Estimates on the Packet Loss Ratio via Queue Tail Probabilities

András György[†]
[†]Dept. of Computer Science and Information Theory,
Budapest University of Technology and Economics
Pázmány P. 1/D, H-1117 Budapest, Hungary

Tamás Borsos[†‡]
[‡]Traffic Analysis and Network Performance
Laboratory, Ericsson Ltd.
Laborc u. 1., H-1037 Budapest, Hungary

***Abstract* — In this paper we consider the connection between the packet loss ratio (PLR) in a switch with a finite buffer of size L and the tail distribution of the corresponding infinite buffer queue Q. In the literature the PLR is often approximated with the tail probability $\mathbf{P}(Q > L)$, and in practice the latter is often a good conservative estimate on the PLR. Therefore, efforts have mainly focused on finding bounds and asymptotic expressions concerning the tail probabilities of the infinite queue. However, our first result shows that the ratio PLR$/\mathbf{P}(Q > L)$ can be arbitrary, in particular the PLR can be larger than the tail probability. We also determine an upper bound on this ratio yielding an upper bound on the PLR using the tail distribution of the infinite queue. The bound is fairly tight for certain traffic patterns. In many situations it clearly improves the estimation with the tail probability, and it is rarely significantly larger than the estimate $\mathbf{P}(Q > L)$, while it is an upper bound. On the other hand, if the PLR is much smaller than $\mathbf{P}(Q > L)$, then our bound is usually loose. For this case a practically good approximation on their ratio is proposed.**

I. Introduction

The emerging integrated services (broadband) telecommunication networks offer a new type of services which, unlike traditional best effort data networks, meet strict Quality of Service (QoS) requirements. From the engineering point of view, one of the most important questions is how to utilize the network resources efficiently, that is, how one can transmit as much traffic as possible while keeping the QoS requirements. To achieve high utilization the burstiness of the sources can be exploited via statistical multiplexing and buffering. However, recent results indicated that the performance of data networks cannot be significantly improved by the use of large buffers [1]. Moreover, the delay requirements of real-time applications also constrain the buffer size. Therefore, the importance of the analysis of finite buffers has considerably increased. In this paper the connection of the finite and infinite buffers is investigated from the point of view of one of the most important QoS parameters, the packet loss ratio (PLR).

The PLR in a buffer of size L is generally estimated with the tail probability $\mathbf{P}(Q > L)$ of the corresponding infinite buffer queue Q. Measurements showed that for generally used traffic models and for some real traffic traces this estimation is in fact an upper bound [2, 3, 4]. Therefore, the tail of queue length distributions (in infinite buffers) has been extensively studied, see, e.g., [5, 6] and the references therein. The asymptotic expressions (or rarely upper bounds) on the tail probabilities are generally obtained for large buffers [7, 8, 9, 10, 11] or many sources [12, 13, 14, 15].

Although Theorem 1 in the next section shows that the ratio PLR$/\mathbf{P}(Q > L)$ can be arbitrary, only a few papers deal with the real PLR. For memoryless arrivals Kelly [5] derived an asymptotic expression for the PLR for large buffers, and in [15] Likhanov and Mazumdar gave an asymptotic formula for the case of many sources.

In Theorem 2 we determine an upper bound on the PLR in a buffer of size L using the tail probability $\mathbf{P}(Q > L)$. The result enables the correct extension of the bounds and asymptotic expressions corresponding to the tail probability to the PLR.

In the third part of the paper the behavior of the new bound is discussed for real traffic traces. The examples also show situations where PLR $> \mathbf{P}(Q > L)$ and where the packet loss is smaller than the tail probability by an order of magnitude. For the latter case an approximation is proposed which proved to be fairly good for real traffic traces.

II. Bounding the packet loss probability

We consider the well-known discrete time model of a switch. Let X_n denote the number of packets arriving in slot n to the switch (the arrival process $\{X_n\}$ is the overall traffic offered by all sources), let Y_n denote the number of packets that can be served in slot n, and let S_n denote the number of waiting packets in the buffer of size L at the end of slot n. Then the queue length can be described by the equation

$$S_{n+1} = \min\left((S_n - Y_{n+1} + X_{n+1})^+, L\right)$$

for $n \geq 0$, where x^+ equals zero if x is negative, and it is x otherwise.

The packet loss ratio is defined as

$$\text{PLR} = \lim_{n\to\infty} \frac{\{\text{no. of packets lost up to slot } n\}}{\{\text{no. of packets offered up to slot } n\}}.$$

To analyze the PLR, an auxiliary infinite buffer queue is introduced:

$$Q_{n+1} = (Q_n - Y_{n+1} + X_{n+1})^+. \tag{1}$$

Concerning the stability of the sequence $\{Q_n\}$, Loynes [16] proved that if the pair $\{X_n, Y_n\}$ is stationary and ergodic, then the queue defined by (1) is stable if

$$\mathbf{E}X_n < \mathbf{E}Y_n \tag{2}$$

for all n. Moreover, there is a unique limit distribution of the sequence $\{Q_n\}$. In what follows we assume that Q has the limit distribution of the sequence $\{Q_n\}$.

As we have mentioned in the introduction, the PLR is often approximated with the tail probability $\mathbf{P}\{Q > L\}$ in the literature. The heuristic considerations behind this decision are the following [17]. The expected number of packets lost in one time slot due to buffer overflow is given by

$$\mathbf{E}(\text{no. of packets lost}) = \mathbf{P}(S \text{ overflows}) \cdot \mathbf{E}(\text{no. of packets arriving while } S \text{ overflows}).$$

0-7803-7206-9/01/$17.00 © 2001 IEEE

The arrivals are approximately independent of the state of the queue, and so the expected number of packets arriving while S overflows is approximately the mean activity of the sources. For stationary and ergodic sources the PLR is

$$\begin{aligned} \text{PLR} &= \frac{\mathbf{E}(\text{no. of packets lost})}{\mathbf{E}(\text{no. of packets arriving})} \\ &= \frac{\mathbf{E}(\text{no. of packets lost})}{\text{mean activity}} \end{aligned}$$

giving

$$\text{PLR} \approx \mathbf{P}(S \text{ overflows}) \le \mathbf{P}(Q > L) \tag{3}$$

However, as the next theorem shows, despite the empirical justifications the above approximate inequality does not hold in general. Moreover, the ratio $\text{PLR}/\mathbf{P}(Q > L)$ can be set arbitrarily.

Theorem 1 *For any $r > 0$ there is a queuing system with constant service rate such that $\text{PLR}/\mathbf{P}(Q > L) < r$, and there is another system such that $\text{PLR}/\mathbf{P}(Q > L) > r$. That is, the ratio $\text{PLR}/\mathbf{P}(Q > L)$ can be arbitrarily small (nonnegative) and arbitrarily large.*

Remark. In particular the theorem implies that the PLR can be significantly larger than $\mathbf{P}(Q > L)$.

Proof. Assume that $Y_n = s$ for all n, $L = Bs$, and let $\{X_n\}$ be a periodic source with one-slot-long peaks followed by $(t-1)$-slot-long low activity periods for some positive integer t, and suppose that the occurrence of the first peak is uniformly distributed in the first t time slots (the latter condition ensures that the sequence $\{X_n\}$ is stationary). The source emits ps packets during peaks and ms packets in all other time slots. (Here we assume that $(p-1)s > L$ to induce packet loss, and $(t-1)m + p < t$ to meet the stability conditions of (2).) More formally, for $k = 1, \ldots, t$ and all positive integer u

$$\begin{aligned} \mathbf{P}(X_1 = ms, \ldots, X_{k-1} = ms, X_k = ps, \\ X_{k+1} = ms, \ldots, X_t = ms) = 1/t \end{aligned}$$

and

$$X_k = X_{tu+k}.$$

Note that the source $\{X_n\}$ is stationary and ergodic. Then in every period of length t starting at a peak activity time slot, $(p-1-B)s$ packets are lost in the finite buffer case if the buffer was originally empty, and at the end of the period the buffer gets empty again (this is guaranteed by the stability condition). On the other hand, $s(p + (t-1)m)$ packets are to be transmitted in every period, hence

$$\text{PLR} = \frac{p-1-B}{p+(t-1)m}.$$

In case of an infinite buffer, at the end of the first slot of the period $(p-1)s$ packets are stored in the buffer, and then emptied at a rate $(1-m)s$. Thus the event $\{Q > L\}$ occurs for $\lceil (p-1-B)/(1-m) \rceil$ time slots. Therefore,

$$\mathbf{P}(Q > L) = \frac{1}{t} \left\lceil \frac{p-1-B}{1-m} \right\rceil.$$

For simplicity assume that $(p-1-B)/(1-m)$ is an integer (the following discussion can also be carried out without this assumption). Then we obtain

$$\frac{\text{PLR}}{\mathbf{P}(Q > L)} = \frac{t(1-m)}{p+(t-1)m}.$$

Now if $m = 0$ and $t \to \infty$, then $\text{PLR}/\mathbf{P}(Q > L) \to \infty$. On the other hand, if $m \to 1$ and, for each m, the value of p is chosen to be very close to $t - (t-1)m$, then $\text{PLR}/\mathbf{P}(Q > L) \to 0$ (for small values of B, otherwise no packet loss occurs). □

The following theorem gives a strict upper bound on the ratio $\text{PLR}/\mathbf{P}(Q > L)$ provided the stability requirements are met. The result can also be used to give an exact upper bound on the PLR when it is combined with different estimations corresponding to the tail distribution of Q.

Theorem 2 *Assume that $\{X_n\}$ is stationary and ergodic, and the service process $\{Y_n\}$ is stationary, memoryless, independent of Q_0 and $\{X_n\}$, and $\mathbf{E}X_n < \mathbf{E}Y_n$ for all n. Let $m \ge 0$ be a real number such that $X_n \ge m$ almost surely. Then*

$$\text{PLR} \le \frac{(\mathbf{E}Y_1 - m)}{\mathbf{E}X_1} \mathbf{P}(Q > L).$$

where Q has the limit distribution of $\{Q_n\}$.

Remarks. Note that (i) equality holds for the source of Theorem 1; (ii) if m, the essential minimum of X_n is unknown, then the theorem yields the (weaker) upper bound

$$\text{PLR} \le \frac{\mathbf{E}Y_1}{\mathbf{E}X_1} \mathbf{P}(Q > L).$$

Proof. The number of packets lost from the finite buffer in slot n is given by

$$X_n^S = (S_{n-1} + X_n - Y_n - L)^+$$

Let $X_n^Q = (Q_n - L)^+$. Since $Q_n \ge S_n$ for all $n \ge 0$ (assuming $Q_0 = S_0$),

$$\begin{aligned} X_n^S &= (S_{n-1} + X_n - Y_n - L)^+ \\ &\le (Q_{n-1} + X_n - Y_n - L)^+ \\ &= (Q_n - L)^+ = X_n^Q. \end{aligned} \tag{4}$$

That is, in each time slot n, the number of lost packets in the finite buffer system can be bounded from above by the number of packets *overflown* in the infinite buffer system (we call a packet overflown if there are at least L packets waiting at the queue when it arrives). Without loss of generality we can assume that first the overflown packets are transmitted from the infinite buffer queue (this assumption clearly does not modify the number of packets waiting at the queue). In each

time slot n the number of overflown packets waiting in the infinite queue is decreased by at most $Y_n - m$. Moreover, observe that if $Q_n \leq L$, then no overflown packets are waiting at the infinite buffer queue at the end of the time slot n. Thus, at most $(Y_n - m)I_{\{Q_{n-1}>L\}}$ overflown packets can be emptied at time n. Therefore, whenever $Q_i \leq L$

$$\sum_{n=1}^{i} X_n^Q \leq \sum_{n=1}^{i} (Y_n - m) I_{\{Q_{n-1}>L\}}. \tag{5}$$

Now let $\{i_k\}$ be the monotone increasing sequence of indices for which $Q_{i_k} \leq L$. Then, by stability, $i_k \to \infty$ almost surely. Thus, combining (4) and (5), we have

$$\begin{aligned}
\text{PLR} &= \lim_{i\to\infty} \frac{\sum_{n=1}^{i} X_n^S}{\sum_{n=1}^{i} X_n} \\
&\leq \lim_{i\to\infty} \frac{\sum_{n=1}^{i} X_n^Q}{\sum_{n=1}^{i} X_n} = \lim_{k\to\infty} \frac{\sum_{n=1}^{i_k} X_n^Q}{\sum_{n=1}^{i_k} X_n} \\
&\leq \lim_{k\to\infty} \frac{\sum_{n=1}^{i_k} (Y_n - m) I_{\{Q_{n-1}>L\}}}{\sum_{n=1}^{i_k} X_n} \\
&= \lim_{k\to\infty} \frac{\frac{1}{i_k}\sum_{n=1}^{i_k} (Y_n - m) I_{\{Q_{n-1}>L\}}}{\frac{1}{i_k}\sum_{n=1}^{i_k} X_n} \\
&= \lim_{n\to\infty} \frac{\mathbf{E}((Y_n - m) I_{\{Q_{n-1}>L\}})}{\mathbf{E}X_n} \\
&= \lim_{n\to\infty} \frac{\mathbf{E}(Y_n - m)\mathbf{P}(Q_{n-1} > L)}{\mathbf{E}X_n} \\
&= \frac{(\mathbf{E}Y_1 - m)\mathbf{P}(Q > L)}{\mathbf{E}X_1}
\end{aligned}$$

almost surely, where we used the ergodicity of $\{X_n, Y_n\}$ and the independence of Y_n and Q_{n-1}. This completes the proof. □

III. Bounds for Real Traffic

In this section the bound of Theorem 2 is applied for different types of real traffic traces. A server with finite and infinite buffers driven by video traces is investigated and the ratio of PLR/$\mathbf{P}(Q > L)$ is compared to the calculated bounding constant. Examples are given for cases where the packet loss exceeds the tail probability and where it is smaller by an order of magnitude.

The video traces used in this paper are captured and encoded by Fitzek and Reisslein [18]. MPEG4 compression method was used for encoding, which involves reduction of both spatial and temporal redundancy. The captured video files were compressed according to variable bit rate (VBR) coding scheme.

Example 1 In this example the frame level version of an MPEG4 trace is used as the input process. There are three types of frames in this trace: I, P and B, which considerably differ in their average sizes, i.e., an I frame is typically a couple of times larger than P and B frames. Due to the MPEG coding technique, the frames are arranged in a deterministic periodic sequence (in this case "IBBPBBPBBPBB"), which is called Group of Pictures (GOP). This coding scheme leads to a highly bursty traffic on the frame level.

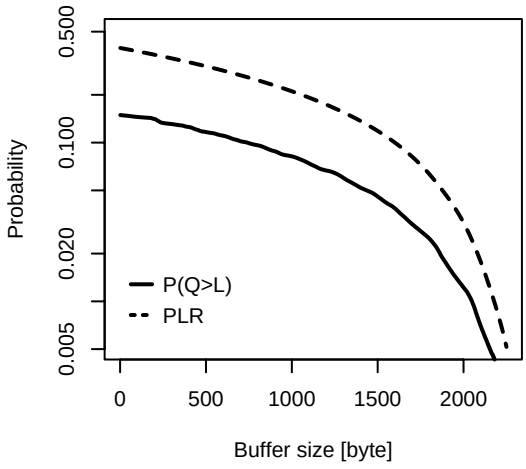

Fig. 1: The PLR and the tail probability for a bursty MPEG source.

In the simulation process the large I frames of 5 Kbytes were fragmented into smaller packets of 300-400 bytes. The average rate is 450 bytes/frame. In order to keep the frame delay in the order of 500 ms a service rate of 2600 bytes/frame slot were chosen. Figure 1 shows the obtained results for the tail probability $\mathbf{P}(Q > L)$ in the infinite buffer and true PLR. It can be seen that the packet loss exceeds the tail probabilities over a wide range of buffer sizes. From this simulation their ratio proves to be ~ 2.5, while the calculated constant is $\frac{\mathbf{E}Y}{\mathbf{E}X} = \frac{2600}{450} \approx 5.7$. In this case, the approximation of the PLR with $\mathbf{P}(Q > L)$ underestimates the actual loss, while the calculated multiplier provides an upper bound that overestimates the real packet loss only by a factor of 2.

Example 2 In practice the input process is a superposition of the traffic offered by different sources. Thus, input traffic generally does not contain such high peaks as those in the previous example. In these cases, the tail probability usually overestimates the packet loss. This scenario was investigated in the second simulation, where the input was an aggregate of 15 different VBR coded MPEG4 traces. The simulation was performed on the GOP level to eliminate the burstiness due to the deterministic MPEG structure. This requires the smoothing of the 12 frame periods ($\sim$ 500 ms). The resulting aggregate traffic has 1.12 Mbyte/s average rate, while the service rate was set to 1.3 Mbyte/s. The obtained results are shown in Figure 2. It can be clearly seen that there is an order of magnitude difference between the tail probability and true PLR. If the minimum rate is known (720 Kbyte/s in this case), a slightly better – but strictly conservative – approximation can be given with the calculated bound $\frac{\mathbf{E}X - \min X}{\mathbf{E}Y}\mathbf{P}(Q > L) \approx .51\,\mathbf{P}(Q > L)$.

IV. Packet Loss Estimation

When the actual packet loss is much below the queue tail probability, Theorem 2 cannot be used for approximating the actual packet

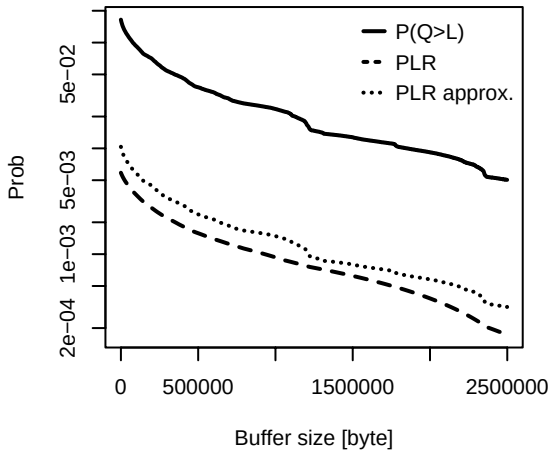

Fig. 2: The the tail probability, PLR and its approximation for an aggregate traffic input.

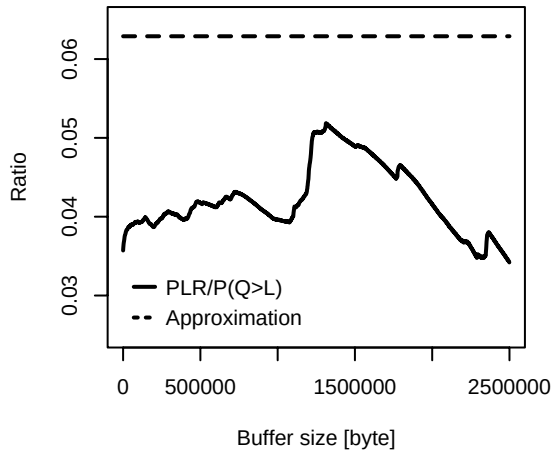

Fig. 3: The ratio of the queue tail to the packet loss of Example 2 for different buffer sizes.

loss. If the minimum rate is not known, the constant multiplier is always greater than one. On the other hand, if the utilization is high, the constant is close to one, which already provides a basis for using the infinite buffer tail probability as a conservative estimate in such cases.

However, it is possible to give a better approximation for packet loss with a heuristic argument based on the proof of Theorem 2, as follows. If in time slot n the finite buffer overflows, i.e., $S_n = L$, the number of arriving packets is X_n and at most $(X_n - Y_n)^+$ of them is lost. Then an upper bound can be given to the number of lost packets up to time i. For all i we have

$$\sum_{n=1}^{i} X_n^S \le \sum_{n=1}^{i} (X_n - Y_n)^+ I_{\{S_n = L\}} \le \sum_{n=1}^{i} (X_n - Y_n)^+ I_{\{Q_n \ge L\}}.$$

Then the PLR is bounded by

$$\begin{aligned} \text{PLR} &\le \frac{\mathbf{E}\{(X_n - Y_n)^+ I_{\{Q_n \ge L\}}\}}{\mathbf{E}X_1} \\ &= \frac{\mathbf{E}\{(X_n - Y_n)^+ | Q_n \ge L\} \mathbf{P}(Q_n \ge L)}{\mathbf{E}X_1}. \end{aligned}$$

For constant service rate s the conditional expectation can be approximated by $\mathbf{E}\{X_n - s | X_n > s\}$ since in general there is a high correlation between the events $\{X_n > s\}$ and $\{Q_n \ge L\}$. On the other hand, $\mathbf{P}(Q \ge L) \approx \mathbf{P}(Q > L)$, and so the PLR can be simply approximated as

$$\text{PLR} \approx \frac{\mathbf{E}\{X_n | X_n > s\} - s}{\mathbf{E}X_1} \mathbf{P}(Q > L).$$

Unfortunately, this is not an upper bound. However, when applying to real traces, it was always conservative for both single and multiplexed input traces. In certain cases it improves the estimation of the packet loss with the tail probability by an order of magnitude, as can be seen in Figure 2. The ratio $\text{PLR}/\mathbf{P}(Q > L)$ and its approximation for Example 2 for different buffer sizes is shown in Figure 3. Since the approximation of this ratio is independent of the buffer size, it is usually expected to overestimate the highest ratio, as depicted in the figure. Several simulations performed for other VBR MPEG and constant bit rate (CBR) coded H.263 traces yielded similar results.

V. Conclusion

In this paper we considered the connection between the packet loss in a finite buffer and the tail probability of the corresponding infinite buffer queue. We showed that the PLR can significantly differ from $\mathbf{P}(Q > L)$ in both directions. An upper bound was given on their ratio, which, in addition, can easily be calculated since it assumes only the knowledge of the average rate. An improved version can be obtained with the use of the minimum rate. In case of high utilization the bound is close to 1, which suggests that the tail probability is indeed a conservative estimate for the PLR. As simulations with real traffic traces showed, the bound is fairly tight for certain traffic patterns. However, if the packet loss is much smaller than the tail probability, the bound is usually loose. Therefore, an approximation on the ratio $\text{PLR}/\mathbf{P}(Q > L)$ is proposed, which turned out to be in the same order as the real ratio for all investigated scenarios performed with real traffic patterns.

Acknowledgments

The authors wish to thank Prof. László Györfi for helpful comments. They would also like to thank the Telecommunication Networks Group at the Technical University of Berlin for making their video traces publicly available.

References

[1] I. Norros, "A storage model with self-similar input," *Queueing Systems*, vol. 16, pp. 387–396, 1994.

[2] N. G. Duffield, J. T. Lewis, N. O'Connel, R. Russel, and F. Foomey, "Entropy of atm traffic streams: a tool for estimating qos parameters," *IEEE Journal on Selected Areas in Communications*, vol. 13, pp. 981–989, 1995.

[3] S. Ramaswamy, T. Ono-Tesfaye, W. Armstrong, and P. Gburzynski, "Equivalent bandwidth characterization for real-time cac in atm networks." Preprint.

[4] M. Krunz and A. M. Ramasamy, "The correlation structure for a class of scene-based video models and its impact on the dimensioning of video buffers," *IEEE Trans. Multimedia*, vol. 2, pp. 27 – 36, 2000.

[5] F. P. Kelly, "Notes on effective bandwidth," in *Stochastic Networks: Theory and Applications* (F. P. Kelly, S. Zachary, and I. B. Ziedins, eds.), vol. 4, Royal Statistical Society Lecture Notes Series, 1995.

[6] A. Weiss, "An introduction to large deviations for communication networks," *IEEE Journal on Selected Areas in Communications*, vol. 13, pp. 938–952, 1995.

[7] P. W. Glynn and W. Whitt, "Logarithmic asymtotics for steady-state tail probabilities in a single-server queue," *Journal of Applied Probability*, vol. 31A, pp. 131–156, 1994.

[8] N. G. Duffield, "Exponential bounds for queues with Markovian arrivals," *Queueing Systems*, vol. 17, pp. 413–430, 1994.

[9] J. Guibert, "Overflow probability upper bound in fluid queues with general on/off sources," *Journal of Applied Probability*, vol. 31, no. 3, pp. 1134–1139, 1994.

[10] N. G. Duffield and N. O'Connell, "Large deviations and overflow probabilities for the general single server queue, with applications," in *Mathematical Proceedings of the Cambridge Philosophical Society*, pp. 363–374, 1995.

[11] N. Duffield, M. Huggard, R. Russel, F. Toomey, and C. Walsh, "Fast bounds for ATM quality of service parameters," in *Proceedings of the 12th IEE UK Teletraffic Symposium*, (Old Windsor), 1995.

[12] D. D. Botvich and N. G. Duffield, "Large deviations, economies of scale, and the shape of the loss curve in large multiplexers," *Queueing Systems*, vol. 20, pp. 293–320, 1995.

[13] C.-S. Chang and J. A. Thomas, "Large deviations approximation for fluid queues fed by a large number of on/off sources," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 6, pp. 1017–1027, 1995.

[14] C. Courcoubetis and R. Weber, "Buffer overflow asymptotics for a switch handling many traffic sources," *Journal of Applied Probability*, vol. 33, no. 3, pp. 886–903, 1996.

[15] N. Likhanov and R. Mazumdar, "Cell loss asymptotics for buffers fed with a large number of independent stationary sources," *Journal of Applied Probability*, vol. 36, March 1999.

[16] R. M. Loynes, "The stability of a queue with non-independent inter-arrival and service times," *Proc. Cam. Phil. Soc*, vol. 58, pp. 497 – 520, July 1962.

[17] B. McGurk and R.Russell, "Simple bounds for queues fed by markovian sources: a tool for performance evaluation," in *Computer Performance Evaluation, Modeling Techniques and Tools*, Lecture Notes in Computer Science 1245, Springer, 1997.

[18] F. H. P. Fitzek and M. Reisslein, "MPEG-4 and H.263 video traces for network performance evaluation." *Technical Report TKN-00-06*, 2000.