On the output process from a finite buffer with long range dependent input

Ilze Ziedins
Department of Statistics
The University of Auckland
Private Bag 92019
Auckland, New Zealand

Abstract

This paper studies the output from a finite buffer with long range dependent input. The particular question of interest is whether the output from the buffer is always long range dependent. We test this using two different kinds of input to the buffer. The first are traces collected by the WAND group at the University of Waikato, and the second are simulated inputs. We simulate input data using both Pareto inter-arrival times (which are heavy-tailed, and therefore long range dependent) and superpositions of processes with Pareto inter-arrival times. Our conclusion is that for sufficiently high arrival rates (and therefore high loss probabilities), the output process no longer shows evidence of long range dependence on the time scales which were considered in the simulations. The point at which long range dependence is no longer evident varies, depending on the characteristics of the input process, but in some examples it was no longer observed when the load on the server, ρ, was less than 1.1.

1 Introduction

It has been known for some time that traffic in networks often exhibits features that are not consistent with traditional Markov and short range dependent models, such as the Poisson process. This has been observed in various settings, both in Local Area Networks and Wide Area Networks, and for various types of traffic (see for instance, [21],[23],[27]). Willinger *et al.* [28] give an excellent bibliographic guide to earlier works on self-similar traffic and performance modelling for modern high-speed networks.

Long range dependent traffic has major consequences for controls and sizing of buffers, and many authors have now studied the effect of long range dependent input on queueing behaviour. Much of this work has concentrated on queues with infinite capacity, addressing such questions as the distribution of the queue length, waiting time in the queue, and the time at which the queue size first exceeds a certain amount, the latter being equivalent to the time of the first buffer overflow for a finite queue (see, for instance, [10], [6], [12], [16], [17], [22], [24]). More recently, queues with finite buffers have been studied directly, often with the aim of obtaining expressions for the loss probabilities (see, for instance, [19], [26], [29]).

If network models are to be built incorporating long range dependent traffic models then it is also important to understand the output processes from such queues, since they (or some aggregation of a collection of them) will often constitue the input process at some other queue. Several authors have considered the output process from queues with infinite capacity. Daley and Vesilio [11] consider the single-server queue with infinite capacity and renewal arrivals and show that the output process can be long range (count) dependent if either the input process is, or if the service times are heavy tailed. Anantharam [3] considers a network of S-reversible infinite capacity queues and shows that if the external arrival processes are long range dependent then so are the internally routed streams of customers. However, relatively little is known about the output process from a *finite* queue with long range dependent input and it is not immediately clear that the property of long range dependence is maintained as traffic passes through buffers and switches with finite capacity. Several authors have shown that for finite buffers the impact on loss probabilities of correlation in the arrival process becomes negligible beyond a time scale that is related to the maximum buffer size. Ryu and Elwalid [25] call this the "Critical Time Scale" and Grossglauer and Bolot [15] the "correlation horizon". It has been suggested that well-designed Markov traffic models can be effective for predicting quality of service (loss probabilities), even in the presence of long range dependence ([15], [18], [25], [14]). If this is so, then it suggests that even when the input process is long range dependent, there may be circumstances in which the output process is not.

In this paper, we use simulations to study the output process of a finite queue with long range dependent input. Two different kinds of input are considered. The first are traces taken from a very extensive collection of data made by the WAND group at the University of Waikato ([8] and http://wand.cs.waikato.ac.nz). These data show clear evidence of long range dependence. The second set of simulations use simulated inputs with Pareto distributed inter-arrival times, and superpositions of processes with Pareto inter-arrival times. The particular question of interest is whether the property of long range dependence is preserved as the traffic passes through the buffer — that is, is the output process from the buffer always long range dependent. Throughout, we use the non-parametric wavelet estimators developed by Abry *et al.* ([1], [2]) to test for the presence of long range dependence.

In section 2 we give some definitions and background. In section 3 we describe the

two trace data sets that we will be using for our simulations. Section 4 contains the queue simulations, using both the trace data sets, and simulated processes with Pareto distributed inter-arrival times as inputs to the queues. Finally, in section 5 we give our conclusions with some further discussion.

2 Definitions and background

Long range dependence occurs when correlations decay as a power law rather than exponentially. More formally, a second order stationary process $\mathbf{X} = \{X(t), t \in Z\}$ with finite mean and variance and covariance function $r(k) = Cov(X(j)X(j+k)), k \in Z)$ is said to be long-range dependent if $r(k) \sim c_r k^{\alpha-1}$ as $k \to \infty$, where $\alpha \in (0,1)$ is the scaling parameter [4], [9]. Thus the autocorrelation function decays hyperbolically with increasing lag and $\sum_k r(k) = \infty$. (Equivalently, the spectrum at the origin diverges according to a power law.) Long range dependence, if present, has serious implications for the design and control of networks.

In the context of queueing processes, Daley and Vesilio[11] distinguish between two kinds of long range dependence. Long range interval dependence (LRiD) is long range dependence of the inter-event times of some process associated with the queue (either the input process or the output process). Long range count dependence (LRcD) is long range dependence of the counts of events in successive intervals.

These two kinds of long range dependence are not at all equivalent. For instance, a renewal process, which by definition has independent and identically distributed inter-event times, cannot be LRiD. Daley and Vesilio show, however, that a renewal process is long range *count* dependent if the inter-event distribution function, $F(\cdot)$ is heavy-tailed in the sense that

$$1 - F(x) = x^{-c} L(x) \qquad\qquad x > 0.$$

where $1 < c < 2$ and $L(x)$ is a function slowly varying at infinity (that is, $\lim_{x\to\infty} L(tx)/L(x) = 1$ for every finite $t > 0$).

Daley and Vesilio then consider the infinite capacity queue with an arrival process that has heavy-tailed inter-event times (that is, a long range count dependent arrival process). They show that such a queue with exponential service times will also have a long range count dependent output process. They also show that queues with Poisson arrivals and a heavy-tailed service time distribution can produce long range count dependent output.

In this paper we will be examining the output from a *finite* queue for the presence of long range dependence, and the approach will be via simulations, in the spirit of [13]. We will therefore need to estimate the scaling parameter α. Several techniques are available to do this, amongst them the Whittle estimator and various graphical methods (the R/S estimator, the variance-time curve and the periodogram [4]). In

this paper we will use the non-parametric wavelet estimator [2]. In the remainder of this section we give a very brief exposition of the wavelet method – this material is drawn from Bruce and Gao [7] and Abry *et al.* [1].

Wavelets are building block functions, analogous to sine and cosine functions. However, unlike sine and cosine functions, they are localized in time or space. Wavelets are designed in families. Father wavelets, ϕ, satisfy $\int \phi(t)dt = 1$, and capture the smooth and low-frequency components of a trace. Mother wavelets, ψ, satisfy $\int \psi(t)dt = 0$ and capture the detail and high-frequency components of a trace. There are many different wavelet families in use – the wavelet estimator proposed by Abry *et al.* ([1], [2]) uses the Daubechies wavelet.

Define

$$\phi_{j,k}(t) = \frac{1}{2^{j/2}}\phi\left(\frac{t}{2^j} - k\right)$$

and

$$\psi_{j,k}(t) = \frac{1}{2^{j/2}}\psi\left(\frac{t}{2^j} - k\right).$$

Thus the functions ϕ and ψ are scaled by a factor 2^j (also known as the dilation or level), and translated by $2^j k$ (also known as the location or shift). The index j is sometimes also referred to as the octave. Then the wavelet series approximation for a continuous time trace, $f(t)$, is given by

$$f(t) \approx \sum_k s_{J,k}\phi_{J,k}(t) + \sum_k d_{J,k}\psi_{J,k}(t) + \sum_k d_{J-1,k}\psi_{J-1,k}(t) + \ldots + \sum_k d_{1,k}\psi_{1,k}(t)$$

where the $s_{J,k}, d_{j,k}$ are the wavelet transform coefficients; J is the number of scales (or multiresolution components); and k runs from 1 to the number of coefficients in the specified component. The transform coefficients are given by

$$s_{J,k} \approx \int \phi_{J,k}(t)f(t)dt$$

$$d_{j,k} \approx \int \psi_{J,k}(t)f(t)dt, \qquad j = 1, 2, \ldots, J$$

and their magnitude is a measure of the contribution of the coresponding wavelet function to the series.

The discrete wavelet transform (DWT) calculates the coefficients of the wavelet series approximation for a discrete, finite, trace $x_1, x_2, \ldots, x_n$. A multiresolution analysis (MRA) separates the DWT into its various components thus:

$$detail_j(t) = \sum_k d_x(j,k)\psi_{j,k}(t)$$

with

$$x(t) = approx_J(t) + \sum_{j=1}^{J} detail_j(t).$$

A stationary finite variance discrete time series, $\mathbf{X}$ is long range dependent if its spectral density $\Gamma_X(\nu)$ satisfies

$$\Gamma_X(\nu) \sim c_f|\nu|^{1-2H}, \qquad \nu \to 0.$$

An estimate of the spectral density is

$$\hat{\Gamma}_X(\nu)\left(\frac{\nu_0}{2^j}\right) = \frac{1}{n_j}\sum_k |d_x(j,k)|^2 \equiv \mu_j$$

where $n_j \approx n/2^j$, n being the length of the data, and ν_0 is an arbitrary frequency determined by the choice of ψ_0, the mother wavelet.

A linear regression of $ln(\hat{\Gamma}_x(\nu)\left(\frac{\nu_0}{2^j}\right))$ on j, that is, of $ln(\frac{1}{n_j}\sum_k |d_x(j,k)|^2) = \hat{c} + (2\hat{H} - 1)j$ on j then gives an estimate of H, or, alternatively, of $\alpha = 2H - 1$. Specifically, let $y_j = log(\mu_j) - g(j)$ where $g(j)$ is a correction factor. If the $d_X(j,k)$ are uncorrelated both within and across scales, and the process $\mathbf{X}$ (and hence also the process $d_X(j,\cdot)$) is Gaussian, then

1. $g(j) = \psi(n_j/2) - log_2(n_j/2)$, where $\psi(z) = \Gamma'(z)/\Gamma(z)$ is the Psi function, $\Gamma(z)$ the gamma function.
2. $E(y_j) = \alpha j + log_2 c_f C$.
3. $Var(y_j) = def\sigma_j^2 = \zeta(2, n_j/2)/ln^2 2$ where $\zeta(z,v)$ is a generalized Riemann Zeta function.
4. The estimator, $\hat{\alpha}$, of α is the slope of a weighted linear regression of the y_j on j given by
$$\hat{\alpha} = \frac{\sum y_j(jS - S_j)/\sigma_j^2}{SS_{jj} - S_j^2}$$
where $S = \sum \sigma_j^{-2}$, $S_j = \sum j\sigma_j^{-2}$ and $S_{jj} = \sum j^2\sigma_j^{-2}$.

The mother wavelet, ψ_0, can be chosen so as to have N vanishing moments, that is, so that

$$\int t^k\psi_0(t)dt = 0, \qquad 0 \le k \le N-1.$$

If it has N vanishing moments, then the Fourier transform of the wavelet about $\nu = 0$ satisfies $|\psi_0(\nu)| = O(\nu^N), \nu \to 0$. Trends in the series can be detected and their effect on the estimation of α or H eliminated, by considering mother wavelets with larger numbers of vanishing moments.

Abry and Veitch[2] show that the estimates of α and/or H obtained using this method are unbiased and efficient. They are superior to to graphical methods such as the R/S statistic, some of which are biased, even asymptotically. They do no worse than the Whittle estimator, which is asymptotically unbiased, and may do better.

3 The data

The two test data sets that we will use in our simulations below are segments taken from the ATM link at the University of Auckland on 8 July 1999, between the hours of 4 p.m and 6 p.m, a time when the link is heavily loaded. Each segment contains 100,000 time stamps. One segment is inbound traffic and the other is outbound traffic. The full data sets for the two hours contain about 2,000,000 time stamps, but since there was some evidence of nonstationarity, a shorter sequence is considered here.

The traces are part of a large collection of GPS-synchronized IP header traces captured with a DAG2 system at the University of Auckland Internet uplink by the WAND (Waikato Applied Network Dynamics) research group, which is based in the University of Waikato Computer Science Department. The University of Auckland ITSS department is operating an OC3 ATM link to carry a variety of services off the main campus. A single ATM channel is used to connect the university to the global Internet, and since it is the only connection, all packets for all external connections pass the measurement point. The connection has a packet peak rate of 2 MBits/sec in each direction. (For more details see the WAND site http://wand.cs.waikato.ac.nz and the NLANR MOAT site, http://moat.nlanr.net/Traces/Kiwitraces, where later traces from November 1999 are available.)

Both traces were examined for evidence of long range interval dependence and long range count dependence using the wavelet analysis described in section 2 above. As described there, the logscale diagram is examined for regions of alignment. (The logscale diagram is a plot of the logarithm of variance estimates of the wavelet details at each scale against the scale at which they are estimated.) A region of alignment at large scales, with slope between 0 and 1, indicates the presence of long-range dependence. An estimate of the scaling parameter α is then obtained by taking a weighted regression over this region. The code to do this was kindly provided by Darryl Veitch.

Both traces showed evidence of long range interval dependence. This can be seen on the logscale diagrams which are given in Figure 1. Many such logscale diagrams were generated to do the work here – these two are given here to illustrate the method. The estimate of the scaling parameter, α, is obtained from the fitted regression line on the logscale diagram. For trace 1 it was 0.8170, with 95% confidence interval $(0.703, 0.931)$. The estimate of the scaling parameter for trace 2 was 0.8575, with 95% confidence interval $(0.685, 1.030)$. We note that the mean interarrival time for the first trace is 0.0047, with sample standard deviation 0.0057; and for the second it is 0.0027 with sample standard deviation 0.0029.

Both traces were then aggregated into counts over intervals in order to test for long range count dependence. The first trace was aggregated into counts over intervals of 0.028 seconds, giving a total of 16,755 intervals. The estimate of α for trace 1 was 0.771 (95% confidence interval $(0.668, 0.873)$), with the number of vanishing moments

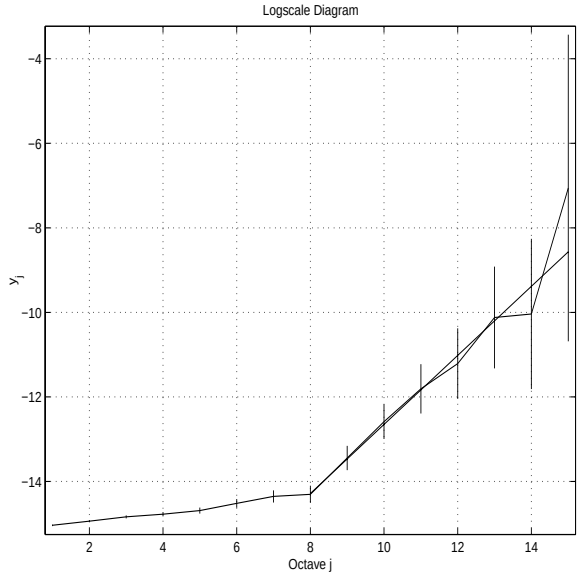


Figure 1.a) Logscale diagram for interarrival times in trace 1.

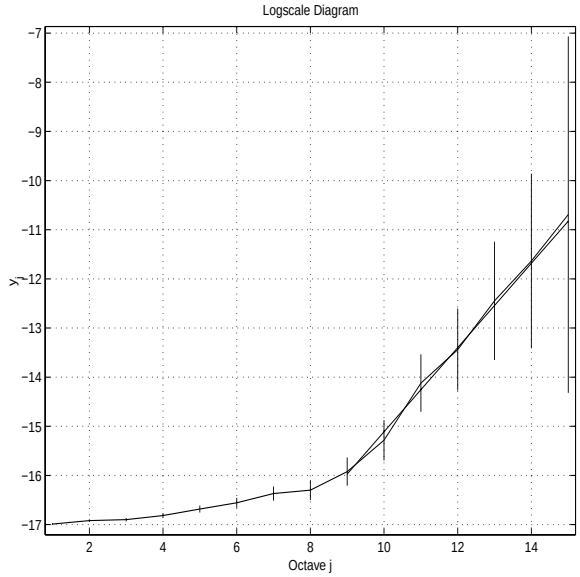


Figure 1.b) Logscale diagram for interarrival times in trace 2.

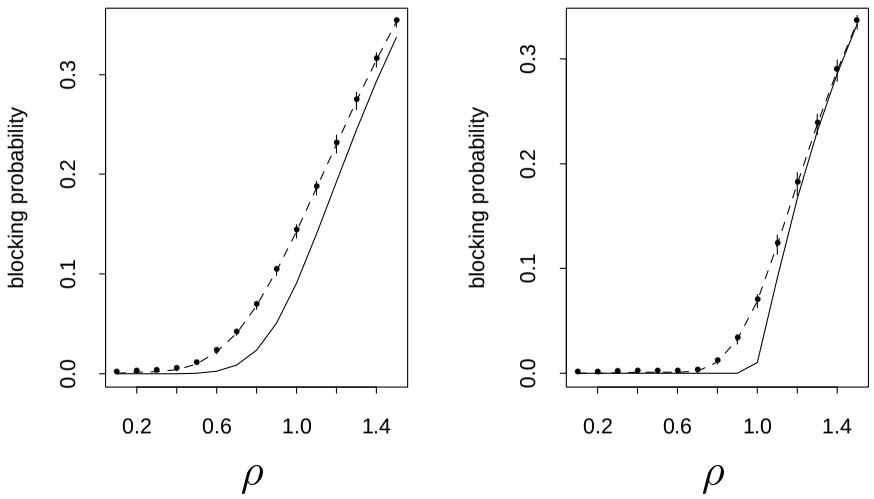
Figure 1: Logscale diagrams for interarrival times in the two data sets.

set to $N = 2$. For the second data set the counts were aggregated into intervals of length 0.016 seconds. The estimate of the scaling parameter, α, for this trace was 0.7435 with 95% confidence interval $(0.604, 0.883)$.

4 Simulations

4.1 Simulations with trace input

The sets of data described in section 3 were used as the input to a single server queue with a finite buffer and simulated exponential service times. We simulated systems with capacity 10 and 100. The estimated loss probabilities, when using data set 1 as input, are given below in Figure 2 for a range of service rates, chosen so that ρ, the arrival rate divided by the service rate (or traffic intensity), ranged from 0.1 to 1.5. Let $\hat{b}_C(\rho)$ be the estimated blocking probability for the buffer with trace input, buffer size C, and traffic intensity ρ. In each case we compare these blocking probabilities with those of an $M|M|1$ queue with the same load, $l_C(\rho) = (1-\rho)\rho^C/(1-\rho^{C+1})$.



simulated loss probability, $\hat{b}_C(\rho)$, – – – – – –
$M|M|1$ loss probability, $l_C(\rho)$ ————————

Figure 2: Blocking probabilities Fig (a), C=10, Fig (b), C=100

We see that the absolute difference in loss probabilities between the simulations and the $M|M|1$ model is greatest around $\rho = 1$. However, if we examine the log of the relative difference, we find that for small ρ, the loss probabilities are of different orders of magnitude (Figure 3).

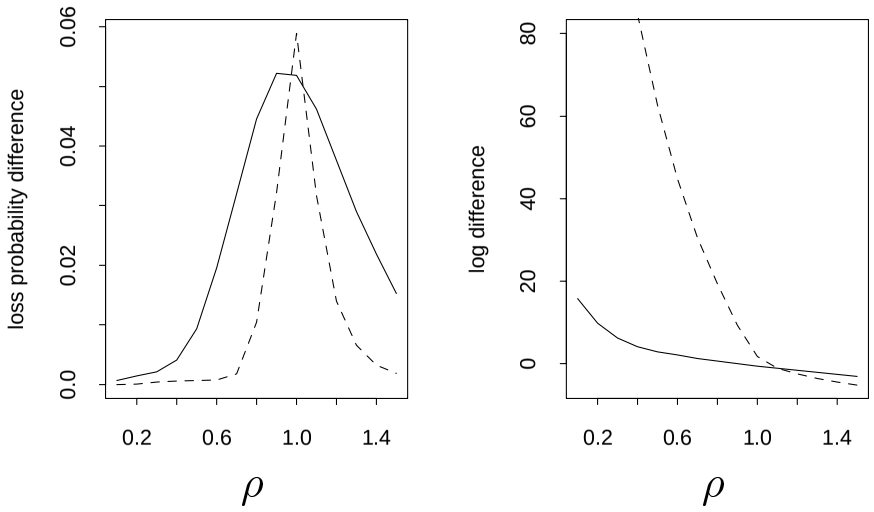


Figure 3: Difference in loss probabilities between simulated queue and $M|M|1$ queue.
Fig (a) $\hat{b}_C(\rho) - l_C(\rho)$
Fig (b) Log of relative difference $log((\hat{b}_C(\rho) - l_C(\rho))/l_C(\rho))$

We now examine the output from these queues for long range interval dependence and long range count dependence. In Figure 4 below we see estimates for the scaling parameter α, for both the inter-departure times of the process, and the counts of departures in successive intervals of length 0.028 seconds for the single-server queue with $C = 100$. The estimates are given with their 95% confidence intervals, all obtained using the wavelet estimators described above. For the binned counts, it was necessary to take the number of vanishing moments, $N = 2$, for $\rho < 1$ — this ties in with the finding above that the binned input data also required $N = 2$ when estimating α. The most striking feature of both plots is that for ρ sufficiently large, there is no evidence of long range dependence of either kind, at the time scales over which the simulations have been made. In addition, it appears that as ρ increases above critical loading ($\rho = 1$), the scaling parameter, α, decreases to 0.

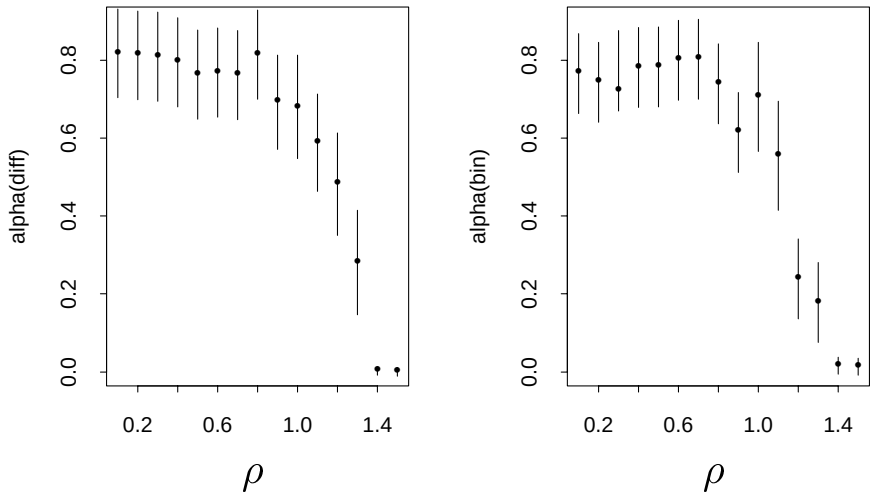


Figure 4: Estimates of α for departure process from queue with $C = 100$.
Fig (a) Inter-departure process (LRid)
Fig (b) Departure counts in intervals of length 0.028 (LRcD)

The same analysis is performed with $C = 10$, and the estimates of α appear in Figure 5. Again, we see that dependence, at least on the time scale at which these simulations are made, is no longer apparent for large ρ. An interesting feature here is that one might have expected the dependence to vanish at lower ρ with smaller capacity, C. However, that appears not to be the case.

There is as yet no automatic means for deciding exactly which of the possible fits for α is best — if the scaling varies for differing time scales then the fit is made to the highest octaves available and it may not always be clear which provides the best fit. However, qualitatively the features observed in the plots above appear for any of the fits that were reasonable.

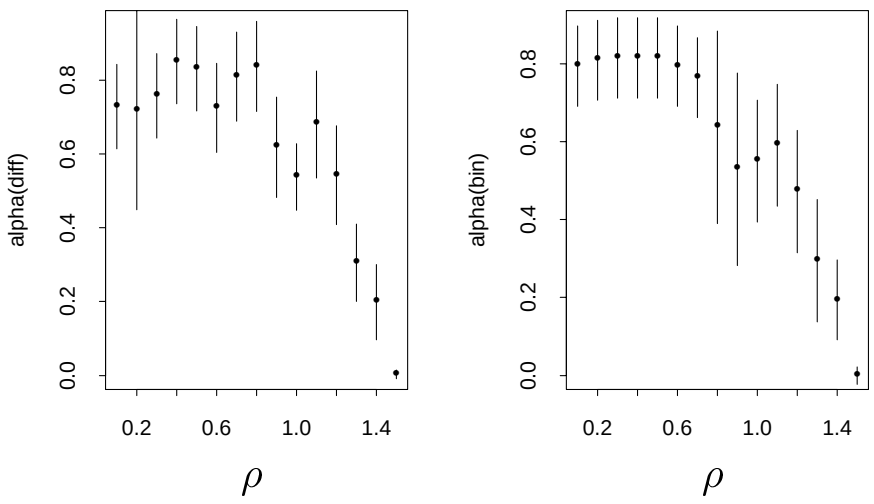


Figure 5: Estimates of α for departure process from queue with $C = 10$.
Fig (a) Inter-departure process (LRid)
Fig (b) Departure counts in intervals of length 0.028 (LRcD)

We do not display results for the second data set here, but note that they were similar, with long range dependence no longer being apparent at a lower traffic intensity than for the first data set (at $\rho = 1.2$ when $C = 100$ and $\rho = 1.4$ when $C = 10$.

4.2 Simulations with Pareto inter-arrival times

In this section we consider the output of a finite queue with capacity $C = 100$ and simulated input, rather than trace input. We considered two models for the input. The first is a model with Pareto inter-arrival times. The Pareto distribution is heavy-tailed and has distribution function, $F(\cdot)$ given by ([20])

$$F(x) = 1 - \left(\frac{k}{x}\right)^c \qquad k > 0, c > 0; x \geq k.$$

The second model that we consider superimposes (or multiplexes) five such processes with Pareto interarrival times. In each case, we fix the parameters $k = 1$, $c = 1.8$. Thus the interarrival times for the first model have finite mean 2.25 and infinite variance. Both arrival processes with Pareto distributed interarrival times, and superpositions of such processes have been frequently suggested as traffic models. Our aim here is to study the output processes that these models will generate. We note that the interarrival times for the datasets considered in the previous section do not follow those of a Pareto distribution (qq plots indicated that the tails were not as

heavy as those of a Pareto distribution), but they are rather better modelled by several Pareto arrival processes multiplexed together.

We consider first the simple renewal input process with Pareto inter-arrival times. We know from [11] that this input process is long range count dependent. However, it is not long range interval dependent and in this respect it differs markedly from the trace data. We again performed simulations for a range of loadings, and several input traces, and did not find any instances where the output process was long range interval dependent. And again, for $\rho \geq 1.1$ there was no evidence of long range count dependence.

The input process derived from the superposition of several Pareto processes is of greater interest, since it does appear to be long range interval dependent, as well as long range count dependent. The simulated input process had mean interarrival time 0.4203 and sample standard deviation 0.3464. It was long range interval dependent with scaling parameter $\alpha = 0.352$ (95% confidence interval (0.326,0.378)) and long range count dependent with scaling parameter $\alpha = 0.388$ (95% confidence interval (0.353, 0.421)). Estimates of α for the output process obtained from a single simulation are displayed in Figure 6 below – these are typical of the several simulations that we considered.

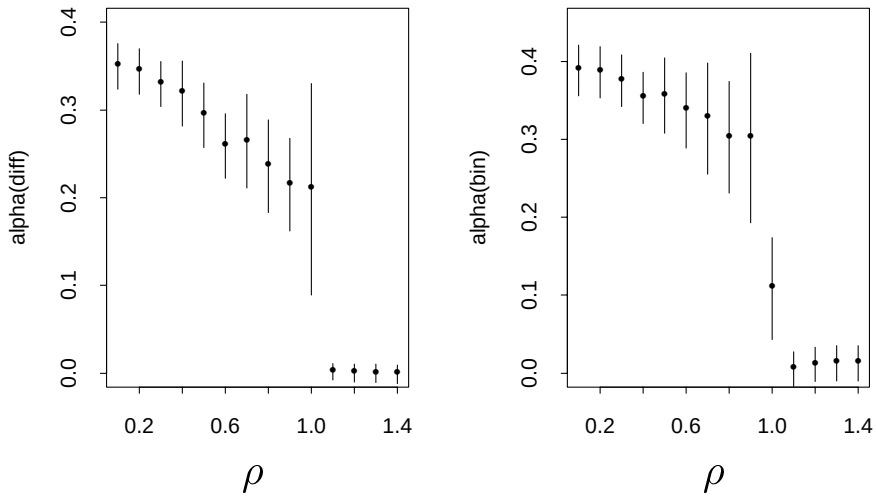


Figure 6: Estimates of α for departure process from buffer with sumperimposed Pareto input process.

Fig (a) Inter-departure process (LRid)

Fig (b) Departure counts (LRcD)

The estimate of the scaling parameter, α, is somewhat lower for these simulations than it was for the data sets and trace-driven simulations given above, which is a

consequence of the lower value of α for the input process. We note that, as with the trace-driven simulations, the long range dependence vanishes for ρ sufficiently large, and that is in this case, as with the renewal Pareto input, that occurs for ρ close to 1.

5 Conclusions and direction for further work

The simple simulations performed in this paper indicate that the output process from a finite buffer will not always exhibit long range dependence. It is true that in the examples studied here this appears to be so only at very heavy loadings with $\rho > 1$ – however, this does suggest that, particularly in heavily congested networks, it may not always be the case that the property of long range dependence is maintained throughout the network. A next step would be to provide analytical results which provide more insight into the output process. It might also be of interest to consider the output from a series of tandem queues, rather than just a single finite queue.

6 Acknowledgements

The author is grateful to the members of the WAND group for providing the data and for encouraging involvement in the project. Thanks are also due to Darryl Veitch for introducing and explaining the wavelet estimator and providing the code to implement it. This work was done with the assistance of a FRST PGSF grant, which is gratefully acknowledged.

References

[1] Abry, P., Flandrin, P., Taqqu, M.S. and Veitch, D. (1998) Wavelets for the analysis, estimation and synthesis of scaling data. Preprint.

[2] Abry, P. and Veitch, D. (1998) Wavelet analysis of long range dependent traffic. *IEEE Transactions on Information Theory*. 1998.

[3] Anantharam, V. (1996) Networks of queues with long-range dependent traffic streams. In: *Stochastic Networks* ed. Glasserman, P., Sigman, K. and Yao, D.D. Springer, 1996. pp. 237-256.

[4] Beran, J. *Statistics for long-memory processes.* Chapman and Hall. 1994.

[5] Beran, J., Sherman, R., Taqqu, M.S. and Willinger, W. (1995) Long-range dependence in variable-bit-rate video traffic. *IEEE Transactions on Communications*, **43**, 1566-1579.

[6] Brichet, F., Roberts, J., Simonian, A. and Veitch, D. (1996) Heavy traffic analysis of a storage model with long range dependent on/off sources. Preprint. 1996.

[7] Bruce, A. and Gao, H.-Y. *Applied wavelet analysis with S-plus*. Springer. 1996.

[8] Cleary, J., Graham, I., McGregor, T., Pearson, M., Ziedins, I., Curtis, J., Donnelly, S., Martens, J. and Martin, S. (1999) High precision traffic measurement by the WAND research group. Working Paper 99/17. The University of Waikato.

[9] Cox, D.R. (1984) Long-range dependence: a review. In H.A. David and H.T. David, editors, *Statistics: An Appraisal*, 55-74. Iowa State University Press. 1984.

[10] Crovella, M.E. and Bestavros, A. (1997) Self-similarity in World Wide Web traffic: evidence and possible causes. *IEEE/ACM Trans. on Networking* **5**, 835-846.

[11] Daley, D.J. and Vesilo, R. (1997) Long range dependence of point processes, with queueing examples. *Stochastic Processes and their Applications*, **70**, 265-282.

[12] Duffield, N.G. and O'Connell, N. (1995) Large deviations and overflow probabilities for the general single-server queue, with applications. *Proc. Cambridge Phi. Soc.*

[13] Erramilli, A., Narayan, O. and Willinger, W. (1996) Experimental queueing analysis with long-range dependent packet traffic. *IEEE/ACM Trans. on Networking* **4**, 209-223.

[14] Feldman, A. and Whitt, W. (1997) Fitting mixtures of exponentials to long-tail distributions to analyze network performance models. *IEEE INFOCOM '97*, 1096-1104.

[15] Grossglauer, M. and Bolot, J-C. (1996) On the relevance of long-range dependence in network traffic *Proceedings SIGCOMM '96*, 15-24.

[16] Heath, D., Resnick, S. and Samorodnitsky, G. (1997) Patterns of buffer overflow in a class of queues with long memory in the input stream. *Ann. Appl. Prob.* **7**, 1021-1057.

[17] Heath, D., Resnick, S. and Samorodnitsky, G. (1998) How system performance is affected by the interplay of averages in a fluid queue with long range dependence induced by heavy tails. Preprint. Cornell University.

[18] Heyman, D.P. and Lakshman, T.V. (1996) What are the implications of long-range dependence for VBR-video traffic engineering? *IEEE/ACM Transactions on Networking*, **4**, 301-317.

[19] Jelenkovic, P. (1999) Long-tailed loss rates in a $GI/GI/1$ queue with applications. *Queueing Systems* **33**, 91-125.

[20] Johnson, N.L. and Kotz, S. *Continuous univariate distributions* Wiley. 1970.

[21] Leland, W.E., Taqqu, M.S., Willinger, W. and Wilson, D.V. (1994) On the self-similar nature of Ethernet traffic (Extended version). *IEEE/ACM Trans. Networking*, **2**, 1-15.

[22] Norros, I. (1994) A storage model with self-similar input. *Queueing systems*, **16**, 387-396.

[23] Paxson, V and Floyd, S. (1995) Wide-area traffic: the failure of Poisson modelling. *IEEE/ACM Trans. on Networking* **3**, 226-244.

[24] Roughan, M., Veitch, D. and Rumsewicz, M. Computing queue-length distributions for power-law queues. Preprint.

[25] Ryu, B.K. and Elwalid, A. (1996) The importance of long-range dependence of VBR video traffic in ATM traffic engineering: myths and realities. *ACM SIGCOMM Computer Communication Review*, **26**, 3-14.

[26] Tsybakov, B. and Georganas, N.D. (1997) On self-similar traffic in ATM queues; definitions, overflow probability bound and cell delay distribution. *IEEE/ACM Trans. on Networking* **5**, 397-409.

[27] Willinger, W., Taqqu, M.S., Leland, W.E. and Wilson, D.V. (1995) Self-similarity in high-speed packet traffic: analysis and modeling of ethernet traffic measurements. *Statistical Science* **10**, 67-85.

[28] Willinger, W., Taqqu, M.S. and Erramilli, A. (1996) A bibliographical guide to self-similar traffic and performance modeling for modern high-speed networks. In: *Stochastic Networks* ed. Kelly, F.P., Zachary, S. and Ziedins, I. Oxford University Press, 1996. pp. 339-366.

[29] Zwart, A.P. (2000) A fluid queue with a finite buffer and subexponential input. *Adv. Appl. Prob.* **32**, 221-243.