Predicting Quality of Service for Traffic with Long-Range Fluctuations [*]

N.G.Duffield[†]
Dublin City University
J.T.Lewis
Dublin Institute for Advanced Studies
Neil O'Connell
Trinity College, Dublin
Raymond Russell and Fergal Toomey
Dublin Institute for Advanced Studies

Abstract

We present the tail asymptotics in a queue serviced at constant rate and whose input process has long-range dependence, for example, fractional Brownian Motion (fBM) with Hurst parameter $H > 1/2$. Some heterogeneous superpositions of such sources are also treated. We examine the experimental basis for self-similar modelling. If fluctuations in traffic levels can be accounted for by variations in time-dependent external parameters, rather than statistically through parsimonious modelling by fBM, then Quality of Service predictions may be achieved by applications of techniques familiar for short-range dependent traffic.

1 Introduction

In a series of recent papers, Leland *et al* [17, 18] propose that Ethernet traffic can be modelled by self-similar possesses and possesses long-range dependence (LRD). Similar proposals have been made for Variable-Bit-Rate sources by Beran *et al* [2]. In the former case the proposal is justified using traffic observations that were taken from an Ethernet local area network at Bellcore Laboratories: we refer to these observations, which have been made publicly available, as the 'Bellcore data'. In this paper we discuss the implications of such models for the problem of estimating loss probabilities in networks, and their validity.

There has been much recent work on estimating rare-event probabilities in queueing networks: see for example [1, 3, 4, 6, 7, 8, 9, 12, 15]. For traffic which is stationary and possesses only short-range dependence (SRD), the tails of the distribution of the queue-length Q in a single server queue with deterministic service should satisfy

$$P[Q > b] \approx e^{-\delta b}, \tag{1}$$

for some positive constant δ that depends on the service rate at the queue and the statistical properties of the arrivals process. This formula has application in congestion and connection admittance control through the notion of effective bandwidth of a source. (See, for example, [4, 11, 13, 14, 22]). These furnish a linear inequality in the numbers of sources of various types which must be satisfied in order to maintain a given quality of service.

The asymptotic (1) holds in great generality, and in particular for any Markovian model. The proposal that LRD traffic models be used instead has been interpreted by some as meaning that the usual SRD models and methods of analysis based on (1) are no longer applicable. In Section 2 the question we address is: *should we expect (1) to hold for self-similar traffic that possesses long range dependence?* The answer is 'no': this follows from a theorem in [8]. In fact, self-similarity is not an issue here: it is the presence of long range dependence that destroys the property (1), which is replaced by

$$P[Q > b] \approx e^{-\delta b^{2(1-H)}} \quad \text{some } H > 1/2. \tag{2}$$

In section 3 we examine the consequences of this for performance prediction for superpositions of LRD sources. Here we find limited reassurance through a simple example of heterogeneous superpositions of LRD sources: Quality of Service (QoS) is maintained if the numbers of the various sources present satisfy a linear inequality.

In section 4 we examine the basis for LRD modelling. The central issue here is stationarity: unless a time-series is drawn from a stationary process it cannot be used to make predictions for times beyond its extent. Level-shifts within the data bring into question the assumption of an underlying stationarity LRD process.

One approach, which has been taken by Leland *et al*, is to regard self-similar modelling as a "parsimonious" scheme in which the statistics of fluctuations are accounted for in a stationary model: self-similarity becomes a phenomenological traffic model specified by a small number of parameters.

In this paper we discuss the implications for performance prediction of the alternative view. The Bellcore data sets exhibit level shifts; it is reasonable to suppose that between level shifts the data may be treated as stationary. This means that over appropriately short time scales we could use (1) as a basis for performance prediction. Prediction over longer time-scales will be possible provided level-shifts can be understood in terms of "deterministic" variations in time-dependent external parameters, such as the number of active connections.

[*] In: Proceedings of ICC'95, Seattle, 18-22 June 1995, pp473-477

[†] Author presenting paper, now at duffield@research.att.com

2 Queue asymptotics in the presence of long range dependence.

Suppose we have a stationary arrival process (X_k) with mean $EX_1 = \mu < \infty$ and a queue with deterministic service rate $s > \mu$. A sufficient condition for (1) to hold is that the scaled cumulant generating function, defined by

$$\lambda(\theta) := \lim_{n\to\infty} n^{-1} \log E e^{\theta \sum_{k=1}^{n} X_k}, \tag{3}$$

exists, is finite in some neighbourhood of the origin and differentiable on the interior of its effective domain. Then the arrivals process satisfies a large deviation principle with rate function given by the Legendre transform λ^* of λ:

$$\lambda^*(x) = \sup_{\theta}\{\theta x - \lambda(\theta)\}. \tag{4}$$

This can be restated informally by saying that the asymptotics of the arrival process distribution are

$$P[\sum_{k=1}^{n} X_k > nx] \approx e^{-n\lambda^*(x)}. \tag{5}$$

The relation between this and δ, the asymptotic decay rate for the tails of the queue-length distribution in (1), is given by

$$\delta = \inf_{c>0} c^{-1}\lambda^*(c+s). \tag{6}$$

An alternative representation of δ is

$$\delta = \sup\{\theta : \ \lambda(\theta) - \theta s \le 0\}. \tag{7}$$

Variants of this result have appeared in the literature: we refer the reader to [15] for a heuristic derivation, and to the recent papers of Glynn and Whitt [12] and Duffield and O'Connell [8] for proofs under very general conditions; further bibliographical details can be found in [6].

Now suppose that the arrivals process X has the property that for large m and n,

$$\sum_{k=1}^{nm}(X_k - \mu) \ \overset{\mathcal{D}}{\approx}\ n^H \sum_{k=1}^{m}(X_k - \mu), \tag{8}$$

for some $H \in [1/2, 1)$. Here $\overset{\mathcal{D}}{\approx}$ means 'approximately equal in distribution'. Then X is said to be *asymptotically self-similar with Hurst parameter* H; if $H > 1/2$ the process exhibits LRD. This is a criterion which Leland *et al.* [17, 18] have employed to infer self-similarity from observed time-series.

If (8) holds in a suitably rigorous sense and the relevant expectations are finite, then the scaled cumulant generating function defined by $\lambda(\theta) = \lim_{n\to\infty}\lambda_n(\theta)$ where

$$\lambda_n(\theta) := n^{-2(1-H)} \log E e^{\theta n^{1-2H} \sum_{k=1}^{n} X_k} \tag{9}$$

exists and is finite in some neighbourhood of the origin. It then follows from [8, Corollary 2.3] that, under mild regularity conditions, the tails of the corresponding queue-length distribution for a queue with constant service rate $s > \mu$ satisfy

$$\lim_{b\to\infty} b^{-2(1-H)} \log P[Q > b] = -\delta, \tag{10}$$

where

$$\delta = \inf_{c>0} c^{-2(1-H)}\lambda^*(c+s), \tag{11}$$

and λ^* is the Legendre transform of λ from (9). In particular, the log-probability of overflow is asymptotically linear in buffer-size if, and only if, $H = 1/2$; otherwise the decay is polynomial and depends on the value of H.

We now describe the effect of LRD alone, without assuming self-similarity. Note that (9) implies

$$\lim_{n\to\infty} n^{-2H} \ \mathrm{var}\left(\sum_{k=1}^{n} X_k\right) \in (0,\infty); \tag{12}$$

the existence of such a power law (for some $H > 1/2$) is often treated as a definition of LRD for finite variance processes. A more general statement is that there exists a divergent sequence u_n with $u_n/n \nearrow +\infty$ and that (12) holds with n^{-2H} replaced by u_n^{-1}; under additional hypotheses on the asymptotic behaviour of higher order moments (see, for example, [5, pp253–]) this yields a large deviation principle for X with scaling coefficients $v_n = n^2 u_n^{-1}$ and, assuming the limit

$$g(c) := \lim_{n\to\infty} \frac{v_{n/c}}{v_n} \tag{13}$$

exists for each $c > 0$, it follows from [8, Theorems 2.1 and 2.2] that

$$\lim_{b\to\infty} \frac{1}{v_b} \log P[Q > b] = -\delta, \tag{14}$$

where

$$\delta = \inf_{c>0} g(c)\lambda^*(c+s), \tag{15}$$

and λ^* is the Legendre transform of the scaled cumulant generating function

$$\lambda(\theta) := \lim_{n\to\infty} v_n^{-1} \log E e^{v_n \theta \sum_{k=1}^{n} X_k/n}. \tag{16}$$

If we thus take the existence of such a sequence u_n (with $u_n/n \nearrow +\infty$) as a working definition of LRD, we conclude that we should not expect (1) to hold in the presence of LRD; the actual behaviour is predicted by (14). We note that a large deviation lower bound for fBM arrivals (i.e. replacing $=$ by $\ge$ in (15)) has been obtained by Norros [21].

3 Queue-tail asymptotics for superposed arrivals.

Let Q^L denote the queue length due to a superposition of L identical sources, service at rate sL. When the v_n is a power-law n^{2-2H} (i.e. $u_n = n^{2H}$) for $H \in [1/2, 1)$, under very general conditions it follows that (see [7])

$$\lim_{b\to\infty} \frac{1}{v_b} \log P[Q^L > b] = -\delta L^{2H-1}, \tag{17}$$

with δ as in (15), using the cgf λ for a single source.

Note that in the absence of LRD, $H = 1/2$, $v_b = b$ and the right-hand side of (17) reduces to $-\delta$. Thus the queue-tail asymptotics (1) are invariant under L-fold superposition, provided the service rate is scaled proportionately. This property remains true for heterogeneous superpositions, and is the basis of the effective bandwidth approximation. (See [4, 11, 13, 14, 22] for development and applications to connection admission control).

Briefly, consider superpositions of numbers L_i of SRD arrivals of type i each with scaled cumulant generating function λ_i. From (1) and (7) it follows that a criterion for a loss ratio no worse than $e^{-\theta b}$ to be achieved in buffer size b served at rate s is

$$\sum_i L_i \lambda_i(\theta) - s\theta \leq 0. \tag{18}$$

This provides a connection acceptance boundary which is affine in the L_i.

In the presence of LRD, the situation is more complex. Under mild technical assumptions, then [7, Theorem 3] (15) is equivalent to

$$\delta = \sup\{\theta : \ \Lambda(\theta) \leq 0\}, \tag{19}$$

where Λ is the Legendre transform of the function $x \mapsto \lambda^*(x^{1/(2-2H)} + s)$ (defined for $x \geq 0$). From (2) (or (14)) it follows that a criterion for a loss ratio of no worse than $e^{-\theta b^{2(1-H)}}$ to be achieved in a buffer of size b served at rate s is

$$\Lambda(\theta) \leq 0 \tag{20}$$

In the case that the arrivals are a fractional Brownian motion with Hurst parameter $H \in (1/2, 1)$ and hence variance $\sigma^2 t^{2H}$, then

$$\lambda(\theta) = \frac{\sigma^2}{2}\theta^2 \qquad \text{and hence} \qquad \lambda^*(\theta) = \frac{1}{2\sigma^2}x^2, \tag{21}$$

whence, after a straightforward calculation, (19) becomes

$$\theta\sigma^2 \leq \delta_H := \frac{1}{2}\left(\frac{s}{H}\right)^{2H}(1-H)^{-2(1-H)}. \tag{22}$$

In the special case that the fBM is a mildly heterogeneous superposition of L_i fBM's of (identical) Hurst parameter $H \in (1/2, 1)$ and variance $\sigma_i^2 t^{2H}$ then the connection acceptance criterion is again affine in the L_i, becoming

$$\theta \sum_i L_i \sigma_i^2 \leq \delta_H. \tag{23}$$

The extent to which such behaviour holds more generally has not been determined. Note that for $H > 1/2$, δ_H is not linear in s. Finer asymptotic properties of the tail of the queue-length distribution are considered in [3, 7].

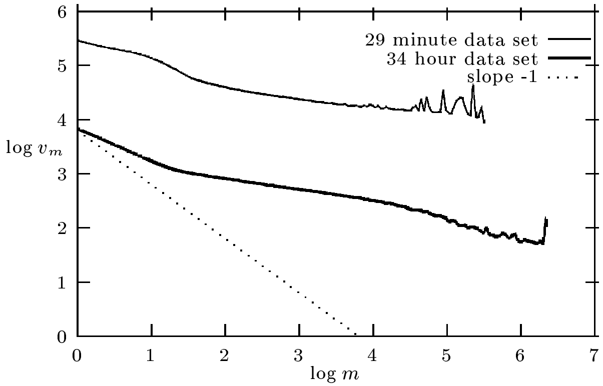

Figure 1: Variance-time plots for two of the Bellcore data sets.

4 Long range dependence: evidence and alternatives.

Having established the tail-asymptotics for queues with LRD sources, we shift the focus to the experimental observations which led to the proposal of the model. Leland *et al.* [17, 18] have analyzed Ethernet traffic measurements taken from local area networks at Bellcore Morristown Research and Engineering Center between August 1989 and February 1992. In these papers they conclude that the traffic is statistically self-similar and exhibits LRD. A variety of statistical tests were performed, including the inspection of 'variance-time plots'. We will concentrate on the use of variance-time plots as a detector of LRD.

Denote by $X_1, \ldots, X_n$ the traffic measurements over a period of observation. A variance-time plot can be produced as follows. Begin by computing, for each m, an aggregated sequence

$$X_k^{(m)} := \frac{1}{m}\sum_{j=(k-1)m+1}^{km} X_j \ , \qquad k = 1, 2, 3, \ldots \tag{24}$$

Then for each m, compute the sample variance v_m of the sequence $X_1^{(m)}, X_2^{(m)}, \ldots$, and plot $\log v_m$ against $\log m$. If the observations are taken from a stationary sequence that does not exhibit LRD we expect on the basis of (12) to observe, given sufficient data, an asymptotic slope of -1; on the other hand, for a stationary sequence with LRD, we expect the observed slope to be strictly greater than -1 for large values of m (a self-similar process with Hurst parameter H will produce an asymptotic slope of $-2(1 - H)$). Leland *et al.* consistently observed slopes that were greater than -1. The variance-time plots for two of their data sets are shown in Figure 1.

One of these data sets is about 30 hours in length; the other is 29 minutes long. Let us take a closer look at the 30-hour data set. Its activity is displayed in Figure 2: we have aggregated the data and displayed number of bytes observed in each 240-second time interval. Clearly, there are varying mean levels of activity on this time-scale. The activity of the 29-minute data set is displayed in Figure 3,

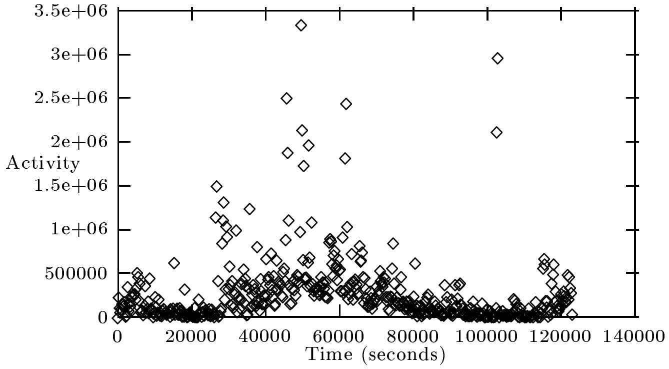

Figure 2: Activity plot for the 30-hour data set, aggregated over 240-second time intervals.

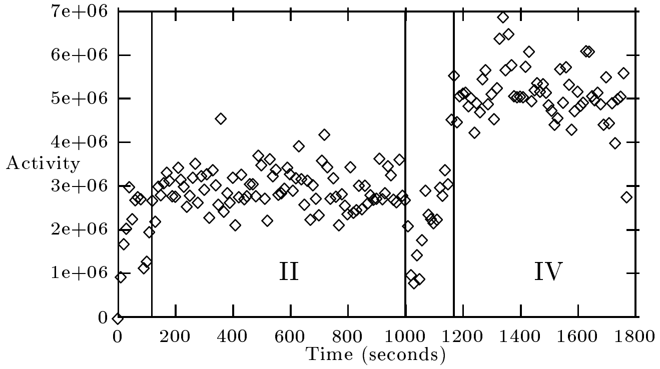

Figure 3: Activity plot for the 29 minute data set, aggregated over 10-second time intervals.

aggregated over 10-second time intervals. There is clearly a level-shift here, between regions labelled II and IV.

As discussed by Beran *et al* [2] and Leland *et al* [17, 18], the presence of large-scale fluctuations lead us to a fundamental dichotomy, which is an old one within statistics. Are we to regard such sequences as embedded in a stationary sequence that possesses LRD, or, should we regard the fluctuations as indicating non-stationarity within the data? For example, the fluctuations in the 30-hour sample may occur because time of day is an important factor in determining levels of activity. Note that 'time of day' is a periodic process. It is clearly important to understand the possible 'deterministic' origins of level shifts within a given data set.

In order to progress, we must ask what can be concluded about a data sequence which yields a high estimated Hurst parameter. Only that there are fluctuations in the data at the time-scale over which the data is observed. Such fluctuations are consistent with a LRD model. However, by the very nature of these fluctuations there is not enough data to say anything about future fluctuations at this time scale, or at even longer time scales. In particular we cannot infer self-similarity i.e. LRD at all longer time-scales. In the field of hydrology, in response to the use of self-similar processes by Mandelbrot and Wallis [20], similar objections have been forcefully stated by Klemes [16, p667]:

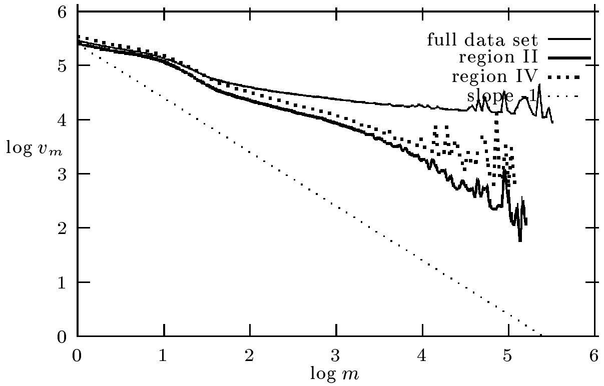

Figure 4: Variance-time plots for regions II and IV in the 29-minute data set.

> "... for if a historic process really does have time-varying parameters, then it is pointless to attempt large sample predictions in the first place, not to speak about the usefulness of predictions based on stationary models."

Where does this leave us in the attempt to solve our original problem of determining loss probabilities? Returning to the 29 minute data-set of Figure 3, the regions labelled II and IV may be treated as stationary. We have created variance-time plots for each of these regions (Figure 4).

The slopes at large aggregation levels are quite close to -1, indicating SRD within each region. For forecasting loss ratios this suggests that (1) rather than (2) can be applied successfully provided that

- the periods of stationarity are longer than the range of dependence within them; and
- the time-scales over which loss occurs are shorter than the period of stationarity.

The first is a property of the traffic stream alone while the second depends of buffer dimensions. From sample-path large-deviation arguments [5] we know that for SRD the time to overflow is asymptotically proportional to the buffer size b for large b.

The Bellcore data is very challenging from the point of view of forecasting. We do not know the origin of the observed level shifts. If these can be linked to external time-dependent parameters (for example, the number of active virtual circuits) they can effectively be removed; this opens the door to using (1) as a tool for prediction on longer time-scales.

5 Conclusion

In the introduction we referred to self-similar traffic models as being phenomenological models based on assumptions concerning the statistics of fluctuation. We conclude by pointing out that both LRD *and* SRD models have a common phenomenological framework provided by the theory of large deviations. Given the assumptions of either case, one can make predictions of QoS through (2) or (1) as appropriate, based only on the large deviation properties of traffic as expressed through its scaled cumulant generating function λ, rather than on the fine details of an underlying model. Elsewhere, [10], we describe a method of estimating δ directly from a time-series of SRD traffic, thus bypassing modelling.

Acknowledgement

We thank Walter Willinger for making the Bellcore data available to us, and for an informative correspondence on self-similar modelling.

References

[1] David Aldous (1989). *Probability Approximations via the Poisson Clumping Heuristic.* Applied Mathematical Sciences 77, Springer-Verlag.

[2] J. Beran, R. Sherman, M.S. Taqqu and W. Willinger (1993). Variable-bit-rate video traffic and long range dependence. To appear in *IEEE Trans. Comm.*

[3] D.D. Botvich and N.G. Duffield (1995). Large deviations, the shape of the loss curve, and economies of scale in large multiplexers. *Queueing Systems,* **20** 293-320

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

[5] Amir Dembo and Ofer Zeitouni (1993). *Large Deviation Techniques and Applications.* Jones and Bartlett, Boston-London.

[6] G. de Veciana, C. Courcoubetis and J. Walrand (1993). Decoupling bandwidths for networks: a decomposition approach to resource management. Memorandum No. UCB/ERL M93/50, University of California.

[7] N.G. Duffield (1994). Economies of scale in queues with sources having power-law large deviation scalings. Preprint DIAS-APG-94-27. To appear in *J. Appl. Prob., 1996*

[8] N.G. Duffield and Neil O'Connell (1995). Large deviations and overflow probabilities for the general single-server queue, with applications. *Math. Proc. Cam. Phil. Soc.* **118**, 363-374.

[9] N.G. Duffield and Neil O'Connell (1994). Large deviations for arrivals, departures, and overflow in some queues of interacting traffic. *Proceedings of the* 11^th*h IEE Teletraffic Symposium,* Cambridge, March 1994.

[10] N.G. Duffield, J.T. Lewis, Neil O'Connell, Raymond Russell and Fergal Toomey (1995). Entropy of ATM traffic streams: a tool for estimating QoS parameters. *IEEE J. Selected Areas in Commun.* **13** 981-990

[11] R.J. Gibbens and P.J. Hunt (1991). Effective Bandwidths for the multi-type UAS channel *Queueing Systems,* **9** 17–28

[12] Peter W. Glynn and Ward Whitt (1994). Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. In: *Studies in Applied Probability* Eds. J. Galambos and J. Gani, *Journal of Applied Probability, Special Volume* **31A** 131–159

[13] J.Y. Hui (1988). Resource allocation for broadband networks. *IEEE J. Selected Areas in Commun.* **6** 1598–1608

[14] F.P. Kelly (1991). Effective bandwidths at multi-type queues. *Queueing Systems,* **9** 5–16

[15] G. Kesidis, J. Walrand and C.S. Chang (1993). Effective bandwidths for multiclass Markov fluids and other ATM Sources. *IEEE/ACM Trans. Networking* **1** 424–428

[16] V. Klemes (1974). The Hurst phenomenon: a puzzle? *Water Resour. Res.,* **10** 675–688

[17] W.E. Leland, M.S. Taqqu, W. Willinger and D.V. Wilson (1993). Statistical analysis of high time-resolution Ethernet LAN traffic measurements. *Proceedings of INTERFACE*, 1993

[18] Will E. Leland, Murad S. Taqqu, Walter Willinger and Daniel V. Wilson (1993). On the self-similar nature of Ethernet traffic. *ACM SIGCOMM Computer Communications Review* **23** 183-193

[19] B.B. Mandelbrot and J.W. Van Ness (1968). Fractional Brownian motions, fractional noises and applications. *SIAM Review*, **10** 422–437

[20] B.B. Mandelbrot and J.R. Wallis (1968). Noah, Joseph, and operational hydrology. *Water Resour. Res.,* **4** 909–918

[21] Ilkka Norros (1994). A storage model with self-similar input. *Queueing Systems,* **16** 387–396

[22] W. Whitt (1993). Tail probabilities with statistical multiplexing and effective bandwidths in multi-class queues. *Telecommunications Systems.* **2** 71–107