the transition between the two-source and the many-source behaviors. Our goal is to ensure that this understanding is adequately incorporated into the simple methodology we propose to determine whether it is safe to multiplex different traffic sources.

The configuration we use for our investigation consists of multiplexing a variable (from 1 to 500) number of voice sources with one video source (Jurassic low). This allows us to explore the evolution from a safe, two-source traffic mix to an unsafe, many-source mix. The loss probability ratios and the additional bandwidth required (in percentage) are plotted for this range of scenarios in Figure 7(a) and Figure 7(b) respectively.

As can be seen from the figure, the model is quite accurate in predicting the loss probability ratio across all scenarios. In addition, the figure shows that the transition from safe to unsafe regions is progressive, roughly following a linear function, but with a reasonably steep slope. In particular, mixing the video source with just 50 voice sources already



(a) Loss Probability Ratio (b) Additional Bandwidth in Percentage

Fig. 7. Aggregating Voice Sources and Jurassic (low): Increasing Number of Voice Sources

yields a loss probability ratio close to 10, and it would take 50% more bandwidth to fix the problem. This is quite significant, especially in light of the fact that the traffic contribution of the video source is about one tenth of the total traffic, i.e., not an insignificant amount. This again highlights the fact that it is better to err on the side of caution when considering mixing traffic sources with rather disparate traffic characteristics, i.e., peak and mean rates.

VI. Discussions

A. Discussions of Prediction Errors

One of the main assumptions behind the approach used in this paper, is that while the underlying analytical model may be too simplistic to accurately compute *absolute loss probabilities*, it is capable of robust predictions when it comes to *loss probability ratios.* The obvious question that this assumption raises is "why would two wrongs make a right?" This section is an attempt at answering part of this question.

1) Prediction Error In The Two-Source Case: The error in computing loss probabilities and, therefore, the loss probability ratio, originates from mapping the actual distribution of the burst duration into an exponential distribution. Specifically, the peak rate and the utilization (mean rate) of a source can usually be captured reasonably accurately, and the burst duration is the main parameter whose estimation in problematic. As a result, modeling a traffic source using a Markov ON-OFF source involves selecting a triplet of the form $< R, b^{'}, \rho >$, where $b^{'}$ is an *equivalent* burst duration chosen to yield the same loss probability as the one experienced by the actual source. Because the choice of $b^{'}$ is very much source dependent (see [6] for details on a method for deriving estimates for $b^{'}$) and although its value is often different from the average burst duration b, we nevertheless make the assumption that $b^{'} = b$ in our evaluation. This therefore represents the major source of error, when it comes to computing loss probabilities and the loss probability ratio. Our goal in this section, is to provide a worst case analysis of the magnitude of this error.

Consider a system consisting of two sources multiplexed onto the same link and where, without loss of generality, source 2 is the one that contributes the smallest amount of traffic. Denote as PLR_{est} the loss probability ratio obtained

from the approach proposed in this paper, and let E_{PLR} be the *worst case* possible difference between PLR_{est} and the actual loss probability ratio PLR. The determination of this worst case difference is based on evaluating the maximum possible value of $|PLR_{est} - PLR|$ as a function of the system parameters. A detailed derivation is provided in Appendix I, and we only review the major steps in this section. The derivation of E_{PLR} relies on two key factors. The first one is that the analytical model reflected in Equation (1) and (2), though not applicable for evaluating the magnitude of the individual and overall loss probabilities for general "ON-OFF" sources, i.e., sources not limited to two-state Markovian sources, can still be useful in deriving bounds for these absolute performance measures. This allows us to further derive bounds for both PLR and PLR_{est} through Equation (6). The second factor is that those bounds can be found to depend only on the sources' peak rates, R_1 and R_2, and the allocated capacity C. Furthermore, it is possible to identify regions for the values of these parameters, within which simple expressions can be obtained for those bounds. Most important, in many regions, those expressions give rise to small values for E_{PLR}, which supports our claim of greater robustness of the approach when it comes to evaluating loss probability ratios.

Specifically, one can distinguish three main regions for which different bounds can be derived for PLR (and PLR_{est}):

- Case 1: $R_1 \leq C, R_2 \leq C$
- Case 2: $R_1(R_2) < C < R_2(R_1)$
- Case 3: $C \leq R_1, C \leq R_2$

where Case 2 can be further divided into two sub-regions depending on whether R_1 is larger than R_2 or not.

After computing bounds for PLR and PLR_{est} in each region, one can then obtain bounds for E_{PLR} itself. However, because the allocated bandwidth C in actual system might be different from C_{est}, the estimated value achieved using the model [7], the number of regions that needs to be considered grows from three to nine (all possible combinations). The resulting values for E_{PLR} are given in Table III, with detailed derivations available from Appendix I.

From the table, we can infer that in many cases the value of E_{PLR} will be reasonably small. Specifically, the only cases where a potentially large estimation error can occur is when the allocated capacity C for either the model or the actual system is smaller than the peak rates of the two sources. In all other configurations, one can find that the maximum error remains relatively small.

For example, when both the model and the actual system correspond to Case 1, one can establish that $E_{PLR} = 0$. This is because the loss probability ratio of the two sources satisfies $P_L^2/P_L^1 = \rho_1/\rho_2$, which is independent of the burst duration, and hence of errors in estimating its statistics. Similarly, when neither system falls into Case 3, it is possible to show $E_{PLR} \leq 1$ if the smaller traffic contributor, source 2, satisfies $\rho_1 < \rho_2$ and $R_1 > R_2$. In this case, one can easily verify that $P_L^1 > P_L^2$, and since it is then the major traffic contributor (source 1) that loses more, $PLR = P_L^1/P_L$. This implies that PLR can never exceed r_S/r_S^1, which is less than 2. This then yields $E_{PLR} \leq 1$. In other scenarios,

[7] This is because the capacity allocated in the model is based on the assumption that the ON period is exponentially distributed with mean b, while the actual capacity allocated incorporate the effect of the real source statistics.

Actual Situation \ Model Prediction	Case 1	Case 2	Case 3
Case 1	0	$\begin{cases} <1 & (r_S^1/R_1 < r_S^2/R_2,\ R_1 > R_2) \\ \max[2, \frac{1}{r_S^2/r_S} * \frac{1}{1+R_1/R_2}] - 1 & (r_S^1/R_1 > r_S^2/R_2,\ R_1 > R_2) \\ \frac{1}{r_S^2/r_S} * (\frac{R_1/R_2}{1+R_1/R_2}) & (R_1 < R_2) \end{cases}$	$\frac{1}{r_S^2 / r_S} - 1$
Case 2	$\begin{cases} <1 & (r_S^1/R_1 < r_S^2/R_2,\ R_1 > R_2) \\ \max[2, \frac{1}{r_S^2/r_S} * \frac{1}{1+R_1/R_2}] - 1 & (r_S^1/R_1 > r_S^2/R_2,\ R_1 > R_2) \\ \frac{1}{r_S^2/r_S} * (\frac{R_1/R_2}{1+R_1/R_2}) & (R_1 < R_2) \end{cases}$	$\begin{cases} <1 & (r_S^1/R_1 < r_S^2/R_2,\ R_1 > R_2) \\ \max[2, \frac{1}{r_S^2/r_S} * \frac{1}{1+R_1/R_2}] - 1 & (r_S^1/R_1 > r_S^2/R_2,\ R_1 > R_2) \\ \frac{1}{r_S^2/r_S} * (\frac{R_1/R_2}{1+R_1/R_2}) & (R_1 < R_2) \end{cases}$	$\frac{1}{r_S^2 / r_S} - 1$
Case 3	$\frac{1}{r_S^2 / r_S} - 1$	$\frac{1}{r_S^2 / r_S} - 1$	$\frac{1}{r_S^2 / r_S} - 1$

TABLE III

WORST CASE PREDICTION ERROR: E_{PLR}

as long as neither system falls into Case 3, it is possible to show that E_{PLR} will be a decreasing function of both the fraction of traffic contributed by source 2 and the difference between R_1 and R_2. This is reflected in the expressions $\max\left\{1, \frac{1}{r_S^2/r_S} \cdot \frac{1}{1+R_1/R_2} - 1\right\}$ and $\frac{1}{r_S^2/r_S} \cdot \frac{R_1/R_2}{1+R_1/R_2}$ for the cases $R_1 > R_2$ and $R_1 < R_2$, respectively. In those cases, E_{PLR} will be large only when both the value of r_S^2/r_S is small (the traffic contribution of source 2 is much smaller than that of source 1) and R_1/R_2 is close to 1.

The only other cases where we commit it is possible for E_{PLR} to be large is, as mentioned earlier, when either the modeled or the actual system correspond to Case 3. In this case, either PLR or PLR_{est} can be anywhere in the range 1 to r_S/r_S^2, so that $E_{PLR} = r_S/r_S^2 - 1$, which can be arbitrarily large if the traffic contribution of source 2 is very small. The question is then to determine how common this configuration is.

To answer this question, we carry on a set of extra experiments in addition to the two cases we test in Section III-A. Specifically, we use four different low bit rate sequences and four different high bit rate sequences, among which one low bit rate sequence is arbitrarily selected and aggregated with another arbitrarily selected high bit rate sequence, resulting in a combination of 16 different scenarios. For each scenario, we again increase the access link speed of the low bit rate sequence while keeping the access link speed of the high bit rate sequence constant. For simplicity, we omit reporting the detailed trace statistics and results, and only summarize our major findings. A careful inspection of our results indicates that in almost all the cases, the real system fell into Case 2 when differences between the peak rates of the two sources is large ($R_2/R_1 > 4$). Meanwhile, only when multiplexing video sources of rather different rates, e.g., Lecture low and Jurassic high, will the modeled system (erroneously) fell into Case 3. This indeed translated into substantial differences between the loss probability ratio predicted by the model and the value of PLR obtained in the

simulations, as is seen in Figure 1. However, it is worth noting that in spite of those differences, the tests performed using the model to determine if it was safe or not to multiplex the sources, still provided a correct answer in most cases. Specifically, among *all* of the additional tests we have conducted, although the modeled system sometimes wrongly fell into Case 3^8, only one scenario fails the consistency test, i.e., the model prediction fails to generate warnings to separate dangerous traffic mixture. This is an evidence that in most cases using the average burst duration b one can still efficiently capture the actual burst process, which warrants the use of our simple methodology for determining whether an aggregation procedure is safe or not.

In a summary, this section together with the results of the many different cases we tested, provides some justification for our claim of robustness in estimating the loss probability ratio, in the two-source case, of the simple analytical model we rely on. In the next section, we briefly explores whether this conclusion extends to the many-source case.

2) Prediction Errors When Aggregating Many Sources: In the many-source case, the equations used to predict the loss probability ratio, i.e., Equation (6) and (7) of [13], are derived assuming a bufferless model. For such a system, the loss probability ratio is only affected by the peak rate and the utilization [10]. This is because data losses occur *only* when the arrival rate is greater than the link capacity, and thus are determined only by their difference. As a result, the loss probability ratio does not depend on the statistics of burst duration. Instead, the model's accuracy depends only on how good a fit a bufferless model is for the real system. This is a function of both the buffer size and the relative burstiness of the source(s).

When the number of sources is large, the ability of a buffer, even a large one, to accommodate many large bursts is very limited, so that a bufferless model is often quite accurate. From the simulations of Section IV and Section V, we have indeed seen that a buffer size of 5000 Bytes gives results that are barely distinguishable from those of a bufferless system, even in cases where the number of sources is moderate. In other simulations that are not included in the paper not shown here, similar observations were made for systems with much larger buffers, i.e., 100 kbytes. In those cases, although there were instances where the absolute prediction was not entirely accurate, it always remained sufficient to clearly identify when it was dangerous to multiplex sources of different types. As a result, we believe that the performance of an approach based on a bufferless model should remain robust across a wide range of buffer sizes, as long as the number of sources being multiplexed is large enough.

VII. Conclusions

This paper is concerned with the problem of deciding when it is possible to multiplex different traffic sources into a common service class, when both resource allocation and performance monitoring is done at the class level. Our goal is to develop a reasonably simple methodology for identifying traffic mixes that can generate significant performance deviations. The methodology we propose is based on a set of analytical tools developed in [13], and which are used together with a simple method for mapping real traffic sources onto source models that the analysis is capable of handling.

[8]This happens mostly when the Lecture low sequence and another high bit rate sequence are aggregated

The effectiveness of the methodology is evaluated through simulations by exploring its prediction ability across a number of scenarios involving real traffic sources. The paper establishes that the methodology, though simple, is quite successful and robust in identifying traffic mixes that can result in severe loss performance deviations. In particular, it reliably identifies cases where large deviations would occur and mostly avoids false alarms. In addition, the region for which it gives somewhat inconclusive results, its so-called threshold region, is relatively narrow. Furthermore, the paper also shows that in most cases where large performance deviations are observed, the amount of additional bandwidth needed to provide all users in the service class with the desired performance, can often be quite large. This indicates that simply over-provisioning the bandwidth allocated to the service class cannot easily eliminate loss performance deviations, when present. As a result, providing a simple yet accurate method for identifying potentially dangerous traffic mixes is important when contemplating deploying class-based services.

VIII. Acknowledgement

The authors would like to acknowledge Wenyu Jiang in Columbia University for providing the voice traffic traces and Frank Fitzek in Technical University of Berlin, Germany for providing the video traffic traces.

Appendix I
Deriving Expressions of E_{PLR} in Table III

In this section, we derive the expressions of E_{PLR} in Table III. In Section VI, E_{PLR} is defined as the maximum possible prediction error, i.e., the maximum possible difference between the estimated loss probability ratio PLR_{est} and the actual loss probability ratio PLR. Hence, a natural direction in deriving E_{PLR} is to attain upper bounds for $|PLR_{est}-PLR|$. In order to achieve this, we first derive intervals that are able to bracket the possible values of PLR_{est} and PLR. The expressions of these intervals though, is dependent on the relationship between the allocated capacity and the peak rates of the two sources and can be shown to be different across the three cases given in Section VI. Furthermore, since the estimated capacity C_{est} may be different from the actual allocated capacity C, the total number of configurations will increase from three to nine in order to cover all the possible combinations. For each of these nine cases, we evaluate the value of E_{PLR} by considering the maximum possible distance between the interval PLR falls in and the interval PLR_{est} falls in. As we will see in later part of this section, in many cases such distance is small and more important, independent of the detailed distribution of the average burst duration b.

A. The Range of PLR

We start our derivation for E_{PLR} by evaluating the possible range that PLR could fall in. We first consider a general system where there are N sources aggregated, which we then specialize to the two-source system by letting $N = 2$. Based on its definition given in Section II-B.3, the loss probability ratio PLR in such a system can be written as:

$$PLR = \max_{i=1:N}\{PLR_i\} \tag{7}$$

where PLR_i is the ratio between the loss probability of a source i and the overall loss probability. Note that an upper bound for the loss probability ratio of each individual source could be easily achieved by generalizing Equation 6 from source N to every source. More specifically, for source $i(1 \leq i \leq N)$, we would have:

$$PLR_i = \frac{r_L^i / r_S^i}{r_L / r_S} \leq \frac{r_L^i / r_S^i}{r_L^i / r_S} = \frac{r_S}{r_S^i} \tag{8}$$

By plugging this upper bound into (7) and considering the fact that the loss probability ratio will always be at least 1, we can obtain a most obvious range that PLR falls in, namely, $PLR \in [1, \max_{i=1:N}\{r_S / r_S^i\}]$.

For the case where there are only two sources aggregated, i.e., $N = 2$, the above results indicate that $PLR = \max\{PLR_1, PLR_2\}$ and $PLR \in [1, \max\{r_S/r_S^1, r_S/r_S^2\}]$. In order to get a narrower range for PLR, we need to derive tighter bounds for both PLR_1 and PLR_2. Not losing generality, we assume that source 2 is the minor traffic contributor, i.e., $r_S^1 > r_S^2$ and first proceed to derive PLR_2. PLR_1 can then be achieved based on the expression of PLR_1 simply by exchanging indices.

The first step in deriving the range for PLR_2 involves decomposing the loss probability ratio into two parts, one of which is only affected by the peak rate R_2 and the utilization ρ_2, while the other can be affected by the burst duration. In particular, PLR_2 could be written as:

$$PLR_2 = \frac{P_L^2}{P_L} = \frac{1}{r_S^2 / r_S} * \frac{1}{1 + r_L^1 / r_L^2} \tag{9}$$

We can see that the LHS of the resulting expression is simply the upper bound of PLR_2 given in Equation 8. In the RHS of the equation, r_L^1 / r_L^2 is the only component affected by the burst duration of the two sources, and thus will be the focus of the following part of derivation.

Notice that the analytical model described in Section II-B, in the form of Equation (1) and (2), can still be utilized to evaluate r_L^1 and r_L^2, as long as the two sources aggregated are ON-OFF sources. In particular, for source i (i=1,2), we will have:

$$r_L^i = \sum_{\substack{\mathbf{S}:(s_i=1,\\ \gamma_{\mathbf{S}} > C)}} (\pi_{\mathbf{S}} - F_{\mathbf{S}}(x))(R_i - C \cdot \frac{R_i}{\gamma_{\mathbf{S}}}) \tag{10}$$

This is because by its definition, r_L^i can be obtained by averaging the loss speed across every state that source i suffers from losses, which is incorporated in Equation 10. However, since the ON and OFF periods do not have to be exponentially distributed, the value of $F_{\mathbf{S}}(x)$ cannot be achieved by previous work such as [8] and [12] that is built upon the Markov ON-OFF sources. Fortunately, the lack of knowledge of $F_{\mathbf{S}}(x)$ will not prevent us from deriving bounds for r_L^i. From Equation 10, one can observe that the value of r_L^i is dependent on the relationship between $\gamma_{\mathbf{S}}$ and C, and thus the relationship between R_1, R_2 and C. Particularly, the value of r_L^1 / r_L^2 will be different across the three different *disjoint* cases described in Section VI, reflected by the following formula:

$$\frac{r_L^1}{r_L^2} = \begin{cases} \frac{(\pi_{<1,1>}-F_{<1,1>}(x))*(R_1-(R_1/(R_1+R_2)*C))}{(\pi_{<1,1>}-F_{<1,1>}(x))*(R_2-(R_2/(R_1+R_2)*C))} & : \quad R_1 \le C \wedge R_2 \le C \\ \frac{(\pi_{<1,1>}-F_{<1,1>}(x))*(R_1-(R_1/(R_1+R_2)*C))}{(\pi_{<1,1>}-F_{<1,1>}(x))*(R_2-(R_2/(R_1+R_2)*C))+(\pi_{<0,1>}-F_{<0,1>}(x))*(R_2-C)} & : \quad R_1 < C < R_2 \\ \left(\frac{(\pi_{<1,1>}-F_{<1,1>}(x))*(R_1-(R_1/(R_1+R_2)*C))+(\pi_{<1,0>}-F_{<1,0>}(x))*(R_1-C)}{(\pi_{<1,1>}-F_{<1,1>}(x))*(R_2-(R_2/(R_1+R_2)*C))}\right) & : \quad (R_2 < C < R_1) \\ \frac{(\pi_{<1,1>}-F_{<1,1>}(x))*(R_1-(R_1/(R_1+R_2)*C))+(\pi_{<1,0>}-F_{<1,0>}(x))*(R_1-C)}{(\pi_{<1,1>}-F_{<1,1>}(x))*(R_2-(R_2/(R_1+R_2)*C))+(\pi_{<0,1>}-F_{<0,1>}(x))*(R_2-C)} & : \quad C \le R_1 \wedge C \le R_2 \end{cases} \tag{11}$$

which can be further written into the following form:

$$\frac{r_L^1}{r_L^2} \begin{cases} = \frac{R_1}{R_2} & : \quad R_1 \le C \wedge R_2 \le C \\ < (>) \frac{R_1}{R_2} & : \quad R_1(R_2) < C < R_2(R_1) \\ \text{uncertain } (> 0) & : \quad C \le R_1 \wedge C \le R_2 \end{cases} \tag{12}$$

Bring the inequalities in Equation (12) into Equation (9) and also consider the simple upper bound $PLR_2 \le r_S/r_S^2$, we can easily attain the range of PLR_2, as is shown in below:

$$PLR_2 \begin{cases} = \frac{r_S}{r_S^2} \cdot \frac{1}{1+R_1/R_2} & : \quad R_1 \le C \wedge R_2 \le C \\ \in \left[\frac{r_S}{r_S^2} \cdot \frac{1}{1+R_1/R_2}, \frac{r_S}{r_S^2}\right] & : \quad R_1 < C < R_2 \\ \in \left[0, \frac{r_S}{r_S^2} \cdot \frac{1}{1+R_1/R_2}\right] & : \quad R_2 < C < R_1 \\ \in \left[0, \frac{r_S}{r_S^2}\right] & : \quad C \le R_1 \wedge C \le R_2 \end{cases} \tag{13}$$

Similarly, for PLR_1, we can also have:

$$PLR_1 \begin{cases} = \frac{r_S}{r_S^1} \cdot \frac{1}{1+R_2/R_1} & : \quad R_1 \le C \wedge R_2 \le C \\ \in \left[0, \frac{r_S}{r_S^1} \cdot \frac{1}{1+R_2/R_1}\right] & : \quad R_1 < C < R_2 \\ \in \left[\frac{r_S}{r_S^1} \cdot \frac{1}{1+R_2/R_1}, \frac{r_S}{r_S^1}\right] & : \quad R_2 < C < R_1 \\ \in \left[0, \frac{r_S}{r_S^1}\right] & : \quad C \le R_1 \wedge C \le R_2 \end{cases} \tag{14}$$

To get the range of PLR, the next step is to merge the expressions of PLR_1 and PLR_2, based on Equation(7) and the simple fact that $PLR \ge 1$. In some cases, the range of PLR is easy to get. For example, when both $R_1 \le C$ and $R_2 \le C$, PLR is simply a value equal to $\max\{\frac{r_S}{r_S^2} \cdot \frac{1}{1+R_1/R_2}, \frac{r_S}{r_S^1} \cdot \frac{1}{1+R_2/R_1}\}$. In the case when $R_1 < C < R_2$, we will have: $\frac{r_S}{r_S^1} \cdot \frac{1}{1+R_2/R_1} < 1 < \frac{r_S}{r_S^2} \cdot \frac{1}{1+R_1/R_2}$, due to the fact that $r_S^2 < r_S^1$ and $R_1/R_2 < 1$. Hence, in such case $PLR = PLR_2 \in \left[\frac{r_S}{r_S^2} \cdot \frac{1}{1+R_1/R_2}, \frac{r_S}{r_S^2}\right]$. In addition, when both $C \le R_1$ and $C \le R_2$, we will have $PLR \in \left[1, \frac{r_S}{r_S^2}\right]$ since $r_S/r_S^1 < r_S/r_S^2$. The sub-case when $R_2 < C < R_1$, though, needs more careful inspection. In particular, this sub-case can be further divided disjointly based on whether ρ_1 is greater than ρ_2. When $r_S^1/R_1 = \rho_1 \le \rho_2 = r_S^2/R_2$, we will have: $\frac{r_S}{r_S^2} \cdot \frac{1}{1+R_1/R_2} \le 1 \le \frac{r_S}{r_S^1} \cdot \frac{1}{1+R_2/R_1} \le \frac{r_S}{r_S^1} < 2$. Hence, $PLR = PLR_1 \in [1, 2]$. Conversely, when $r_S^1/R_1 = \rho_1 > \rho_2 = r_S^2/R_2$ we will have: $\frac{r_S}{r_S^2} \cdot \frac{1}{1+R_1/R_2} \ge \frac{r_S}{r_S^1} \cdot \frac{1}{1+R_2/R_1}$, and thus

$PLR \in \left[1, \max\left\{\frac{r_S}{r_S^1}, \frac{r_S}{r_S^2} \cdot \frac{1}{1+R_1/R_2}\right\}\right] \subseteq \left[1, \max\left\{2, \frac{r_S}{r_S^2} \cdot \frac{1}{1+R_1/R_2}\right\}\right]$. In a summary, depending on the relationship between the link capacity allocated and the peak rates of the two sources, PLR can be bounded using the following formula:

$$PLR \begin{cases} = \max\left\{\frac{r_S}{r_S^2} \cdot \frac{1}{1+R_1/R_2}, \frac{r_S}{r_S^1} \cdot \frac{1}{1+R_2/R_1}\right\} & : \quad R_1 \le C \wedge R_2 \le C \\ \in \left[\frac{r_S}{r_S^2} \cdot \frac{1}{1+R_1/R_2}, \frac{r_S}{r_S^2}\right] & : \quad R_1 < C < R_2 \\ \in [1,2] & : \quad R_2 < C < R_1 \wedge \rho_1 \le \rho_2 \\ \in \left[1, \max\left\{2, \frac{r_S}{r_S^2} \cdot \frac{1}{1+R_1/R_2}\right\}\right] & : \quad R_2 < C < R_1 \wedge \rho_1 > \rho_2 \\ \in \left[1, \frac{r_S}{r_S^2}\right] & : \quad C \le R_1 \wedge C \le R_2 \end{cases} \tag{15}$$

B. The Range of PLR_{est}

The derivation of PLR_{est} is essentially similar to the derivation of PLR. In particular, we could follow the same steps in the previous section to derive PLR_{est}, if all the labels of variables obtained by model estimation are tagged using a subscript "est". For example, PLR, PLR_i and C will be substituted by PLR_{est}, PLR_{est}^i and C_{est} respectively. Therefore, for sake of simplicity, we omit presenting the detailed derivations and only report the results. Specifically, the range of PLR_{est} can be reflected by the following equation:

$$PLR_{est} \begin{cases} = \max\left\{\frac{r_S}{r_S^2} \cdot \frac{1}{1+R_1/R_2}, \frac{r_S}{r_S^1} \cdot \frac{1}{1+R_2/R_1}\right\} & : \quad R_1 \le C_{est} \wedge R_2 \le C_{est} \\ \in \left[\frac{r_S}{r_S^2} \cdot \frac{1}{1+R_1/R_2}, \frac{r_S}{r_S^2}\right] & : \quad R_1 < C_{est} < R_2 \\ \in [1,2] & : \quad R_2 < C_{est} < R_1 \wedge \rho_1 \le \rho_2 \\ \in \left[1, \max\left\{2, \frac{r_S}{r_S^2} \cdot \frac{1}{1+R_1/R_2}\right\}\right] & : \quad R_2 < C_{est} < R_1 \wedge \rho_1 > \rho_2 \\ \in \left[1, \frac{r_S}{r_S^2}\right] & : \quad C_{est} \le R_1 \wedge C_{est} \le R_2 \end{cases} \tag{16}$$

C. Expressions of E_{PLR}

In this section, we derive expressions of E_{PLR} for all the possible combinations of regions to which C and C_{est} might belong, as are reflected by the cells in Table III. As mentioned before, the value of E_{PLR} is computed as the maximum distance between any two arbitrary points in the ranges bounding PLR and PLR_{est} respectively.

1) Both Model and Actual Situation Fall into Case 1: In this case, $PLR = PLR_{est} = \max\left\{\frac{r_S}{r_S^2} \cdot \frac{1}{1+R_1/R_2}, \frac{r_S}{r_S^1} \cdot \frac{1}{1+R_2/R_1}\right\}$, thus $E_{PLR} = 0$.

2) Model Falls into Case 1 and Actual Situation Falls into Case 2, or Vice Versa: Not losing generality, we first consider the case when model falls into Case 1 and the actual situation is case 2.

In this case, if $R_1 < R_2$, based on Equations (15) and (16) : $PLR \in \left[\frac{r_S}{r_S^2} \cdot \frac{1}{1+R_1/R_2}, \frac{r_S}{r_S^2}\right]$, $PLR_{est} = \frac{r_S}{r_S^2} \cdot \frac{1}{1+R_1/R_2}$. Hence $E_{PLR} = \frac{r_S}{r_S^2} \cdot \left(1 - \frac{1}{1+R_1/R_2}\right)$.

If $R_2 < R_1$ and $\rho_1 \leq \rho_2$, $PLR \in [1, 2]$ and $PLR_{est} = \frac{r_S}{r_S^1} \cdot \frac{1}{1+R_2/R_1} \leq 2$. Hence $E_{PLR} \leq 1$.

At last, if $R_2 < R_1$ and $\rho_1 > \rho_2$, we will have: $PLR \in \left[1, \max\left\{2, \frac{r_S}{r_S^2} \cdot \frac{1}{1+R_1/R_2}\right\}\right]$ and $PLR_{est} = \frac{r_S}{r_S^2} \cdot \frac{1}{1+R_1/R_2}$. Hence, $E_{PLR} = \max\left\{2, \frac{r_S}{r_S^2} \cdot \frac{1}{1+R_1/R_2}\right\} - 1$.

Combining the results for all of the above sub-cases, we will have:

$$E_{PLR} \begin{cases} \leq 1 & : \quad R_2 < R_1 \wedge \rho_1 \leq \rho_2 \\ = \max\left\{2, \frac{r_S}{r_S^2} \cdot \frac{1}{1+R_1/R_2}\right\} - 1 & : \quad R_2 < R_1 \wedge \rho_1 > \rho_2 \\ = \frac{r_S}{r_S^2} \cdot \left(1 - \frac{1}{1+R_1/R_2}\right) & : \quad R_1 < R_2 \end{cases} \tag{17}$$

if the model falls into Case 1 and the real situation is Case 2.

If the model falls into Case 2 and actual situation is Case 1, one can achieve the the new expressions of PLR and PLR_{est} simply by exchanging their values in the previous case. It is easy to see that in this scenario, E_{PLR} can also be represented by Equation 17.

3) Both Model and Actual Situation Fall into Case 2: In this case, both PLR and PLR_{est} will belong to same range for each sub-case in Case 2. That is, PLR_{est} and PLR will satisfy:

$$PLR = PLR_{est} \begin{cases} \in \left[\frac{r_S}{r_S^2} \cdot \frac{1}{1+R_1/R_2}, \frac{r_S}{r_S^2}\right] & : \quad R_1 < R_2 \\ \in [1, 2] & : \quad R_2 < R_1 \wedge \rho_1 \leq \rho_2 \\ \in \left[1, \max\left\{2, \frac{r_S}{r_S^2} \cdot \frac{1}{1+R_1/R_2}\right\}\right] & : \quad R_2 < R_1 \wedge \rho_1 > \rho_2 \end{cases}$$

Hence, E_{PLR} in such case will be equal to the distance between the left and right boundaries of the interval PLR and PLR_{est} belong to. A simple inspection will demonstrate that the expression of E_{PLR} is still the same as what is given in Equation (17)

4) Either Model or Actual Situation Falls into Case 3: In this case, either PLR or PLR_{est} is in the interval $\left[1, \frac{r_S}{r_S^2}\right]$. Not losing generality, assume that $PLR \in \left[1, \frac{r_S}{r_S^2}\right]$. Then for PLR_{est}, in all the cases except the one where both R_1 and R_2 are less than C_{est}, either the low boundary of PLR_{est} is equal to 1, or the upper boundary of PLR_{est} is equal to $\frac{r_S}{r_S^2}$, which indicates that $E_{PLR} = \frac{r_S}{r_S^2} - 1$. If both R_1 and R_2 are less than C_{est}, $PLR_{est} = \max\left\{\frac{r_S}{r_S^2} \cdot \frac{1}{1+R_1/R_2}, \frac{r_S}{r_S^1} \cdot \frac{1}{1+R_2/R_1}\right\}$. In such case one can still achieve that $E_{PLR} \leq \frac{r_S}{r_S^2} - 1$.

REFERENCES

[1] The network simulator – ns-2. http://www.isi.edu/nsnam/ns/.

[2] Video trace library. http://www-tkn.ee.tu-berlin.de/research/trace/trace.html.

[3] D. Anick, D. Mitra, and M. M. Sondhi. Stochastic theory of a data-handling system with multiple sources. *Bell Sys. Tech. Journal (BSTJ)*, 61(8):1871–1894, October 1982.

[4] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss. An architecture for differentiated services. Request For Comments (Proposed Standard) RFC 2475, IETF, December 1998.

[5] Frank H. P. Fitzek and Martin Reisslein. MPEG-4 and H.263 video traces for network performance evaluation. *IEEE Network Magazine*, 15(6):40–54, November 2001.

[6] L. Gün. An approximation method for capturing complex traffic behavior in high speed networks. *Performance Evaluation*, 19(1):5–23, January 1994.

[7] W. Jiang and H. Schulzrinne. Analysis of on-off patterns in voip and their effect on voice traffic aggregation. In *The 9th IEEE International Conference on Computer Communication Networks*, 2000.

[8] D. Mitra. Stochastic theory of a fluid model of producers and consumers coupled by a buffer. *Adv. Appl. Prob*, 20:646–676, 1988.

[9] I. Norros and J. Virtamo. Who loses cells in the case of burst scale congestion. In *Proceedings of the 13th Intl. Teletraffic Congress (ITC-13)*, pages 59–64, Copenhagen, Denmark, June 1991.

[10] J. W. Roberts, U. Mocci, and J. Virtamo, editors. *Broadband Network teletraffic - Final Report of Action COST 242*, volume 1155. Springer-Verlag, 1996.

[11] Henning Schulzrinne. Voice communication across the Internet: a network voice terminal. Technical Report 92–50, Department of Computer Science, University of Massachusetts, Amherst, MA, USA, 1992.

[12] T. E. Stern and A. I. Elwalid. Analysis of separable Markov-modulated rate models for information-handling systems. *Adv. Appl. Prob.*, 23:105–139, 1991.

[13] Y. Xu and R. Guérin. Individual QoS versus aggregate QoS: A loss performance study. In *Proceedings of the IEEE Infocom*, NYC, USA, June 2002.