Real Time Cell Loss Ratio Estimation for Bursty and Self-Similar Traffic—(Extended Abstract)

A. Vidács, Sz. Malomsoky*
High Speed Networks Laboratories
Dept. of Telecommunications and Telematics
Technical University of Budapest
Sztoczek u. 2., Budapest, H-1126, Hungary
E-mail: {vidacs,malomsok}@ttt-atm.ttt.bme.hu

H. Saito
NTT Multimedia Networks Laboratories
3-9-11, Midori-cho, Musashino-shi, Tokyo, 180, Japan.
E-mail: saito@hashi.tnl.ntt.co.jp

1 Introduction

The asymptotics of queue length distribution in the regime of large buffers are exponential and can be characterized by two parameters, the asymptotic constant and asymptotic decay rate [5]:

$$P(Q \geq k) = \beta e^{-\eta k}. \tag{1}$$

This result is very general, provided that the arrival process does not possess long range dependence. As for the long range dependent case (or equivalently, when the increment of the traffic process is self-similar), the CLR decreases with the buffer size sub-exponentially. Recent results from the literature show that for self-similar traffic the tail of the stationary queue length distribution is Weibullian [4, 8]:

$$P(Q \geq k) = \beta e^{-\eta k^{\gamma}}, \quad \gamma \in (0, 1]. \tag{2}$$

Using these results, this paper proposes an algorithm for estimating the CLR in real time based on buffer measurements, which works for both the long range and the short range dependent case.

2 Large deviations results for cell loss ratio asymptotics

Consider a single server queue that has an infinite buffer with stationary and ergodic cell arrivals from a single source. Let A_t denote the number of arrivals in $(0,t]$, and C is the number of cells that can be served in unit time. According to [4], under suitable hypotheses the logarithm of the queue length distribution has the following asymptotics:

$$\lim_{k \to \infty} h_k^{-1} \log P(Q \geq k) = -\eta, \tag{3}$$

if there exist the appropriate scaling function h_t. For example, a model called fractional Brownian motion has been proposed in the literature (e.g., in [8]) to account for self-similarity and long range

*This work has been done while working at NTT Multimedia Networks Laboratories

dependence. For that, $h_t = t^{2-2H}$, and from Eq.(3) we get

$$p_k \stackrel{\text{def}}{=} P(Q \geq k) \approx \beta e^{-\eta k^{2-2H}}. \tag{4}$$

In this case, it is rather complex to calculate the asymptotic constant β and asymptotic decay rate η.

For linear scalings ($h_t \equiv t$), Eq.(1) is valid. With the help of the cumulant generating function of A_t, calculated as

$$M_A(\theta) = \lim_{t\to\infty} M_A(t,\theta) = \lim_{t\to\infty} t^{-1} \log E e^{\theta A_t}, \tag{5}$$

the asymptotic constant β and asymptotic decay rate η can be derived as [4]

$$\eta = \sup\left\{\theta : \frac{M_A(\theta)}{\theta} \leq C\right\}, \qquad \log\beta = \lim_{t\to\infty}\left\{tM_A(t,\eta) - C\eta t\right\}. \tag{6}$$

According to Chang [2], provided that the input process satisfies some mild technical conditions, η is given by the unique solution of $M_A(\eta)/\eta = C$. $M_A(\eta)/\eta$ can be expanded by η (Cumulant Expansion), and from the reversion of that expansion we get [3]:

$$\eta = \sum_{k=1}^{\infty} c_k^{(2)} \left(\frac{1-\rho}{\rho}\right)^k, \qquad \log\beta = \sum_{k=1}^{\infty} c_k^{(1)} \left(\frac{1-\rho}{\rho}\right)^k, \tag{7}$$

where $\{c_i^{(1)}\}$ and $\{c_i^{(2)}\}$ are given in terms of the cumulants of A_t. It is known that the CLR of a single server queue with a finite buffer size, K, is less than p_K [4]. Once we know the coefficients, $\{c_1^{(1)}\}$ and $\{c_1^{(2)}\}$, these expansions enable us to estimate p_K from Eq.(1) under different utilizations. Choudhury and Whitt showed in [3], that taking only the first term in Eq.(7) is a good approximation as $\rho \to 1$ (Heavy Traffic Expansion). If we take only the first terms of the sums in Eq.(7), from Eq.(1) we get

$$\log p_K \approx \frac{1}{d}\left(\frac{1-\rho}{\rho}\right), \quad d = \frac{1}{c_1^{(1)} - c_1^{(2)} K}. \tag{8}$$

Once we obtain the parameter d somehow (see [7]), Eq.(8) enable us to estimate the CLR under different utilizations.

3 Real time buffer measurement

Shioda and Saito showed in [9], that for the case when Eq.(1) is valid, by monitoring the buffer occupancy levels at two different thresholds, the parameters β and η can be estimated.

When the tail asymptotics of the queue length distributions are Weibullian [4, 8] (see Eq.(2)), since $\log p_K$ is the function of K^γ instead of K, we need three different thresholds to monitor the queue lenght instead of two. It is easy to see, that for $\gamma = 1$ Eq.(2) gives back the result for the short range dependent case (Eq.(1)). Hence, using the three-point measurement method, the same calculation can be used for short range and long range dependent traffic, which is useful for implementations.

4 Simulation

In our simulation experiments, three cases were examined. First, the aggregated traffic offered to the VP was short range dependent with low burstiness ($\beta \simeq 1$). Second, the offered traffic was short range dependent, but more bursty, implying $\beta << 1$. Finally, the traffic was long range dependent.

In the simulations, cells from N identical but independent on-off sources arrived at the output buffer of 128 cells dedicated to the VP. As for the short range dependent input, 10 sources were multiplexed, and cells were offered from each source at rate of 2 Mbps during the exponentially distributed on periods with mean 0.543 ms in the 'non-bursty' case, and 5.43 ms in the 'bursty case'. The off periods

were also exponentially distributed, and the average rate of each source was set to 0.2 Mbps. At three different thresholds (5, 10, 20), the frequencies in which the number of cells in the buffer was greater than or equal to the thresholds were measured every 20 minutes. The VP bandwidth was choosen such that the estimated CLR was approximatelly 10^{-4} (2.35 Mbps for the 'non-bursty' case, and 6.05 Mbps for the 'bursty' case).

Figure 1 and Figure 3 show the results: the measured CLR is plotted together with the estimated values using Weibullian and exponential approximations for the queue length distributions. In the case of 'non-bursty sources', the parameter γ in the Weibullian distribution (Eq.(2)) becomes $\simeq 1$, and the fitted distribution is close to the real one (Figure 2). Our method works well, and gives only slightly higher CLR estimates as the exponential one.

In the case of 'bursty sources' (Figure 4)—although the tail asymptotics for the distribution *are* exponential,— monitoring the queue at two points and estimating the queue length distribution by an exponential distribution is not appropriate here, the CLR is underestimated. The three-point measurement method works better in this case (Figure 3).

To introduce long range dependence into the on-off model, assume a Pareto-type distribution with heavy tail for the activity periods T_{on} [1]. In the limit of a large number of such sources and high load, it is shown in [6] that the aggregated traffic converges to an exactly self-similar Gaussian process, and in [1] Brichet *et al.* showed that the tail of the queue length distribution is Weibullian. In our numerical example the number of sources was set to 100. The mean duration of the on period was 0.1 ms, and the parameters of the Pareto distribution were set with Hurst parameter $H = 0.7$. The VP bandwidth was 26 Mbps, and all the other parameters were kept as they had been before.

What we can see from the results (Figure 5) is that the difference between the exponential and Weibullian estimates is about four orders of magnitude. The measured CLR values are bounded by the two estimates, and are in the order of 10^{-5}. Figure 6 gives us the reason why the exponential fitting underestimates the CLR by 3–4 orders of magnitude. *Any* exponential estimate would underestimate the CLR. On the other hand, the Weibullian approximation fits the queue length distribution relatively well, and the estimate of p_K again turns out to be an upper bound on the actual CLR.

References

[1] F. Brichet, J. Roberts, A. Simonian and D. Veitch. Heavy traffic analysis of a storage model with long range dependent On/Off sources. *Queueing Systems*, 23:197–215, 1996.

[2] C.S. Chang. Stability, queue length, and delay of deterministic and stochastic queueing networks. *IEEE Trans. on Automatic Control*, 39(5):913–931, 1994

[3] G.L. Choudhury and W. Whitt. Heavy-traffic asymptotic expansion for the asymptotic decay rates in the BMAP/G/1 queue. *Commun. Statist.—Stochastic Models*, 10(2):453–498, 1994.

[4] N.G. Duffield and N. O'Connell. Large deviations and overflow probabilities for the general single-server queue, with applications. *Proc. Cam. Phil. Soc.*, 118:363–374, 1994.

[5] P.W. Glynn and W. Whitt. Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. *J. Appl. Prob.*, 31A:131–159, 1993.

[6] W.E. Leland, M.S. Taqqu, W. Willinger and D.V. Wilson. On the self-similar nature of Ethernet traffic (Extended version). *IEEE/ACM Trans. on Networking*, 2(1), 1994.

[7] Sz. Malomsoky, A. Vidács and H. Saito. Real time VP bandwidth control for long range dependent traffic. *submitted to IEICE Trans. on Networking*, 1997.

[8] I. Norros. A storage model with self-similar input. *Queueing Systems*, 16:387–396, 1994.

[9] S. Shioda and H. Saito. Real-time cell loss ratio estimation and its application to ATM traffic controls. *Proc. IEEE INFOCOM'97*, pages 1074–1081, Kobe, Japan, 1997.

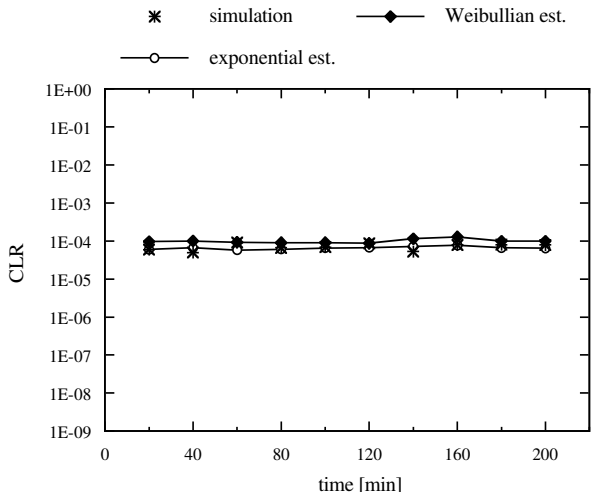

Figure 1: CLR estimates using the proposed measurement method (non-bursty case).

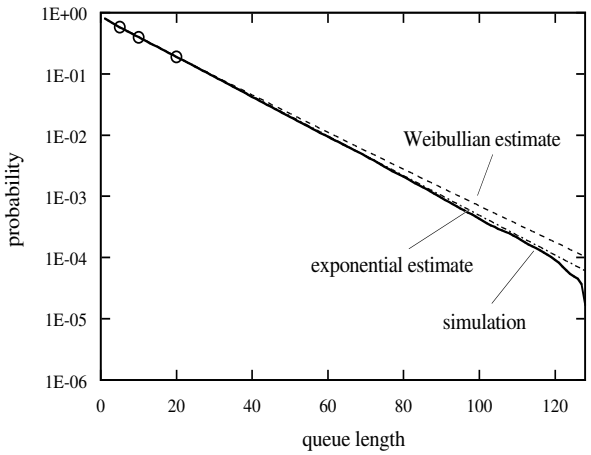

Figure 2: Queue length distribution for non-bursty on-off sources.

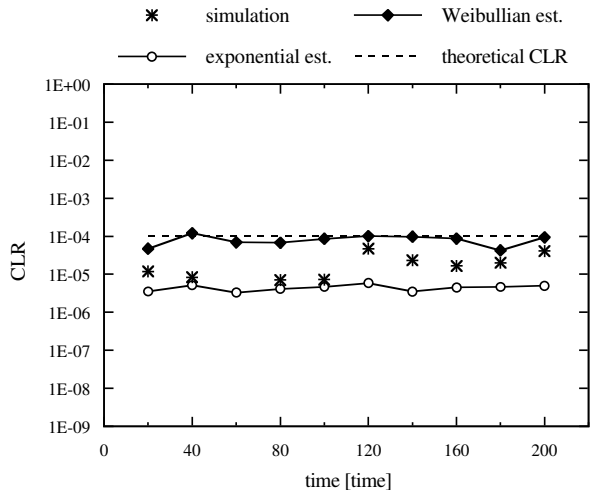

Figure 3: CLR estimates using the proposed measurement method (bursty on-off sources).

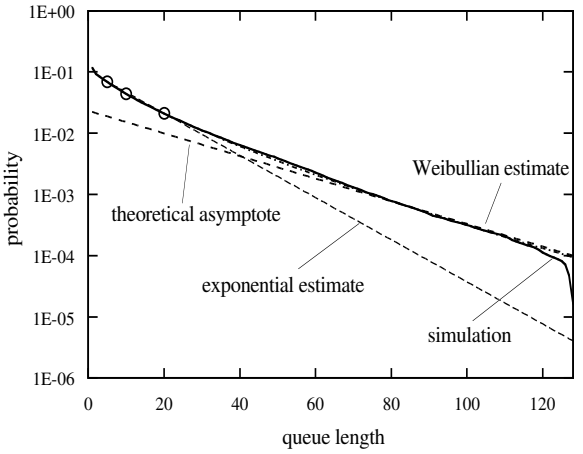

Figure 4: Queue length distribution for bursty on-off sources.

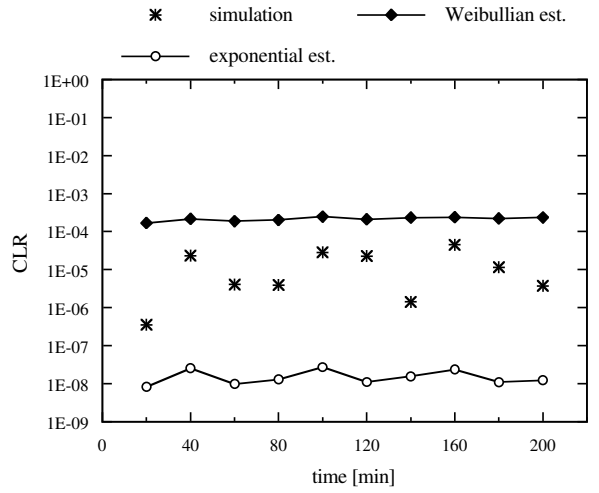

Figure 5: CLR estimates using the proposed method for Pareto-type sources.

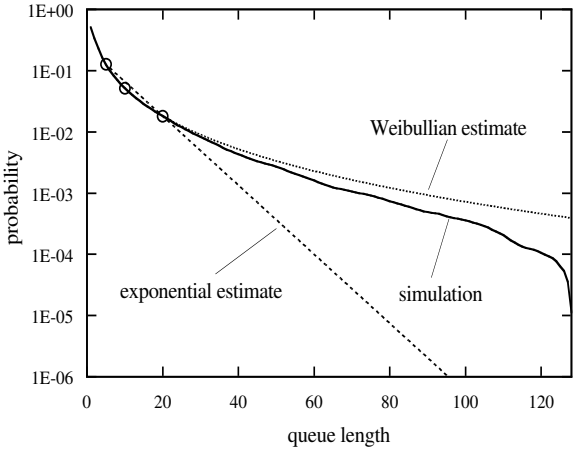

Figure 6: Queue length distribution for Pareto-type sources.