Performance Analysis of a Poisson-Pareto Queue over the Full Range of System Parameters

Ronald G. Addie[1], Timothy D. Neame[2] and Moshe Zukerman[3]

1. Department of Mathematics and Computing, University of Southern Queensland, Australia, Tel +61 7 46 31 5520, Fax: +61 76 46 31 5550.
2. ARC Special Research Centre for Ultra-Broadband Information Networks (CUBIN), The University of Melbourne, Parkville, Victoria 3010, Australia.
3. Electronic Engineering Department, City University of Hong Kong, Hong Kong.

Abstract

There have been many queuing analyses for a single server queue fed by an M/G/∞ traffic process, in which G is a Pareto Distribution, that focus on certain limiting conditions. In this paper we enhance the so-called Quasi-Stationary (QS) approximation – a queuing analysis introduced previously that provides an algorithm for computation of an accurate approximation for the stationary queue distribution, applicable to the entire range of system parameters. By numerical evaluation of the QS approximation and the asymptotic approximations (large buffer, many sources, and heavy traffic) over an extremely wide range of parameter values we are able to graphically display consistency of the QS approximation with all the asymptotic results. We demonstrate that the accuracy of the asymptotic approximations is satisfactory only in limited regions of the system parameter space.

Key words: Large Deviation Theory, Long Range Dependence, Queueing Theory, Pareto Distribution, PPBP

1 Introduction

The discovery that Internet traffic has Long Range Dependent (LRD) characteristics [1] has resulted in much attention given to analyses of queues with LRD input [2–28]. A popular family of traffic models that exhibits the LRD phenomenon and also captures the behaviour of Internet traffic is the one based on a Poisson arrival stream of random, heavy-tailed, or more specifically Pareto distributed bursts [2, 3, 10, 14–16, 18, 19, 22, 25, 26, 28]. This model is widely referred to as the M/G/∞ traffic process. Roberts *et al.* [29] use the name *Poisson burst process* for it. The traffic model we consider is a special case of the M/G/∞ traffic process in which G is a Pareto distribution. Accordingly, and following [2, 30], we call it the *Poisson Pareto Burst Process (PPBP)*.

The PPBP takes the form of overlapping bursts. The arrival times of the bursts form a Poisson process, each burst generates bits at a constant rate r, and the length of each burst is Pareto distributed with scale parameter δ and shape parameter γ. We consider a single server queue fed by a PPBP

Table 1
Classification of results on queues with $M/G/\infty$ traffic

Method	Scaling	Publications
Large buffer large deviations limit	buffer size $\to \infty$	[5, 6, 15, 21–23, 25, 26, 31]
Many sources large deviations limit	$n \to \infty$; buffer size & C linear in n	[10, 17–19, 32–34]
Heavy Traffic & CLT limit	$n \to \infty$; buffer size & net mean input linear in $\sqrt{n}$	[7, 8, 39, 40]
Heavy Traffic non CLT limit	$\rho \to 1$ and buffer size $\sim (1-\rho)^{\frac{-1}{\gamma-1}}$	[27]
Quasi-stationary approximation	NA	[2, 41]

process. Let ρ be the ratio of the total arrival rate [bits/sec] of all active bursts to the service rate [bits/sec] of the server.

The papers [5, 6, 15, 21–23, 25, 26, 31] consider asymptotic regimes where the buffer size, or buffer threshold (in an infinite buffer case), tends to infinity while the number of sources and the server rate C (and consequently the offered load) are fixed. This asymptotic regime is widely referred to as the *Large Buffer Limit*. The papers [10, 17–19, 32–34] consider the case where the buffer size and server speed are linear in the number of sources, which tends to infinity. This is generally known as the *Many Sources Limit*. In both these asymptotic regimes the probability of overflow tends to zero, as either the buffer level increases or the number of sources increases, and therefore these situations are referred to as *Large Deviations* limits.

Another important case which has received attention is where the buffer size grows in proportion to $\sqrt{n}$, where n is the number of sources. This has been proposed as a practical way to provision buffers to cope with growth in traffic [35] and similar or related comments are made in [36–38]. In this literature it is assumed that the server rate also increases with the number of sources, n, in such a way that queueing performance tends to a limit not depending on n. In order that the limit of the buffer distribution exists, in these models, the server rate increases in such a way that the net mean input rate (i.e. the difference between rate of arrival of work and the server speed) also increases linearly with $\sqrt{n}$. As n increases, the *utilization* tends to 1, and consequently these results are also termed, here and elsewhere, *heavy traffic approximations*. This literature often relies explicitly on convergence of the input traffic process to a Gaussian process, and hence we associate these results with the Central Limit Theorem (CLT). In these cases the server rate increases along with the burst arrival intensity in such a way that the CLT applies.

A different heavy traffic limit is provided in [27] where the server speed remains constant as the intensity of burst arrivals increases and the buffer level is scaled to ensure that a limit occurs. As a consequence asymptotically power-law behaviour is exhibited, much as in the large buffer large deviations results. Because in this asymptotical regime system utilization is increased towards 1, with server speed held constant, the limit is consistent with the large buffer limit in the special case where one more than the average number of flows is sufficient to overload the server. This paper also makes use of a *light traffic approximation*, which provides an estimate of the probability of queue non-emptiness which is asymptotically accurate as traffic becomes lighter.

Over a decade ago Choudhury *et al.* [42, 43] demonstrated that in many cases the tail may not be characteristic of the entire distribution. For the case of LRD queues, results obtained for LRD Gaussian queueing models in [30, 44] based on the CLT indicate a different shape of the queue distribution

than that of the tail obtained by Large Deviations Theory, suggesting that LRD queues might also have the feature that the tail behaviour of their queue distributions is not characteristic of the distribution as a whole.

In addition to these asymptotic approximations, the Quasi-Stationary (QS) approximation for the buffer level stationary distribution was presented in [2], which we here show to be a *lower bound*. This approximation was validated against a specially tailored type of simulation in [2]. Simulation has only rarely been used in studies of PPBP queueing systems, possibly because conventional simulations cannot successfully include details at the wide range of time scales needed to provide accuracy. In this paper we explore the boundaries between the regions where one or the other asymptotic approximation is more accurate. The QS approximation is consistent with all the asymptotic results, although the buffer level at which the large buffer approximation becomes approximately the same as the QS approximation may be, depending on the parameters of the system, rather large. All the approaches considered in this paper are summarized in Table 1.

The remainder of the paper is organised as follows. In the next section, we describe the model of a single server queue fed by PPBP input. In Section 3, we review the asymptotic results available in the literature for this system: the large buffer asymptote (where buffers increase but traffic levels stay constant), the many sources asymptote (in which buffer sizes increase in proportion to traffic), the heavy traffic approximation in which buffers increase with the square root of traffic, and the heavy traffic approximation ([27]), in which the server speed remains fixed as utilization approaches 1 and convergence to a limit occurs by scaling buffer levels. In Section 4 the QS approximation of [2], which relies on separation of traffic into long and short bursts, is further developed. A more rigorous derivation, including a demonstration that it provides a lower bound, is provided and some numerical refinements are introduced which enable the method to be used to evaluate very small probabilities, in order to be able compare the QS approximation to the large buffer asymptote in the remote regions where the two approach each other.

In Section 5, two arguments are used to show that the power-law behaviour ($\sim c_\lambda x^\kappa$, where x is buffer level) of the stationary queue distribution in a PPBP single server queue is only exhibited for a very remote region in the parameter space. First, it is shown that any power-law upper or lower bounds on the tail of the stationary complementary distribution function (CDF), or an exact asymptotic power-law approximation for the queue stationary CDF, necessarily diverges unboundedly from the stationary queue CDF as the rate of the PPBP increases. Secondly, it is shown that the level, x_λ, where power-law behaviour begins, is unbounded as a function of λ, the arrival rate of bursts, as it varies even over a finite range, let alone as $\lambda \longrightarrow \infty$.

In Section 6, the large buffer asymptote, the CLT (heavy traffic) limit, and the QS estimates are all shown on the same graph, in Figure 8, which illustrates how these estimates relate to each other. This graph shows clearly that although the heavy traffic limit and the large buffer asymptote appear to have very different (apparently contradictory) characteristics, the QS estimate is consistent with both, illustrating thereby that these two limits are not in contradiction. The QS approximation is also compared to the many sources asymptote. Plots of c_λ, the weight of the power-law tail of the CDF, as a function of λ are also presented here which demonstrate that c_λ is highly sensitive to the parameters of the system. Concluding remarks are presented in Section 7.

2 The queueing model

We consider a single server queue with constant service rate, C [bits/sec], and PPBP input. As discussed, the PPBP traffic model is made up of bursts. The arrival times of each burst form a Poisson process with rate λ [bursts/sec]. Let d [seconds] be a random variable representing the burst duration. We assume that the rate at which data is generated *during* each burst is constant for the duration of each burst and the same for all bursts, hereafter denoted by r [bits/sec]. This assumption is common in the literature, with the exception of [15, 25].

Throughout this paper, we focus on the PPBP case where d follows a Pareto distribution. The CDF of the Pareto distribution used in this paper takes the form:

$$\Pr(d > x) = \begin{cases} \left(\frac{x}{\delta}\right)^{-\gamma}, & x \geq \delta, \\ 1, & \text{otherwise}, \end{cases} \tag{1}$$

in which $\delta > 0$ [seconds] and $\gamma > 0$. As mentioned in the Introduction, δ is the scale parameter and γ is the shape parameter of the Pareto distribution. We have $E(d) = \infty$ for $0 < \gamma \leq 1$, and for $\gamma > 1$, $E(d) = \frac{\delta\gamma}{(\gamma-1)}$. For $0 < \gamma < 2$, the variance of d is infinite. The lower the value of γ, the heavier the tail of the Pareto distribution becomes. In the sequel we shall generally assume that $1 < \gamma < 2$ unless otherwise indicated. Thus, the PPBP queueing systems considered here are characterized by five parameters: λ, r, δ, γ, and C. In several places in the sequel we will consider a scaling in which the mean and second order statistics of the the input process minus the service process, which is termed the *net input process*, will be held constant. The mean of this process shall be termed the *net mean input*.

An alternative definition of the Pareto distribution, in which the density is non-zero for all $x \in (0, \infty)$, could be used without significantly affecting the conclusions of this paper. The definition (1) has the practical feature that there is a non-zero shortest burst length.

Let Y_t be the amount of work [bits] that arrives between time 0 and time t, multiplied by (-1) if $t < 0$. Then for any real numbers s, t, with $t > s$, positive or negative, $Y_t - Y_s$ is the amount of work arriving between time s and time t. Let Q_t be the queue size process, which is also the virtual waiting time in our case where work arrives and is served continuously. By Reich's formula Q_t is given by

$$Q_t = \sup_{s \leq t} \{Y_t - Y_s - C(t-s)\}.$$

If at time t, we have that $Q_t > 0$, the *busy period* that includes time t, is said to start at the time s such that the above supremum occurs for this particular value of s. By the above definitions, Y_t is stationary, therefore Q_t is stationary, so henceforth we omit the index t, and use the random variable Q to denote the stationary queue size.

The terms *queue* and *buffer* are treated as synonymous, and we shall refer to the *queue distribution*, and sometimes Complementary Distribution Function (CDF) of a queue, to mean the same thing as the distribution, or CDF, of buffer level. In an equation such as $P(Q > x) = y$, we shall refer to Q as the stationary buffer level or queue size and x as the threshold that Q exceeds.

Because of the relationship between losses in finite buffers and overflows (exceeding a certain level) in infinite buffers (see Appendix A), we deliberately blur the distinction between the two in general discussions. This conveys the motivation for the analysis much more effectively. However, all details

of our analysis are accompanied by technically precise description and no blurring of the distinction between loss and overflow is employed in any derivations. With the exception of the discussion in subsection 3.1, we concentrate on *Overflow Probability* $P(Q > x)$ rather than loss probability, $P_{loss}(x)$, as our key performance indicator.

3 Review of asymptotic results

Here we review existing performance analyses of Poisson-Pareto queues in four asymptotic regimes: (i) as buffer thresholds become larger and larger, with traffic and server rate fixed; (ii) as the number of sources becomes larger and larger, with the buffer thresholds and server speed increasing in proportion; (iii) as the number of sources becomes larger and larger, with buffer thresholds increasing in size in proportion to the square root of the number of sources; and (iv) as the number of sources becomes larger and larger, approaching the level where the server is fully occupied, while server speed remains fixed but buffer thresholds are scaled. The first two of these asymptotic regimes apply with the considered probabilities approaching zero, and Large Deviations theory therefore supplies an effective analysis method. The third of the asymptotic approximations applies to systems with overflow probabilities tending to a finite positive limit other than zero and so Large Deviations theory is not applicable. The CLT applies in this case. This case has also been described as a heavy traffic approximation in [7,8], which name is justified by the fact that as λ increases, and the other system parameters are adjusted so that the probability of buffer overflow converges to a constant, system *utilization* tends to 1. The last case is a more traditional heavy traffic approximation in which the overflow probability for a fixed buffer threshold approaches 1, but by introducing a scaling of buffer thresholds the shape of the probability distribution of the overflow probability is determined for heavy traffic.

3.1 The large buffer estimate

Upper and lower power-law bounds for the loss probability, $P_{loss}(x)$, in a single server queue with PPBP input process were obtained in [26]. The decay coefficient of the tail is shown to be identical for both upper and lower bounds. Large Deviations Theory was used in [22] to obtain a consistent result for overflow probabilities. Related results have been obtained in [15, 18].

The upper bound for the loss probability from [26] is

$$P_{loss}(x) \leq \frac{a(\lambda,C)\left(\lambda\gamma\delta^{\gamma}(\gamma-1)^{-\gamma}\left(\frac{C}{r}+2\right)^{\gamma-1}r^{\gamma-1}\right)^{k}x^{(-\gamma+1)k}}{\lambda E(d)k!} \tag{2}$$

where

$$a(\lambda,C) = e^{2\lambda} + e^{(\lambda E(d)-\lambda)}\frac{[\lambda E(d)-\lambda]^{C+1}}{(1+C)!}. \tag{3}$$

The lower bound from [26] is

$$P_{loss}(x) \geq \frac{\gamma^{k}\delta^{\gamma k}r^{(\gamma-1)k}x^{(-\gamma+1)k}}{\lambda r E(d)\gamma(\gamma-1)^{k}\left(E(d)+(1-e^{-\rho^{*}/E(d)})^{-1}-1\right)^{\gamma+k}}, \tag{4}$$

where $E(d)$ is the mean burst duration. The parameter k is given by $k = 1 + \left\lfloor \frac{C}{r} - \lambda E(d) \right\rfloor$. Finally, the value of ρ^* depends upon $\lambda E(d)$. If $\lambda E(d) \leq 1$ then $\rho^* = \lambda E(d)$, otherwise, ρ^* may be any value in the range

$$0 \leq \rho^* < \begin{cases} 1 + \delta_p - \Delta, & \text{if } \Delta \geq \delta_p, \\ \delta_p - \Delta, & \text{if } \Delta < \delta_p. \end{cases} \tag{5}$$

The terms δ_p and Δ introduced in this definition for ρ^* are given by $\delta_p = \lambda E(d) - \lfloor \lambda E(d) \rfloor$, and $\Delta = \frac{C}{r} - \left\lfloor \frac{C}{r} \right\rfloor$.

Combining (32) (from Appendix A) with (4) gives a lower bound for the queue level distribution in the infinite queue which is of the form

$$P(Q > x) \geq W x^{(-\gamma+1)k}, \tag{6}$$

where W is a value constant with respect to the buffer size, x. We cannot make the same assertion with regard to an upper bound for $P(Q > x)$.

The fact that the upper and lower bounds, (2) and (4), decay at the same power-law rate gives us some confidence that the true "asymptotic shape" of the CDF has been identified. This form is also supported by a result which includes both shape *and weight* for the asymptotic form of the PPBP overflow probability, as given in [15] and also, more recently, in [6]. That is to say, in these papers a function is given explicitly which is neither an upper nor a lower bound, but whose ratio to the overflow probability tends to one as the buffer threshold tends to infinity. Another somewhat simpler expression for the weight of the tail is independently derived in Subsection 6.1.

In Subsection 5.1, it is shown that bounds of the form of (2) and (4) must inevitably diverge from each other as $\lambda \longrightarrow \infty$. A plot of the ratio of these bounds which confirms this result in a specific example is provided in Figure 3 in that Subsection.

3.2 Many sources large deviations estimate

Large Deviations Theory has been applied to the problem under study with the asymptotic regime considered having buffer threshold growing linearly with the number of sources, n, by a number of authors [5, 6, 10, 17, 19, 32]. The ground for this approach was laid in [33] and the mathematical framework by which these results can be obtained is also presented in [45].

The results obtained in this work take the form [10, Eq (4)], [19]

$$P\left(Q^{\{n\}} > nx\right) \approx \begin{cases} e^{-nI(x)} & n \text{ large} \\ e^{-n\varepsilon v(x)} & n \text{ and } x \text{ large} \end{cases} \tag{7}$$

in which $Q^{\{n\}}$ denotes the buffer level in a system with n times the intensity of arrivals, ε is a constant (denoted by δ in [10]), $I(x)$ is the *shape function* which depends upon the burst length distribution, $v(x)$ is $\ln G_e(x)$ where $G_e(x) = (1 - G(x))/M$, and M is the mean of G, which is the distribution of burst lengths. In the present instance, $v(x) = k \ln(x)$, for some constant k and the shape function, $I(x) \approx \varepsilon v(x)$ for large x, and so we obtain again a result in which

$$P\left(Q^{\{n\}} > nx\right) \approx x^{-k_1} \tag{8}$$

for a certain constant k_1.

The many sources asymptote is numerically evaluated in Subsection 6.3, where we shall see that although it is able to provide much better accuracy for buffer thresholds near zero, this accuracy very quickly evaporates as larger thresholds are considered.

3.3 The Heavy Traffic Limits

The performance of a PPBP single server queue can be modelled by a Gaussian process with the same mean and autocovariance [30]. As shown in [8, 39], for any PPBP, if the intensity of the process is increased while maintaining the net mean (the mean arriving work per second minus the server rate) and autocovariance unchanged, the stationary buffer distribution will tend to the Gaussian result. A consistent result was obtained in [7] without explicitly showing that the traffic process converges to a Guassian process. This result will be demonstrated in a numerical experiment in Subsection 6.2.

This approach to modelling a PPBP queue can be described as a heavy traffic approximation because if we take any PPBP and increase λ (the intensity of burst arrivals), the PPBP will tend to a Gaussian process and if we rescale the server *and* buffer thresholds in such a way that first and second order statistics of the net input process are preserved, utilization will tend towards 1 as $\lambda \longrightarrow \infty$. But this is not the only way in which we can rescale a PPBP queue so that as utilization tends to 1, the distribution of buffer levels tends to a limit. Another approach, used in [27], is to keep the server speed constant and rescale buffer thresholds.

In [27], the distribution of rescaled buffer thresholds approaches a limit which can be expressed in terms of the Mittag-Leffler special function. Because the Mittag-Leffler special function is asymptotically similar to x^{-1} as $x \longrightarrow \infty$, the distribution is shown to take the form $\sim \text{const} \times x^{1-\gamma}$ in this case, which is consistent with the large buffer asymptote discussed earlier and developed independently in §6.1.

The paper [27] also uses a light traffic asymptote, which provides an estimate of the probability of buffer emptiness which is accurate for light traffic, to complement the heavy traffic approximation and thereby obtain a result which is potentially accurate for a full, or at least a much wider, range of system parameters. However, the results rely on the assumption either that burst lengths have finite variance, or that one additional burst, above the mean load, is sufficient to overload the server.

4 The Quasi-Stationary (QS) approximation

In this section we first describe (in Subsection 4.1) the approach of [2] for performance evaluation of queues with PPBP input. In order to compare the QS approximation to the large buffer asymptote in a region where the two become similar it has been necessary to find algorithms for both methods which are accurate for the logarithm of extremely low probabilities (as low as 10^{-200} – see Figure 8). This has placed quite severe demands upon the design of the QS algorithm which has therefore been developed in the following ways: (i) the optimization task on which the algorithm is based has been explored in detail in order to improve understanding of how the algorithm converges, and thereby improving its accuracy and robustness (in Subsection 4.2); (ii) a more rigorous derivation has now

been developed, which shows that the QS algorithm provides a lower bound (in Subsection 4.3); and (iii) the accuracy with which logarithms of very small probabilities can be computed by the algorithm has been enhanced considerably by using large deviations based approximations for logarithms of some of the component probability formulae (in Subsection 4.4).

4.1 The Quasi Stationary algorithm

The Quasi-stationary algorithm makes use of an idea which was used in [19] to find the rate function for a large deviations characterization of multi-source heavy-tailed on-off traffic as the number of sources increases. This idea is to separate the bursts of the PPBP into long and short bursts. If we consider the PPBP over a finite interval of length W, i.e., the period $[t, t+W]$, for arbitrary t, then any burst which last for the entire time period, we label as a *long burst*. All other bursts are called *short bursts*.

This separation into short and long bursts leads to the formula:

$$\ln P\{Q > x\} \geq \max \begin{cases} \sup_{\eta \geq 1, \tau \geq 0} \{\ln P(\ell b[\eta, \tau]) + \ln P(S_\tau(-\tau, 0) > C\tau + x - r\eta\tau)\}, \\ \sup_{\tau \geq 0} \ln P(S_\infty(-\tau, 0) > C\tau + x), \end{cases} \tag{9}$$

where $\ell b[\eta, \tau]$ denotes the event that η or more bursts which began before $-\tau$ have continued to the present time, and $S_\tau(-\tau, 0)$ denotes the traffic contributed by bursts of length less than τ during the interval $(-\tau, 0)$. This formula will be derived in Subsection 4.3, but before we undertake that derivation, let us *illustrate* the concept by plotting the most likely configuration of long bursts which give rise to specified overflow states, together with the evaluation of the approximation (9).

4.2 Demonstration of the quasi-stationary approximation

Examples of the application of this algorithm are shown in Figures 1 and 2. The parameters which remain fixed in all of the systems under study in these diagram are: $\delta = 1$, $\gamma = 1.3$, $r = 0.2$. The parameters which vary from one diagram to the next are arrival rate of bursts (λ) and the net mean input, measured in units of the standard deviation of the traffic.

The net mean input is $-0.9\sigma_1$ in the first example and $-3\sigma_1$ in the second example, where σ_t denotes the standard deviation of the quantity of work (traffic) delivered in an interval of time t. In the first example the net mean input of the system is $-2.54558r$, and in the second this number is $-12r$. The number of long bursts which are sufficient to cause overload also depends upon the *length* of these long bursts. When the length, W, of a long burst is lower, the mean traffic contributed by the short burst process is reduced (because we have redefined what it means to be a short burst), and so a larger number of long bursts are required to cause overload. This explains why, in the example shown in Figure 2, the most likely number of long bursts associated with an overflow exhibits a peak before it falls to its asymptotic limit.

The number of long bursts most likely to cause overload is critical to the observed behaviour of each system. The algorithm estimates this quantity by finding the most likely compound event, made up of a certain number of long bursts, of a certain length, in combination with a certain quantity of

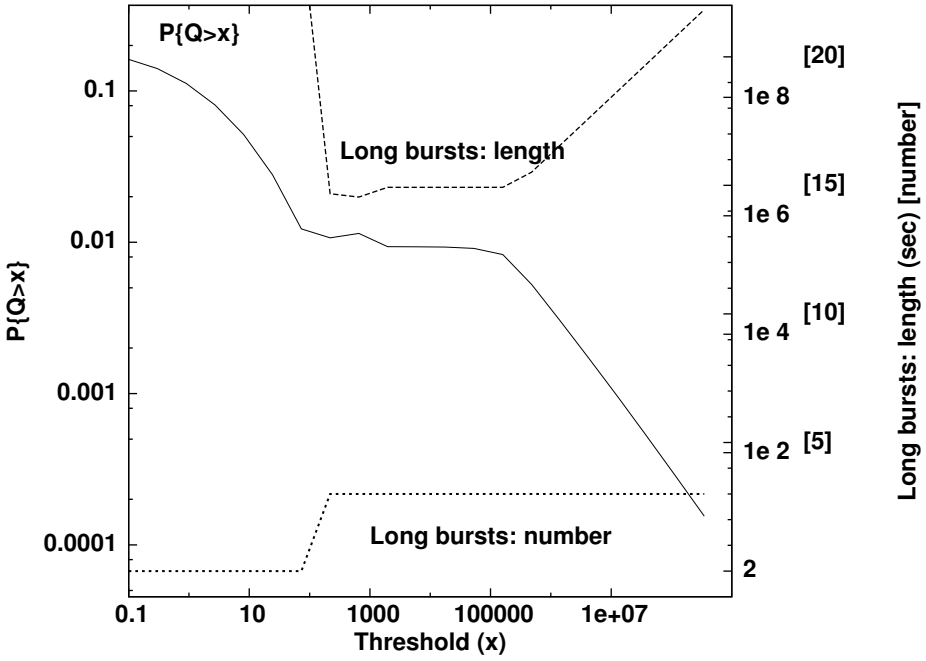

Fig. 1. P(buffer level $> x$), the most likely number of long bursts when buffer level $> x$, and the most likely long burst length [in seconds] when this occurs, when $\lambda = 2$, net mean input$= -0.9\sigma_1 = -0.509117 = -2.54558r$, in which σ_1 denotes the standard deviation of the number of bytes arriving in an interval of length 1.

traffic from the short bursts, to cause an overflow. In each case, for low values of x the most likely number of long bursts to cause overload is low, at or near zero. As x grows, the most likely number of long bursts that will cause overflow climbs, and its maximum may significantly exceed the smallest number of long bursts which drives the system into overload, but then drops back to the smallest integer value larger than $-1\times$ the net mean input divided by r, i.e. the smallest number sufficient to drive the system into overload.

In these figures, the stationary queue distribution clearly exhibits the power-law tail (which is characterized, on a log-log graph, by appearing as a straight line). In the case presented in Figure 1, 3 sufficiently long bursts will overload the system. In this case, the power-law tail appears to emerge from the point $x = 1,000,000$, by which time the overflow probability has dropped to below 0.01. In the second example, shown in Figure 2, 13 sufficiently long bursts will overload the system. In this case, the power-law tail appears to emerge from the point $x = 1,000$, by which time the overflow probability has dropped to below 10^{-10}.

In both these examples, the most likely *length* of the long bursts involved in a congestion event increases approximately linearly with x (buffer threshold) once the point has been reached where the tail behaviour of the system has set in, which occurs when the number of bursts most likely to cause the congestion event has achieved its limiting value.

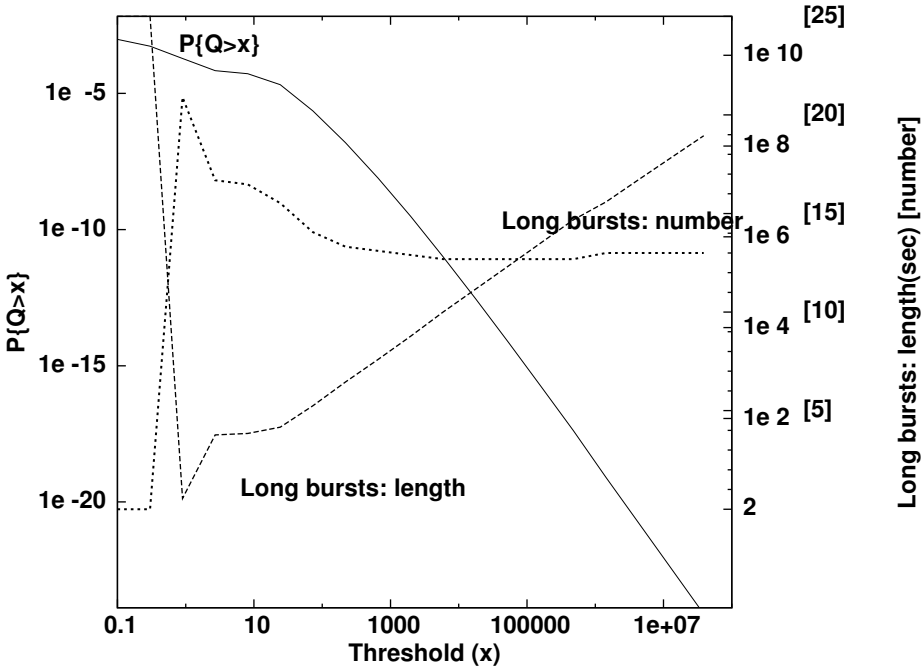

Fig. 2. P(buffer level $> x$), the most likely number of long bursts when buffer level $> x$, and the most likely long burst length [in seconds] when this occurs, when $\lambda = 4$, net mean input$= -3\sigma_\delta = -2.4 = -12r$, in which σ_1 denotes the standard deviation of the number of bytes arriving in an interval of length 1.

4.3 Justification of the quasi-stationary approximation

Whenever a buffer overflow occurs, the long bursts associated with this event can be *uniquely* identified as follows:

(i) trace back the evolution of the traffic and buffer process from now, τ_0 say, to the last time when the buffer was empty, $\tau_1 < \tau_0$ say;
(ii) identify any bursts which were continuously active during this entire period of time, from τ_1 to τ_0; let us say that the number of these bursts is η;
(iii) trace back the evolution of the traffic and buffer process, observing the long bursts, till the *start* of one of these bursts is observed, call this $\tau_2 < \tau_1$.

In this way we can find a more complex event, in which (a) at least $\eta - 1$ bursts are simultaneously active; (b) a new burst arrives, at τ_2; (c) while all η of these bursts continue, a busy period starts, at time τ_1; (d) while all η of these bursts continue, an overflow occurs (the buffer exceeds level x), at time τ_0.

Let us denote this compound event by $\ell b[x, \eta, \tau_0, \tau_1, \tau_2]$.

As we have already observed, all exceedance events ($\{Q > x\}$) correspond to one of these events, and conversely, so if we can determine the probability of this type of event we can also determine the

overflow probability $P(Q > x)$. The stationary overflow probability is therefore

$$P\{Q > x\} = \sum_{\eta=0}^{\infty} P\{\text{for some } \tau_1, \tau_2, \ell b[x, \eta, 0, \tau_1, \tau_2]\}. \tag{10}$$

During an overflow event, the long bursts provide a constant load on the server, so the buffer and the remaining traffic (the *short burst traffic*) is identical to a different model in which the server has reduced capacity, and the traffic has only short bursts.

If the largest term in the sum at (10) is much larger than the remaining terms, and numerical experiments confirm this appears to be frequently the case, the bound:

$$P\{Q > x\} \geq \sup_{\eta \geq 0, \eta \in \mathbf{Z}, \tau' \geq \tau \geq 0} P\{\ell b[x, \eta, 0, -\tau, -\tau']\}. \tag{11}$$

will be close to the true value of $P\{Q > x\}$.

Moreover, so long as $\eta > 0$, the probability of a long burst event decreases with increasing τ' – there is no point in the long bursts being longer than necessary, so except when $\eta = 0$, the supremum in (11) will always occur when $\tau' = \tau$. On the other hand, when $\eta = 0$ the optimal value for τ' is ∞ because the only effect of τ' in this case is to define which component of the traffic is regarded as the short bursts. This produces the somewhat simpler lower bound:

$$P\{Q > x\} \geq \max \begin{cases} \sup_{\eta \geq 1, \eta \in \mathbf{Z}, \tau \geq 0} P\{\ell b[x, \eta, 0, -\tau, -\tau]\}, \\ \sup_{\tau \geq 0} P\{\ell b[x, 0, 0, -\tau, -\infty]\}. \end{cases} \tag{12}$$

For simplicity of calculation it will be useful to separate the long burst aspect of a long burst overflow event from the exceedance aspect.

When we take this step of seeking the largest term of the sum at (10), and therefore replacing (10) by (12), we might as well also slightly alter the definition of the long-burst overflow event that we seek to include any cases where η *or more* bursts occur during the interval of time τ. This will still be a lower bound because the event where η or more long bursts occur and the short burst process is sufficient to cause an overflow to occur even if only η long bursts occur is still a sub-event of the event where an overflow occurs. Let us now define $\ell b[\eta, \tau]$ to be the event in which η or more bursts have been consistently active for at least time τ and let us denote the *short burst traffic* by $S_\tau(t_1, t_2)$, i.e. this is the quantity of traffic made up of bursts starting after $-\tau$ or finishing before 0 in the interval (t_1, t_2). This leads to

$$P\{Q > x\} \geq \max \begin{cases} \sup_{\eta \geq 1, \tau \geq 0} P\{\ell b[\eta, \tau] \& S_\tau(-\tau, 0) > C\tau + x - r\eta\tau\}, \\ \sup_{\tau \geq 0} P\{S_\infty(-\tau, 0) > C\tau + x\}. \end{cases} \tag{13}$$

Since the short and long burst events are conditionally independent, given a specific choice of η and τ, taking logs gives (9).

The overflow probability for the short-burst process could be calculated from the exact probability distribution, which is compound Poisson, however because the distribution of each burst is not heavy-tailed (except in the case where $\eta = 0$) this compound Poisson distribution can be expected to be well

approximated by a Gaussian distribution. Even in the case $\eta = 0$, it is appropriate to use a Gaussian estimate because we need an approximation which is accurate for moderate deviations from the mean; large deviations will necessarily involve $\eta > 0$.

In order to calculate probabilities associated with the short bursts we will need to make use of their variance, which can be found easily once we have an expression for the variance of the arriving bytes in a PPBP. This was derived in [2], but because of its importance we reproduce it here:

$$\sigma_t^2 = \sigma_t^2(\lambda,\gamma,\delta,r) = \begin{cases} 2r^2\lambda t^2\left(\frac{\delta\gamma}{2(\gamma-1)} - \frac{t}{6}\right), & 0 \le t \le \delta, \\ 2r^2\lambda\left\{\frac{\delta^3\gamma}{6(3-\gamma)} - \frac{\delta^2 t\gamma}{2(2-\gamma)} - \frac{t^{3-\gamma}\delta^\gamma}{(1-\gamma)(2-\gamma)(3-\gamma)}\right\}, & t > \delta. \end{cases} \tag{14}$$

The number of long bursts is Poisson distributed with mean, β, equal to λ (the burst intensity) times the probability that the backward recurrence time of the Pareto distribution of burst lengths is longer than τ, the nominated length of a long burst, i.e.

$$\beta = \frac{\lambda\tau^{1-\gamma}}{\delta(\gamma-1)}.$$

4.4 Accurate calculation of logarithms of small probabilities

Since (9) provides an estimate for $\ln P(Q > x)$ in terms of logarithms of probabilities we can reduce numerical error in the evaluation of this formula, when the probabilities are small, by working exclusively in logarithms of probabilities. If the number of long bursts which are to have occurred is very large (6 standard deviations more than the mean), in order to obtain satisfactory accuracy, we should use a large deviations estimate for the logarithm of its probability [46, p10]:

$$\ln P\{\text{long bursts} \ge \eta\} \sim -\eta(\ln\eta - \ln\beta) - (\beta - \eta). \tag{15}$$

Similarly, when computing logarithms of the Normal distribution, when the standard score is larger than 5, the formula from [47], i.e.

$$\ln P\{\text{bursts} \ge \eta\} = \ln P\left\{Z > \frac{\eta-\beta}{\sqrt{\beta}}\right\} \sim -\frac{1}{2}\left(\frac{\eta-\beta}{\sqrt{\beta}}\right)^2 - \ln\left(\frac{\eta-\beta}{\sqrt{\beta}}\right) - \ln 2, \tag{16}$$

has been used.

More details of the QS approximation are provided in [48]. The Mathematica code which has been used to compute the QS approximation is included in [49].

5 Inherent limitation of power-law approximations

In the limits as $\lambda \to \infty$ considered in this section, we consistently refer to a sequence of PPBP queueing systems, $\{S_\lambda\}_{\lambda>0}$ say, which represents the natural outcome of growth in the traffic, to-

gether with an increase in the server speed and buffer capacity chosen so that queue level distribution, appropriately scaled, converges to a certain limit.

The sequence $\{S_\lambda\}_{\lambda>0}$ maintains the burst length distribution parameters, γ and δ, as fixed values but the parameters r_λ (the rate of each burst) and C_λ (the server rate) change with λ as follows:

$$\begin{aligned} r_\lambda &= r_1/\sqrt{\lambda}, \\ C_\lambda &= C_1 + \frac{(r_\lambda\lambda - r_1)\delta\gamma}{(\gamma-1)} = C_1 + \frac{r_1(\sqrt{\lambda}-1)\delta\gamma}{(\gamma-1)}. \end{aligned} \tag{17}$$

This rescaling of PPBP systems undergoing linear growth of burst arrival rate, with constant burst characteristics, has been chosen so that the net mean and autocovariance of the input process are the same for all λ. The stationary cumulative distribution function of the buffer level in this system, when the traffic intensity is λ, is denoted by ϕ_λ.

In Subsection 5.1 we show that for any rule which provides a pair of upper and lower power-law bounds on the stationary PPBP CDF over a fixed region of the form $[x_0,\infty)$, the ratio of the bounds grows without bound as λ increases. In addition to the unboundedness of the ratio of the upper and lower bounds, we show that the worst value of the ratio of the upper bound to $\phi_\lambda(x)$ is unbounded over (x_0,∞) as $\lambda \longrightarrow \infty$ and the same applies to the ratio of $\phi_\lambda(x)$ to a power-law lower bound. Finally, we use these results to show that an exact power-law asymptote, $A_\lambda x^{-f(\lambda)}$, for $\phi_\lambda(x)$ cannot be uniform in λ, i.e. if $\phi_\lambda(x)/A_\lambda x^{-f(\lambda)} \longrightarrow 1$ as $x \longrightarrow \infty$, then for any $x_0 > 0$, either $\sup_{x>x_0} \frac{A_\lambda x^{-f(\lambda)}}{\phi_\lambda(x)} \to \infty$ as $\lambda \to \infty$ or $\inf_{x>x_0} \frac{A_\lambda x^{-f(\lambda)}}{\phi_\lambda(x)} \to 0$ as $\lambda \to \infty$.

It is well understood in the literature, and we saw in Subsection 4.2, that power-law behaviour of the tail of the PPBP CDF comes into play when the overflow events under consideration are caused by a group of long bursts in association with short burst traffic arriving *at its expected rate*. In Subsection 5.2 we estimate how large buffers need to be in order that the overflow is most likely to occur in this way, rather than by the long bursts and short bursts acting in concert. It is shown in Subsection 5.2 that this threshold of power-law tail behaviour is unbounded, as a function of the burst arrival rate (λ), even for modest values of λ.

5.1 Unbounded separation of bounds for large λ

In the scaling (17), the stationary waiting time CDF for a single server queue fed by a PPBP converges weakly, as the intensity of the Poisson process, λ, increases, to the waiting time CDF of the Gaussian queueing system with the same net mean input and autocovariance [8, 39]. The fact that the Gaussian system has a Weibull tail [20] appears to contradict the power-law tails of individual functions making up the limit, however this is not necessarily a contradiction because as $\lambda \longrightarrow \infty$, the remoteness of the power-law tails may increase. Proposition 1 confirms that this must occur.

Proposition 1 *For any functions, A_λ, B_λ of λ, and any increasing function, $f(\lambda)$, defined on $[0,\infty)$, such that*

$$A_\lambda x^{-f(\lambda)} \le \phi_\lambda(x) \le B_\lambda x^{-f(\lambda)}, \tag{18}$$

for all $x > x_0$, necessarily, $\frac{B_\lambda}{A_\lambda} \to \infty$ as $\lambda \to \infty$.

Note that the proposition applies for any increasing function, $f(\lambda)$, however the case of most interest is where the LHS and RHS of (18) tend, in the sense of a ratio for fixed λ, as $x \longrightarrow \infty$, to $\phi_\lambda(x)$. In this case, we shall see below, in Equation (30) of Section 6.1, that $f(\lambda) = \eta_1(\lambda)(1-\gamma)$ where $\eta_1(\lambda)$ is the smallest number of bursts which exceeds the net capacity of the server, after the mean load from the arriving traffic is subtracted, i.e.

$$\eta_1(\lambda) = 1 + \left\lfloor \frac{C}{r} - \frac{\lambda\gamma}{\delta(\gamma-1)} \right\rfloor. \tag{19}$$

The proofs of this proposition and the following two propositions are given in Appendix B.

The ratio between the upper and lower power-law bounds from [26] presented earlier, in Equations (2) and (4), is plotted as a function of λ in Figure 3. The parameters in this example are $\gamma = 1.5$, $\delta = 1$, $r = 1$, net mean input $= -2 \times \sigma_1$ and the buffer level where the bounds are evaluated is 100. The ratio increases extremely quickly as a function of λ, as predicted by Proposition 1. Proposition 1 shows that the behaviour shown is not dependent on the specific bounds used, and that any pair of upper and lower bounds proposed for this system will grow further apart as λ increases. The ratio between the bounds appears to be discontinuous, an explanation for which is provided in Subsection 6.1.

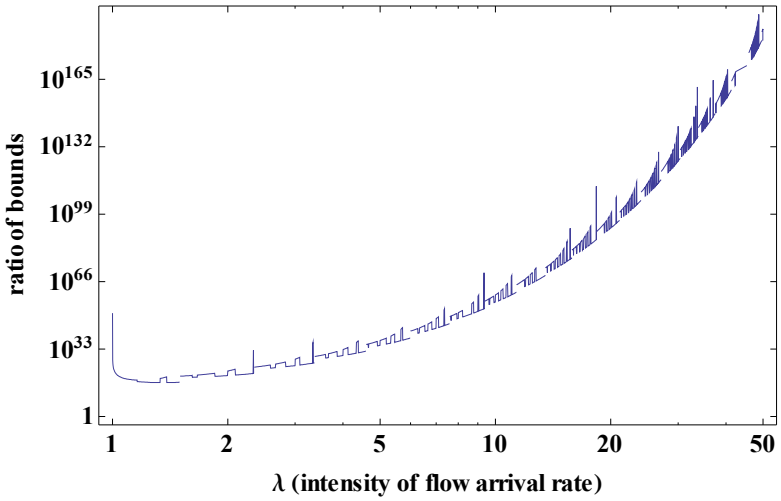

Fig. 3. The ratio between upper and lower bounds obtained by Tsybakov and Georganas.

Proposition 1 leaves open the possibility, for a rule for assigning upper and lower bounds, that although the bounds become increasingly far apart, one of the bounds, for example the upper bound, is nevertheless a good approximation uniformly in λ. The next proposition shows that this cannot occur.

Proposition 2 *In the same context as Proposition 1*

For any mapping, $\lambda \mapsto B_\lambda$, (or $\lambda \mapsto A_\lambda$) and function, $f(\lambda)$, defined on $[0,\infty)$, such that

$$\phi_\lambda(x) \leq B_\lambda x^{-f(\lambda)} \tag{20}$$

$$\left(A_\lambda x^{-f(\lambda)} \leq \phi_\lambda(x)\right), \tag{21}$$

for all $x > x_0$, necessarily, for any $\bar{x}_0 > 0$, $\sup_{x>\bar{x}_0} \frac{B_\lambda x^{-f(\lambda)}}{\phi_\lambda(x)} \to \infty$ ($\inf_{x>\bar{x}_0} \frac{\phi_\lambda(x)}{A_\lambda x^{-f(\lambda)}} \to 0$) as $\lambda \to \infty$.

Proposition 3 *In the same context as Proposition 1, if $A_\lambda x^{-f(\lambda)}$ is a power-law asymptote for a PPBP queue, in the sense that for any PPBP system, $\frac{A_\lambda x^{-f(\lambda)}}{\phi_\lambda(x)} \longrightarrow 1$ as $x \longrightarrow \infty$, the convergence is not uniform in λ. For any $x_0 > 0$, either $\sup_{x>x_0} \frac{A_\lambda x^{-f(\lambda)}}{\phi_\lambda(x_0)} \longrightarrow \infty$ or $\inf_{x>x_0} \frac{A_\lambda x^{-f(\lambda)}}{\phi_\lambda(x_0)} \longrightarrow 0$ as $\lambda \longrightarrow \infty$.*

5.2 The boundary where the tail becomes power-law

For a given buffer threshold, x, let $\kappa(x)$ denote the number of simultaneous long bursts which is most likely to be present during the busy period leading up to the event that level x is exceeded. The figures in Subsection 4.2 show how $\kappa(x)$ varies with x, and as we would expect, this number approaches η_1 (see (19)), the number of extra bursts required to cause system overload, as $x \longrightarrow \infty$. But how large does x have to be in order that $\kappa(y) = \eta_1$ for all $y \geq x$?

In a system with a very low arrival rate of bursts, but high utilization, so that for example one burst is sufficient to fully load the server, power law behaviour may apply for all $x \geq 0$. However, such systems are not relevant to the question we wish to pursue here, so we assume henceforth in this subsection that the utilization level of our system is sufficiently moderate and the burst arrival rate is sufficiently high that power-law queueing behaviour is not exhibited for all x.

We now use the QS approximation, but with a further simplification which is justified when x is large. In the QS approximation, the variance of the short-burst traffic was reduced to take into account the absence of the long bursts. In this simplified model, with the sole purpose of finding the buffer threshold beyond which the power-law asymptote applies, we ignore this effect, and assume that the short burst traffic variance has the same variance-time curve as the complete traffic, no matter what value we consider separates the short from the long bursts.

The variance due to these longer bursts is quite small anyway. In addition, we shall soon see that with this simplification, there is a unique first location where the likelihood of an overflow event is a maximum when the number of long bursts equals η_1.

Let x_h denote the threshold level where the overflow probability begins to follow a power law as a function of x. We seek to estimate x_h by the characteristic feature that $\kappa(x) = \eta_1$ for $x \geq x_h$ but $\kappa(x_h-) < \eta_1$.

Define event $A_k \equiv \{$the number of long bursts is $k < \eta_1\}$, and event $B \equiv \{$there are precisely η_1 long bursts$\}$. Let us now develop formulae for $P(Q > x \cap A_k)$ and $P(Q > x \cap B)$, respectively. The long burst boundary is denoted by $t_k^*(x)$ in both cases, with $k < \eta_1$ in the first case and η_1 in the second.

Evaluation of $P(Q > x \cap A_k)$

In this case, the mean input to the system will be $m_1 = m + kr$, below the server rate, C, and in order for the level x to be exceeded, the short bursts must contribute work to the system by an amount

in excess of their normal rate. From [30, (1.1)], in accordance with a principle established earlier in [20], the most likely way in which this exceedance event $Q > x$ will occur is that the short bursts in aggregate contribute at above their usual rate over a period of duration $t^*(x)$ which can be estimated by approximating the PBPP variance-time curve by that of FBM to be

$$t_k^*(x) \approx \frac{Hx}{(1-H)(C-m_1)} = \frac{Hx}{(1-H)(C-m-kr)}, \qquad x \geq 0. \tag{22}$$

Because it will almost certainly take this long for the short bursts to make their contribution to the overflow event, the long burst boundary must be at least as large as $t_k^*(x)$.

Since the threshold x could be exceeded, also, by a combination of short bursts and k long bursts in which the boundary between long and short bursts was *less* than $t_k^*(x)$, this formula provides a lower bound for $P(Q > x \cap A_k)$. In the case where $k = \eta_1 - 1$, and especially when $\eta_1 r - C \ll 1$, a choice of a much shorter long burst boundary might produce a much more likely event. Also, any inaccuracy in estimation of $t^*(x)$ due to our adopting the FBM variance-time curve instead of the correct PPBP time curve, when obtaining the formula for $t_k^*(x)$, will merely produce a lower estimate of the probability of an overflow event due to k long bursts, and therefore a lower estimate of the power-law threshold.

In this situation, the probability of there being k long bursts will be

$$\frac{\left(\frac{\mathrm{E}(d)\lambda(t_k^*(x)/\delta)^{1-\gamma}}{\gamma}\right)^k e^{\frac{-\mathrm{E}(d)\lambda(t_k^*(x)/\delta)^{1-\gamma}}{\gamma}}}{k!} \tag{23}$$

and we can estimate the probability that the short bursts contribute sufficient to combine with the long bursts to exceed the level x as the probability of exceeding $x + (C - rk)t_k^*(x)$ from a Gaussian distribution with mean $\frac{\lambda r t_k^*(x)\gamma\delta}{\gamma - 1}$ and variance $\sigma^2_{t_k^*(x)}(\lambda, \gamma, \delta, r)$ as defined at (14). This Gaussian probability is therefore $\frac{1}{2}\operatorname{erfc}\left(\frac{x + \left(C - r\left(k + \frac{\lambda\gamma\delta}{\gamma - 1}\right)\right)t_k^*(x)}{\sqrt{2}\sigma_{t_k^*(x)}(\lambda,\gamma,\delta,r)}\right)$. The resulting estimate of the probability of exceeding x with k long bursts is therefore

$$P(Q > x \cap A_k) = \frac{1}{2}\frac{\left(\frac{\mathrm{E}(d)\lambda(t_k^*(x)/\delta)^{1-\gamma}}{\gamma}\right)^k e^{\frac{-\mathrm{E}(d)\lambda(t_k^*(x)/\delta)^{1-\gamma}}{\gamma}}}{k!} \operatorname{erfc}\left(\frac{x + \left(C - r\left(k + \frac{\lambda\gamma\delta}{\gamma - 1}\right)\right)t_k^*(x)}{\sqrt{2}\sigma_{t_k^*(x)}(\lambda, \gamma, \delta, r)}\right). \tag{24}$$

Evaluation of $P(Q > x \cap B)$

In this case, the mean input of the system is more than the capacity of the server and therefore the most likely way for the buffer to exceed x is that the short bursts contribute at their usual mean and the overflow will most probably occur by the buffer simply filling at the rate by which the server fails to meet the demand of the arriving traffic. This will lead to an event of duration

$$t_{\eta_1}^*(x) = \frac{x}{(m + \eta_1 r - C)}, \qquad x \geq 0. \tag{25}$$

The probability that threshold x is exceeded when there are precisely η_1 simultaneous long bursts

is approximately

$$P(Q > x \cap B) = \frac{\left(\frac{\mathrm{E}(d)\lambda(t^*_{\eta_1}(x)/\delta)^{1-\gamma}}{\gamma}\right)^{\eta_1} e^{-\frac{\mathrm{E}(d)\lambda(t^*_{\eta_1}(x)/\delta)^{1-\gamma}}{\gamma}}}{\eta_1!}. \tag{26}$$

A choice of boundary between long and short bursts larger than $t^*_{\eta_1}(x)$ will produce a lower probability, and a shorter boundary would not lead to level x being exceeded without a contribution from short bursts.

Plots of the overflow probability due to a variety of scenarios, calculated using (24) and (26) are shown in Figure 4. The true overflow probability is given by the maximum of all the curves displayed

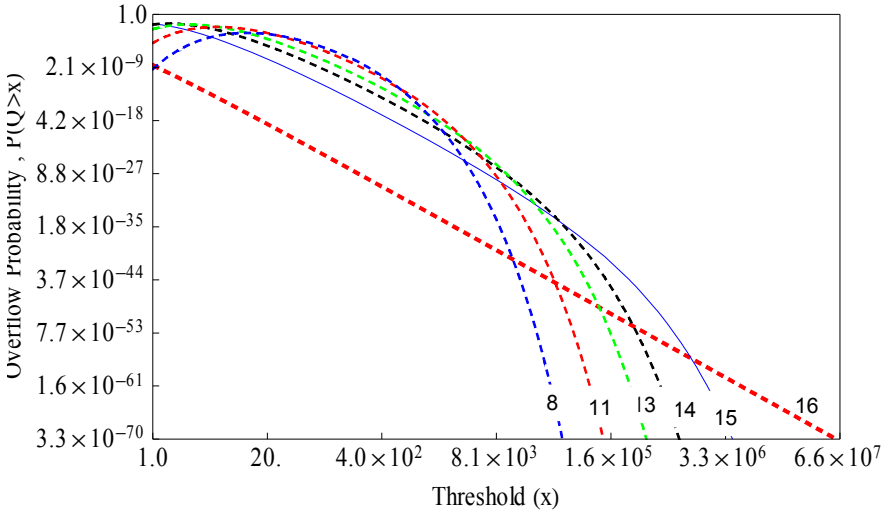

Fig. 4. Plot of the overflow probability under six scenarios: 16 long bursts, 15 long bursts with help from short bursts, 14, 13, 11 and 8 long bursts with help from short bursts, with $\lambda = 10.65$, $\delta = 1$, $r = 1$. In all cases $C = \frac{\lambda\gamma\delta r}{\gamma - 1} + 3\sigma_1$, where σ_1^2 is the variance of the number of bytes arriving in one second in the PPBP input process.

in Figure 4 and therefore the location where power-law behaviour starts, i.e. a straight line in this plot, will usually be where the two curves with the highest and the second highest number (k) of long bursts intersect.

It follows that usually a lower bound on the buffer threshold at which the power-law behaviour of the tail first sets in, x_h, is provided by the solution (for x) of the equation (24) = (26), in which $k = \eta_1 - 1$. In cases where the server is just a tiny bit too slow to fully deal with $\eta_1 - 1$ extra flows the most likely length of an overflow event, estimated just by considering the short flows, ie $t^*_{\eta_1 - 1}(x)$, is unrealistically long. In this situation it is preferable to solve (24) = (26) with $k = \eta_1 - 2$ to provide a lower bound on the power-law tail threshold.

Mathematica's secant method for solving equations was used to solve (24) = (26), with $k \leq \eta_1 - 1$. The maximum of the cases $k = \eta_1 - 1$ and $k = \eta_1 - 2$ has been used as an estimate of the threshold of power-law behaviour. The resulting estimate is shown in Figure 5.

The unusual saw-tooth shape of the curve in Figure 5, showing the dependence of the lower bound for x_h on λ, can be explained as follows. When the capacity of the server is only a *small* amount above an integral multiple of r (the rate of one burst) it is *much* more likely that an overflow event

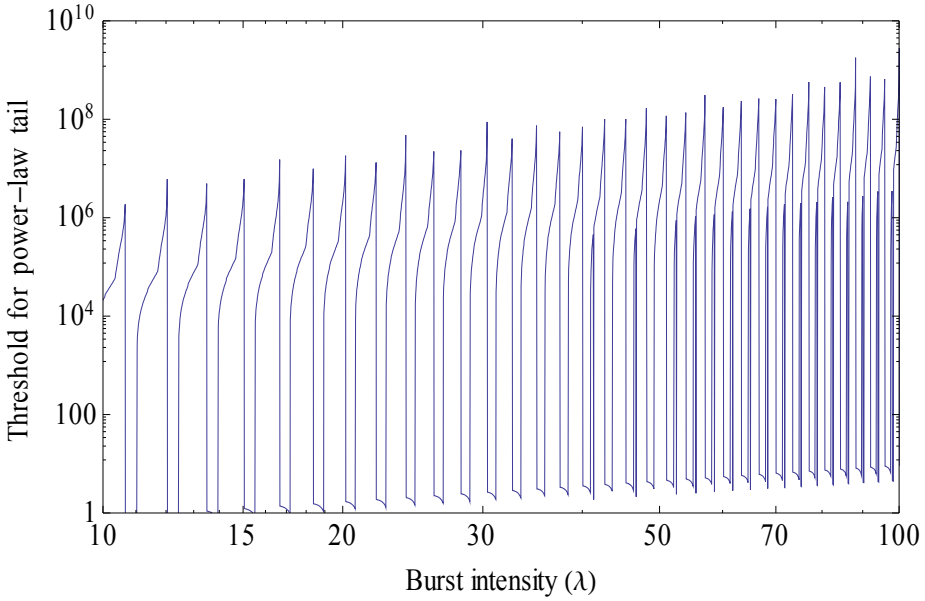

Fig. 5. Lower bound for the threshold for power-law behaviour of the tail, in Mbytes ($\delta = 1$ second, $r = 1$ Mbyte/s, $\gamma = 1.5$, $C = \frac{\lambda\gamma\delta r}{\gamma-1} + 3\sigma_1$, where σ_1^2 is the variance of the number of bytes arriving in one second in the PPBP input process), computed by solving (26)=(24) by the Secant method.

will occur by a joint contribution of long and short bursts – hence the threshold, x_h, where power-law behaviour begins, becomes very large. In addition, even if we only consider systems where the capacity of the server, C, exceeds the arriving traffic by nearly r, x_h increases steadily as $\lambda \longrightarrow \infty$. Because the estimate of x_h shown here is a *lower bound* on power-law behaviour, no significance should be given to the teeth projecting downward in Figure 5.

6 Comparison of approximations

In this section the four approximations for the stationary CDF of a queue with PPBP input are numerically evaluated over a very wide range of parameters and compared, showing explicitly how the asymptotic regimes relate to each other and to the QS approximation. In preparation for this, in Section 6.1, we derive formulae for weight and decay of the power-law asymptote. We see there that the weight, c_λ, of the power-law asymptote is highly sensitive to other system parameters. The QS approximation is compared to the Large buffer and Gaussian approximations in Subsection 6.2. The QS approximation and the Many sources Large Deviations result are compared in Subsection 6.3. The CLT asymptote appears to have better accuracy than either of the Large Deviations results in the central region of parameter space, however its accuracy is still unsatisfactory except for quite large numbers of sources (See Figure 8).

6.1 The weight of the power-law tail

Expressions for the weight of the power-law tail provided in [6, 15, 28] (the result in [28] is applicable for a slightly different context), are difficult to evaluate numerically, except in the case where one burst is sufficient to overload the server.

Using the analysis of Subsection 4.3, and in particular equation (15), we obtain the approximation, which applies for large x:

$$\ln P(Q_t > x) \sim -\eta_1 (\ln \eta_1 - \ln \beta) + (\eta_1 - \beta) \tag{27}$$

where $\eta_1 = 1 + \left\lfloor \frac{C}{r} - \frac{\lambda\gamma}{\delta(\gamma-1)} \right\rfloor$ is the least number of long bursts which cause an overflow of the buffer at threshold x when combined with the mean short burst load (we assume that the x is so large that virtually all bytes are in short bursts) and β is the mean number of long bursts (which will be defined more precisely in a moment).

The duration, τ, of this event leading to the overflow must stand in a simple relationship to the buffer threshold reached in the overflow, x, namely:

$$\tau = \frac{x}{r\eta_1 - C + \frac{\lambda r\gamma\delta}{\gamma-1}}, \tag{28}$$

because, when x is large, the short burst component of the probability to be maximised rapidly moves from zero to one as τ increases in the near vicinity of $x/(r\eta_1 - C + \frac{\lambda r\gamma\delta}{\gamma-1})$. Putting this more directly: the large bursts in combination with the short bursts are supplying work at a rate $r\eta_1 - C + \frac{\lambda r\gamma\delta}{\gamma-1}$ in excess of the servers capacity, during this overflow event. Therefore, the simplest, and most likely, way for the overflow to occur is for the buffer to start empty, and fill at this rate, until it reaches x. This will therefore happen over a period of time τ, as defined at (28).

From this we conclude that the "long bursts" are all the bursts longer than τ and therefore the mean number of long bursts is $\beta = \frac{\lambda\tau^{1-\gamma}}{\delta(\gamma-1)}$. From (27), therefore

$$\begin{aligned} \ln P(Q_t > x) &\sim -\eta_1 \left(\ln \eta_1 - \ln \left(\frac{\lambda\tau^{1-\gamma}}{\delta(\gamma-1)} \right) \right) + \eta_1 - \frac{\lambda\tau^{1-\gamma}}{\delta(\gamma-1)} \\ &\sim -\eta_1 \left(\ln \eta_1 - \ln \left(\frac{\lambda\tau^{1-\gamma}}{\delta(\gamma-1)} \right) \right) + \eta_1, \end{aligned} \tag{29}$$

because for large x, τ is also large and so $\ln(\tau^{1-\gamma})$ is large in magnitude, whereas $\tau^{1-\gamma} \to 0$.

Exponentiating both sides of (29) gives:

$$P(Q_t > x) \approx c_\lambda x^{\eta_1(1-\gamma)}, \tag{30}$$

where

$$c_\lambda = \left(\frac{\eta_1^{-1} e\lambda}{\delta(\gamma-1)} \right)^{\eta_1} \left(r\eta_1 - C + \frac{\lambda r\gamma\delta}{\gamma-1} \right)^{-\eta_1(1-\gamma)} \tag{31}$$

is the *weight* of the large buffer asymptote and the exponent of x, $\eta_1(1-\gamma)$, is the *decay*.

This asymptotic formula is consistent with known results as regards the *decay*; a formula for the *weight* of the tail has been provided in [6], however equivalence of these two formulae is not clear and numerical evaluation of the formula in [6] has not been demonstrated in [6] or attempted here. The context in [6] is significantly different from here, so deducing a specific form, from this result, for the weight of the tail which applies in the present case would be difficult. This formula is also consistent with [27] in regard to the decay but because the latter formula is a heavy-traffic approximation the *weight* of the tail is not comparable to the weight given here. The derivation of c_λ relies on an assumption that events in which a coincidence of long bursts overload the link *overlap* is sufficiently rare that its probability can be neglected. For sufficiently heavy traffic this is not the case, so the formula given here for c_λ cannot be expected to be consistent with the heavy traffic result in [27].

Plots of c_λ vs λ are shown in Figures 6 and 7. From these graphs it seems that the weight of the large buffer asymptote can vary between 0 and 10^{28} over a very small range of λ values. As λ increases from $r \times k$ below server capacity to $r \times (k-1)$ below server capacity, the weight of the tail associated with events where there are precisely k long bursts, gradually reduces to zero. The behaviour of c_λ shown in Figures 6 and 7 is an essential feature of this system, not a numerical aberration.

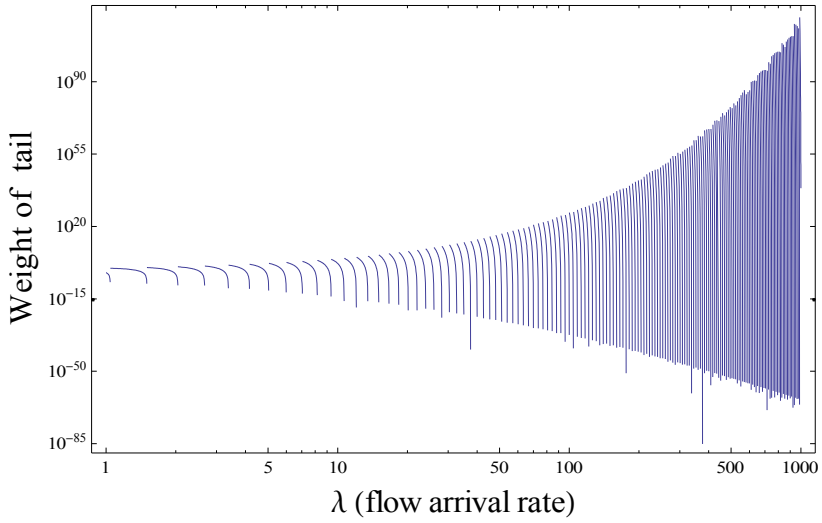

Fig. 6. Weight of Large Buffer Asymptote as a function of λ (broad view)

6.2 Comparison of the heavy traffic limit, large buffer limit and QS approximation

A series of similar PPBP queueing systems, with increasing burst rates, have been analysed by three methods (large buffer asymptote, heavy traffic Gaussian approximation, and QS approximation) and the results are shown in Figure 8. The parameters of the models under consideration in this figure are $\gamma = 1.5$, $\delta = 0.2$ [seconds], $r = 1$ [Mbit/sec] and $\lambda_v = 1$, 64, and 32000 [bursts/sec]. The server capacity, C_v, for each model has been adjusted so that the net mean, $C_v - \frac{\lambda_v r \delta}{\gamma - 1}$, and the autocorrelation of the traffic process is exactly the same, when expressed as a multiple of the standard deviation of the quantity of traffic arriving in an interval of length 1, for all the traffic models. In order to demonstrate convergence to the Gaussian case, the x-axis in Figure 8 adopts units of buffer level divided by $\sqrt{\lambda}$, rather than buffer level itself.

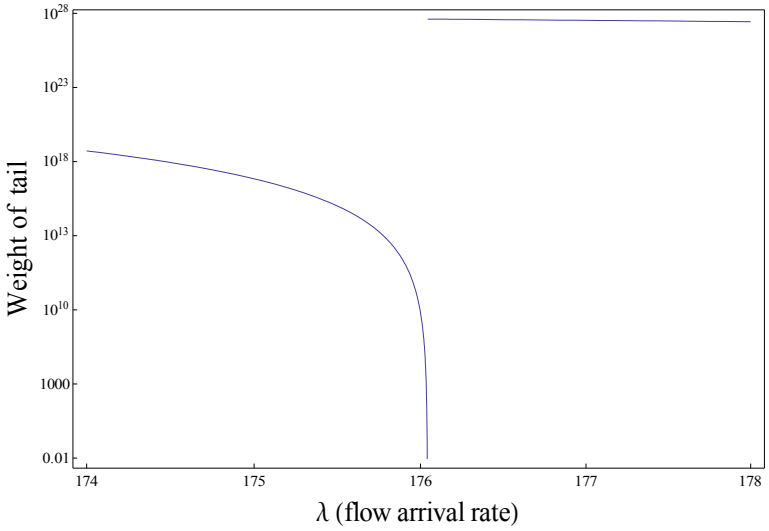

Fig. 7. Weight of Large Buffer Asymptote as a function of λ (close view)

The capacities of the three systems compared in Figure 8 have been set to $\frac{\lambda_k r \delta \gamma}{\gamma - 1} + \frac{3\sigma_\delta(\lambda_k, \gamma, \delta, r)}{\delta}$. The server rates which emerge from this calculation are: 2.79089, 55.9271, and 19591.9 Mbit/s.

Because Gaussian systems include, in general, significant amounts of "negative" traffic, the concept of system utilization is not meaningful for them. Instead, it is preferable to use the *net mean* input. For this reason, even for PPBP systems, the net mean input, expressed in standard deviations over a certain time period, is a more meaningful parameter than system utilization for characterizing system behaviour. However, the system utilization is nevertheless of interest, particularly so that we can observe the gain in system efficiency as traffic load increases. The utilizations of the three systems shown in Figure 8 are $\left(\frac{\lambda_k r \delta \gamma}{\gamma - 1}\right) / \left(\frac{\lambda_k r \delta \gamma}{\gamma - 1} + \frac{3\sigma_\delta(\lambda_k, \gamma, \delta, r)}{\delta}\right)$, $k = 1, 2, 3$, which turns out to be 21.5%, 69%, and 99%, respectively.

Figure 8 shows clearly that the quasi-stationary solution method is consistent with the Gaussian limit and with the large buffer limit. Although the large buffer asymptote appears to be an upper bound for the three choices of λ displayed in Figure 8, cases where the large buffer asymptote are not an upper bound have also been observed in other experiments. This graph is plotted using log-of-log scale for the Y axis and a log scale for the X axis. These scales enable us to see results over a very wide range. The horizontal axis extends to buffer sizes up to $10^{10} \times \sqrt{\lambda}$. The Y axis extends to probabilities as low as 10^{-1000}. This range has been used so that the relationship between the different asymptotic regimes can be seen clearly. Calculating estimates by the different methods over such a wide range of parameter values is quite challenging and cannot be achieved without special effort.

The figure shows clearly the convergence of the queueing model to a limiting behaviour of a queue fed by Gaussian noise, which is also shown in the figure. The Gaussian model, with exactly the same autocorrelation as this PPBP, has been analysed in [30]. The method derived in that paper has been used here to compute the stationary distribution.

Because these plots use a log-log y-scale and a log x-scale, a Weibull distribution would appear in this figure as a straight line and conversely any plots in this figure which are asymptotically linear correspond to distributions which have a Weibull tail. It is clear, therefore, as expected, that the stationary distribution of the Gaussian model has a Weibull tail, and that none of the other curves have Weibull tails. If a pure Weibull distribution was fitted to the model under consideration, for example by fitting fractional Brownian motion to the traffic and using the Weibull distribution which has been

obtained as the large-deviations limit for the buffer distribution of a system handling this traffic, the resulting straight line curve would appear in almost the same position as the curve labelled CLT.

Plots of this type for a wider range of λ values have shown that for Poisson arrival rates greater than about 4000 per second, a Gaussian model should be satisfactory. This figure for when a Gaussian model becomes satisfactory will depend critically upon assumptions concerning expected performance, the rate of each burst, and so on.

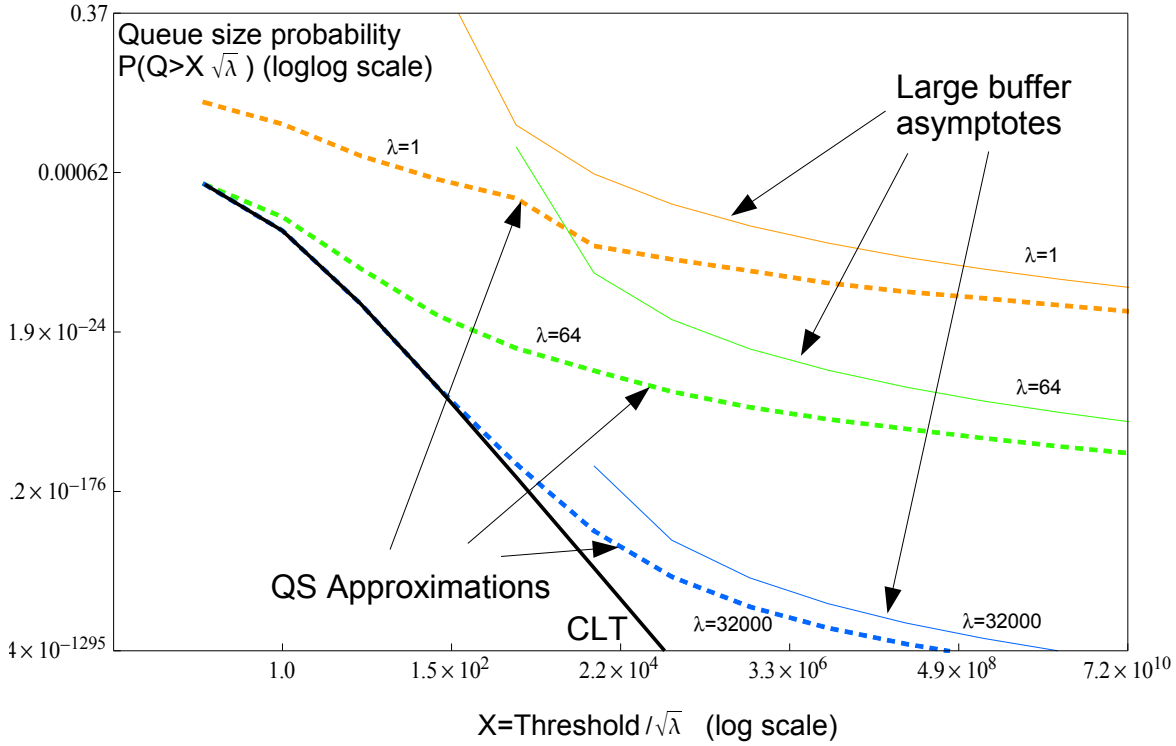

Fig. 8. The stationary CDF as a function of buffer threshold for varying burst arrival rates (λ), estimated by QS algorithm, large buffer asymptote, and the CLT limit; C = mean traffic $+3\sigma_\delta/\delta$; buffer thresholds are measured in units of $\sqrt{\lambda}$.

Figure 8 clearly demonstrates how the heavy traffic *and* large buffer limits can apply despite appearing to lead to contradictory conclusions since, although the CDF's clearly converge to the Gaussian result as λ grows, each one individually exhibits a completely different tail behaviour (which we know to be power-law).

6.3 Comparison with the many sources limit

Large Deviations Theory has been applied in two different ways to the system under study in this paper. One of these approaches was discussed in the previous subsection. The other approach has been used in [10, 15, 17–19, 32–34]. In this approach the limit is taken as $n \longrightarrow \infty$ where n is the number of sources, and the buffer threshold is assumed to grow in linear proportion to n also.

Figure 9 shows some results reported in [10] compared with results derived from the method recommended in the present paper. The case considered is the one in which there are 100 sources – see

Figure 1 of [10]. The many sources result diverges from the results of the present paper very significantly for any non-zero buffer threshold. Since the many sources approximation lies below the quasi-stationary approximation, which is itself a lower bound, it is clear that the discrepancy reveals a problem with the many sources large deviations result. To further emphasize the fact that the many sources approximation is clearly an underestimate, for $x > 0$, we have included the stationary distribution for an FBN queueing system in this figure as well. The FBN input was chosen so that the input process had the same Hurst parameter as the PPBP process and the same variance at the time interval δ, which is equal to 0.2 in this instance. We expect, from experience with simulations for example [2], that the FBN system will exhibit better performance than the PPBP system. However, the many sources approximation exhibits better performance than the FBN system, which, in turn, is better than the quasi-stationary approximation.

It might be hoped that the difficulties of finding a satisfactory solution for the PPBP queue by the large buffer large deviations asymptotic approximation could be addressed by using the many sources asymptote instead. However, the many sources and the large buffer large deviations results must be consistent to a fairly high degree. As a consequence, the fact that the tail behaviour of the stationary queue distribution is not typical of practical buffer thresholds is forced to be a feature of the many sources approximation as well, for any buffer threshold greater than zero.

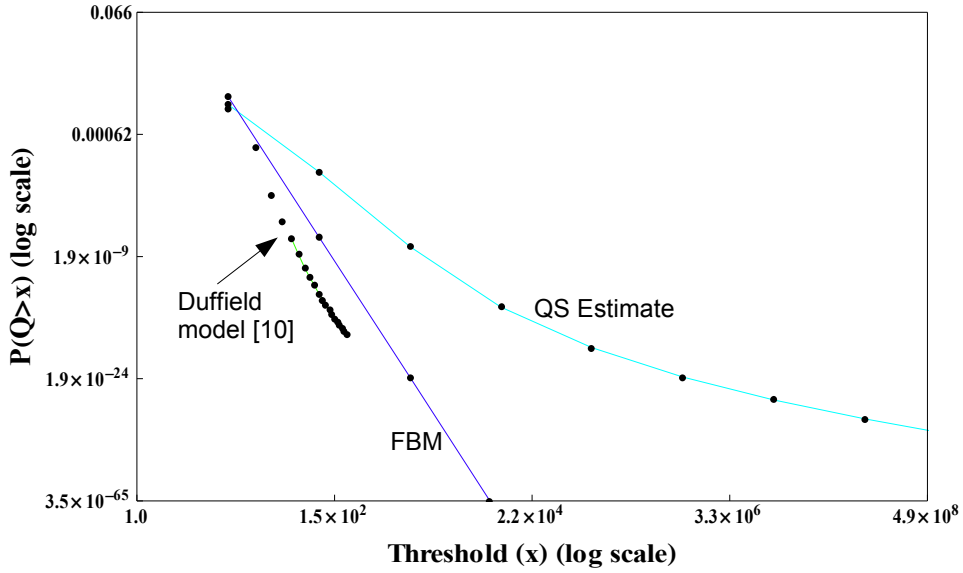

Fig. 9. Comparison with the many sources large deviations results.

7 Concluding remarks

We have shown that the widely known power-law form queueing formulae for PPBP queueing systems, in many cases, will not give us acceptable answers for practical problems. These asymptotic approximations are only accurate for very remote regions of the space of system parameters, except in special cases. In addition, over a substantial region of the space of system parameters the coefficients of these asymptotic approximations are highly sensitive to the system parameters. We have demonstrated that the QS approximation, on the other hand, can be used for the entire range of system parameters and gives results that are consistent with all the existing asymptotic results.

The M/G/∞ traffic model can be further improved, by allowing more complex behaviour for the individual bursts. Such an extension is natural to consider and the methods of this paper appear to be readily applicable to such models.

Appendix A: Overflow Probability $P(Q > x)$ and Loss Probability $P_{loss}(x)$

As observed in [50], for the N$*$D/D/1 queue and its finite buffer N$*$D/D/1/x equivalence, that $P(Q > x)$ provides an upper bound on $P_{loss}(x)$ as follows:

$$P_{loss}(x) \leq \frac{P(Q > x)}{\rho} \tag{32}$$

where ρ is the ratio of the mean arrival rate, denoted $E(A)$, to the service rate C. In [29], it is shown that (32) applies also to the G/D/1 queue and its finite buffer equivalence. This observation was based on the definition of the *overflow period* which is a maximal continuous period of time that the event $Q > x$ occurs. Following [50], it can be shown that (32) applies also to the PPBP queue and its finite buffer equivalence. Since the PPBP queue size Q in the infinite queue system is the same at the beginning as at the end of the overflow period, the amount of work that joined the queue during an overflow period must be equal to the the amount of work served during the overflow period. Suppose $\varepsilon > 0$ is arbitrary. Then, for sufficiently large $L_t > 0$, the mean amount of work lost in the finite buffer PPBP queue (with buffer size x) during L_t is in the interval $(E(A)L_tP_{loss} - \varepsilon, E(A)L_tP_{loss} + \varepsilon)$. This must be lower or equal to the amount of work that *arrived* during the overflow period of the infinite buffer PPBP queue which is equal, as discussed above, to the amount of work *served* during that overflow period. Therefore, since the process Q_t is stationary,

$$E(A)L_tP_{loss} - \varepsilon \leq CL_tP(Q > x).$$

Since $\varepsilon > 0$ was arbitrary, (32) follows.

Appendix B: Proof of Propositions 1–3

As $\lambda \to \infty$, in propositions considered in this appendix, the parameters r, the rate associated with each burst, and C, the capacity of the server, are scaled with λ according to the formulae (17), as explained at the start of Section 5. The remaining parameters, δ and γ of the system of PPBP traffic processes under consideration remain fixed.

Proposition 1 *For any functions, A_λ, B_λ of λ, and any increasing function, $f(\lambda)$, defined on $[0,\infty)$, such that*

$$A_\lambda x^{-f(\lambda)} \leq \phi_\lambda(x) \leq B_\lambda x^{-f(\lambda)}, \tag{18}$$

for all $x > x_0$, necessarily, $\frac{B_\lambda}{A_\lambda} \to \infty$ as $\lambda \to \infty$.

The proof relies on four lemmas which we state and prove prior to presenting the proof.

Lemma 1 *Suppose that a Gaussian process having the same autocovariance as the PPBP and net mean $-\mu$ supplies input to a stationary queue. Then, if Q is the buffer level in this system, for any*

$\varepsilon > 0$ *there exists* $x_0 > 0$ *such that for* $x > x_0$,

$$e^{-(D+\varepsilon)x^{\gamma-1}} \leq P(Q > x) \leq e^{-(D-\varepsilon)x^{\gamma-1}}, \tag{33}$$

where

$$D = \frac{4\lambda r^2 \delta^\gamma (3-\gamma)^{\gamma-2}}{(2-\gamma)3(\gamma-1)^\gamma} \mu^{3-\gamma}.$$

Proof

Since the PPBP variance-time curve, σ_t^2, takes the form shown in (14), as $t \longrightarrow \infty$

$$\sigma_t^2 \sim \mathcal{D} t^{3-\gamma} = \frac{2r^2\lambda t^{3-\gamma}\delta^\gamma}{(\gamma-1)(2-\gamma)(3-\gamma)}$$

where $\mathcal{D} = \frac{2r^2\lambda\delta^\gamma}{(\gamma-1)(2-\gamma)(3-\gamma)}$, and in fact, for typical parameter values, this term is dominant even for relatively small t. In particular, for any $\varepsilon > 0$ there exist $t_0 > 0$ such that for all $t > t_0$,

$$(\mathcal{D} - \varepsilon) t^{3-\gamma} < \sigma_t^2 < (\mathcal{D} + \varepsilon) t^{3-\gamma}. \tag{34}$$

Observe therefore that the limit $g(c) = \lim_{t\to\infty} \frac{\sigma_t^2}{c^2 \sigma_{t/c}^2}$ exists and $g(c) = c^{1-\gamma}$, $c > 0$. Now

$$\inf_{c>0} g(c)(c+\mu)^2/2 = \frac{2(\gamma-1)(3-\gamma)^\gamma}{(3-\gamma)3(\gamma-1)^\gamma} \mu^{3-\gamma},$$

so, by [9, Theorems 2.1 & 2.2],

$$\lim_{b\to\infty} \frac{\sigma_b^2}{b^2} \ln P(Q > b) = -\frac{2(\gamma-1)(3-\gamma)^\gamma}{(3-\gamma)3(\gamma-1)^\gamma} \mu^{3-\gamma}.$$

That is to say, for any $\varepsilon' > 0$ we can find $b_0 > 0$ such that for $b > b_0$,

$$\left(-\frac{2(\gamma-1)(3-\gamma)^\gamma}{(3-\gamma)3(\gamma-1)^\gamma} \mu^{3-\gamma} - \varepsilon' \right) \frac{b^2}{\sigma_b^2} < \ln P(Q > b) < \left(-\frac{2(\gamma-1)(3-\gamma)^\gamma}{(3-\gamma)3(\gamma-1)^\gamma} \mu^{3-\gamma} + \varepsilon' \right) \frac{b^2}{\sigma_b^2}.$$

Using (34), therefore, for any $\varepsilon'' > 0$, there is an $x_0 > 0$ such that for $x > x_0$,

$$\left(-\frac{2(\gamma-1)(3-\gamma)^\gamma}{(3-\gamma)3(\gamma-1)^\gamma} \mu^{3-\gamma} \mathcal{D} - \varepsilon'' \right) x^{\gamma-1} < \ln P(Q > x) < \left(-\frac{2(\gamma-1)(3-\gamma)^\gamma}{(3-\gamma)3(\gamma-1)^\gamma} \mu^{3-\gamma} \mathcal{D} + \varepsilon'' \right) x^{\gamma-1}$$

from which the assertion of the lemma follows by exponentiation. □

Lemma 2 *Given arbitrary* B, $D_1 > D_2 > 0$, $\beta > 0$, *for any* $K > 0$, $x_0 > 0$, *there exists* $x_L > 0$ *such that*

$$\inf_{\lambda,\alpha>0} \sup_{x_0 \leq x \leq x_L} \max \left(\frac{\alpha x^{-\lambda}}{B e^{-D_2 x^\beta}}, \frac{B e^{-D_1 x^\beta}}{\alpha x^{-\lambda}} \right) > K.$$

Proof

Essentially, this lemma says that a graph of the mapping $x \mapsto \alpha x^{-\lambda}$ is fundamentally different in shape to a graph of $x \mapsto Be^{-Dx^{\beta}}$, no matter how the parameters α, λ, B, D, and β are chosen. The detailed proof has been omitted in the interests of brevity, but can be found in the long version of this paper [48].

□

Lemma 3 *The CDF of a Gaussian queueing system with non-zero idle probability is continuous on* $(0,\infty)$.

Proof

This result follows from Theorem 11 in [51, Section 11]. In [51] the main conclusion of Theorem 11 is that the CDF, $F(r)$ say, is absolutely continuous on (r_0,∞) where $r_0 = \inf\{r : F(r) > 0\}$. In [51], $F(r)$ is the CDF of convex functional defined on an Gaussian process on an arbitrary index set whereas in the present case we shall apply this with $F(r)$ as the CDF of

$$\sup_{s \le t} X_t - X_s = X_t - X_{S_t}$$

where S_t denotes the time, s, previous to t where $X_t - X_s$ achieves its maximum value. As noted in the Corollary to Theorem 11 in [51], this sup is a convex linear functional, so Theorem 11 applies to it. In the present case, $r_0 = 0$, since the system has a non-zero probability of being idle. □

Lemma 4 *The sequence* $\phi_\lambda(x)$ *converges to the CDF of the Gaussian queueing system whose input process has the same first and second order statistics as the PPBP uniformly in x on any finite interval in* $[0,\infty)$*, as* $\lambda \longrightarrow \infty$.

Proof

Point-wise convergence follows from the CLT [8, 39] and Lemma 3. Choose a finite interval $[a,b] \subseteq [0,\infty)$. Let $\phi(x) = \lim_{\lambda\to\infty} \phi_\lambda(x)$, $x > 0$. This is a uniformly continuous function on $[a,b]$ (because $[a,b]$ is compact). The rest follows from the fact that both $\phi(x)$ and all the $\phi_\lambda(x)$, are increasing functions of x. □

Proof of Proposition 1

Choose $K > 1$. By Lemma 3, $\phi_\lambda(x)$ converges to the complementary waiting time distribution of a Gaussian queueing system ($\psi(x)$ say) in which the traffic has the same first and second order statistics, and this distribution is continuous.

By Lemma 1, for any $\varepsilon > 0$, for some $x_1 > x_0$, (33) holds for all $x > x_1$. We could choose $\varepsilon = D/4$ for example. Let $D_1 = D + \varepsilon$ and $D_2 = D - \varepsilon$, so, by (33), for all $x > x_1$,

$$Ce^{-D_1 x^{\gamma-1}} \leq \psi(x) \leq Ce^{-D_2 x^{\gamma-1}}. \tag{35}$$

Now, by Lemma 2, we can find $x_L > x_1$ such that for any α, κ, for some $x \in (x_1, x_L)$, either $\frac{\alpha x^{-\kappa}}{Ce^{-D_2 x^{\gamma-1}}} > K^2$ or $\frac{Ce^{-D_1 x^{\gamma-1}}}{\alpha x^{-\kappa}} > K^2$.

Applying this with B_λ for α, and $f(\lambda)$ for κ, we see that over the range x_1 to x_L, the upper bound $B_\lambda x_\lambda^{-f(\lambda)}$ must, for some value of x, fail to approximate the functions $Ce^{-D_1 x^{\gamma-1}}$ and $Ce^{-D_2 x^{\gamma-1}}$ by a ratio of at least K^2, i.e. for any $\lambda > 0$, we can find $x_\lambda \in (x_1, x_L)$ such that either

$$\frac{B_\lambda x_\lambda^{-f(\lambda)}}{Ce^{-D_2 x_\lambda^{\gamma-1}}} > K^2 \tag{36}$$

or

$$\frac{Ce^{-D_1 x_\lambda^{\gamma-1}}}{B_\lambda x_\lambda^{-f(\lambda)}} > K^2. \tag{37}$$

By Lemma 4, $\phi_\lambda(\cdot)$ converges to its limit $\psi(\cdot)$ as $\lambda \to \infty$ uniformly on any finite interval of x values. So we can choose λ_K sufficiently large that for all $\lambda > \lambda_K$ the ratio $\phi_\lambda(x)/\psi(x) > 1/K$ and $\psi(x)/\phi_\lambda(x) > 1/K$ over (x_1, x_L). Using (35), we now see that for all $\lambda > \lambda_K$

$$Ce^{-D_2 x^{\gamma-1}}/\phi_\lambda(x) > 1/K \tag{38}$$

and

$$\phi_\lambda(x)/Ce^{-D_1 x^{\gamma-1}} > 1/K \tag{39}$$

over (x_1, x_L).

If (37) were to hold, multiplying it by (39) gives

$$\frac{B_\lambda x_\lambda^{-f(\lambda)}}{\phi_\lambda(x_\lambda)} < \frac{1}{K},$$

for $\lambda > \lambda_K$, which contradicts the assumption that $B_\lambda x_\lambda^{-f(\lambda)}$ is an upper bound. Hence (37) does not hold.

So (36) holds. Multiply it by (38) to give

$$\frac{B_\lambda x_\lambda^{-f(\lambda)}}{\phi_\lambda(x_\lambda)} > K$$

for $\lambda > \lambda_K$. Since $K > 0$ was arbitrary, this proves that $\frac{B_\lambda x_\lambda^{-f(\lambda)}}{\phi_\lambda(x_\lambda)} \longrightarrow \infty$ as $\lambda \longrightarrow \infty$ and hence, since $A_\lambda x_\lambda^{-f(\lambda)} < \phi_\lambda(x_\lambda)$, also that $\frac{B_\lambda}{A_\lambda} \longrightarrow \infty$ as $\lambda \longrightarrow \infty$. This completes the proof. □

Proposition 2 *In the same context as Proposition 1*

For any mapping, $\lambda \mapsto B_\lambda$, (or $\lambda \mapsto A_\lambda$) and function, $f(\lambda)$, defined on $[0,\infty)$, such that

$$\phi_\lambda(x) \leq B_\lambda x^{-f(\lambda)} \tag{40}$$

$$\left(A_\lambda x^{-f(\lambda)} \leq \phi_\lambda(x)\right), \tag{41}$$

for all $x > x_0$, necessarily, for any $\bar{x}_0 > 0$, $\sup_{x>\bar{x}_0} \frac{B_\lambda x^{-f(\lambda)}}{\phi_\lambda(x)} \to \infty$ ($\inf_{x>\bar{x}_0} \frac{\phi_\lambda(x)}{A_\lambda x^{-f(\lambda)}} \to 0$) as $\lambda \to \infty$.

Proof

Select $\bar{x}_0 > 0$. Except that when we pick x_1, we need to select it to be larger than $\max(x_0, \bar{x}_0)$, instead of just larger than x_0, the proof is the same as for Proposition 1 up to the penultimate conclusion. The penultimate conclusion was that $\frac{B_\lambda x_\lambda^{-f(\lambda)}}{\phi_\lambda(x_\lambda)} \longrightarrow \infty$ as $\lambda \longrightarrow \infty$. Since $x_\lambda > x_1$, it is also larger than $\bar{x}_0$. The primary statement of the proposition follows. The secondary statement follows by an entirely parallel argument. □

Proposition 3 *In the same context as Proposition 1, if $A_\lambda x^{-f(\lambda)}$ is a power-law asymptote for a PPBP queue, in the sense that for any PPBP system, $\frac{A_\lambda x^{-f(\lambda)}}{\phi_\lambda(x)} \longrightarrow 1$ as $x \longrightarrow \infty$, the convergence is not uniform in λ. For any $x_0 > 0$, either $\sup_{x>x_0} \frac{A_\lambda x^{-f(\lambda)}}{\phi_\lambda(x_0)} \longrightarrow \infty$ or $\inf_{x>x_0} \frac{A_\lambda x^{-f(\lambda)}}{\phi_\lambda(x_0)} \longrightarrow 0$ as $\lambda \longrightarrow \infty$.*

Proof

If it is not the case that $\inf_{x>x_0} \frac{A_\lambda x_0^{-f(\lambda)}}{\phi_\lambda(x_0)} \longrightarrow 0$ as $\lambda \longrightarrow \infty$, then for some $K > 0$, for all $x > x_0$, $\frac{A_\lambda x^{-f(\lambda)}}{\phi_\lambda(x)} > K$. Then $(K^{-1} A_\lambda x^{-f(\lambda)})$ will be an upper bound satisfying the conditions of Proposition 2, and therefore $\sup_{x>\bar{x}_0} \frac{K^{-1} A_\lambda x^{-f(\lambda)}}{\phi_\lambda(x)} \to \infty$ as $\lambda \to \infty$, from which the conclusion follows. □

Acknowledgement

This work was supported by the Australian Research Council (ARC). The work on the paper was conducted when M. Zukerman was with the ARC Special Research Centre for Ultra-Broadband Information Networks, EEE Dept, The University of Melbourne, Australia.

References

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

[2] R. G. Addie, T. D. Neame, M. Zukerman, Performance evaluation of a queue fed by a Poisson Pareto burst process, Computer Networks 40 (2002) 377–397.

[3] R. G. Addie, M. Zukerman, T. D. Neame, Broadband traffic modeling: simple solutions to hard problems, IEEE Commun. Mag. 36 (8) (1998) 88–95.

[4] R. G. Addie, M. Zukerman, T. D. Neame, Fractal traffic: Measurements, modelling and performance evaluation, in: Proc. IEEE INFOCOM '95, Vol. 3, 1995, pp. 977–984.

[5] S. Borst, B. Zwart, A reduced-peak equivalence for queues with a mixture of light-tailed and heavy-tailed input flows, Adv. in Appl. Probab. 35 (2003) 793–805.

[6] S. Borst, B. Zwart, Fluid queues with M/G/∞ input, Math. Oper. Res. 30 (4) (2005) 852–879.

[7] F. Brichet, J. Roberts, A. Simonian, D. Veitch, Heavy traffic analysis of a storage model with long range dependent on/off sources, Queueing Syst. 23 (1-4).

[8] K. Debicki, Z. Palmowski, On-off fluid models in heavy traffic environment, Queueing Syst. 33 (4) (1999) 327–338.

[9] N. G. Duffield, N. O'Connell, Large deviations and overflow probabilities for the general single-server queue, with applications, Math. Proc. Cambridge Philos. Soc. 118 (1995) 363–374.

[10] N. G. Duffield, Queueing at large resources driven by long-tailed M/G/∞-modulated processes, Queueing Syst. 28 (1–3) (1998) 245–266.

[11] M. W. Garrett, W. Willinger, Analysis, modeling and generation of self-similar VBR video traffic, ACM SIGCOMM Computer Communication Review 24 (4) (1994) 269–280.

[12] M. Grossglauser, J.-C. Bolot, On the relevance of long-range dependence in network traffic, IEEE/ACM Trans. Networking 7 (5) (1999) 629–640.

[13] D. P. Heyman, T. V. Lakshman, What are the implications of long-range dependence for VBR-video traffic engineering?, IEEE/ACM Trans. Networking 4 (3) (1996) 301–317.

[14] M. M. Krunz, A. M. Makowski, Modeling video traffic using M/G/∞ input processes: A compromise between Markovian and LRD models, IEEE J. Sel. Areas Commun. 16 (5) (1998) 733–748.

[15] N. Likhanov, R. R. Mazumdar, Loss asymptotics in large buffers fed by heterogeneous long-tailed sources, Adv. in Appl. Probab. 32 (4) (2000) 1168–1189.

[16] N. Likhanov, B. Tsybakov, N. D. Georganas, Analysis of an ATM buffer with self-similar ("fractal") input traffic, in: Proc. IEEE INFOCOM '95, Vol. 3, 1995, pp. 985–992.

[17] M. Mandjes, J. H. Kim, Large deviations for small buffers: An insensitivity result, Queueing Syst. 37 (4) (2001) 349–362.

[18] M. Mandjes, A note on queues with M/G/∞ input, Oper. Res. Lett. 28 (5) (2001) 233–242.

[19] M. Mandjes, S. Borst, Overflow behavior in queues with many long-tailed inputs, Adv. in Appl. Probab. 32 (4) (2000) 1150–1167.

[20] I. Norros, A storage model with self-similar input, Queueing Syst. 16 (1994) 387–396.

[21] M. Parulekar, A. M. Makowski, Tail probabilities for a multiplexer with self-similar traffic, in: Proc. IEEE INFOCOM '96, Vol. 3, 1996, pp. 1452–1459.

[22] M. Parulekar, A. M. Makowski, Tail probabilities for M/G/∞ input processes (i): preliminary asymptotics, Queueing Syst. 27 (3-4) (1997) 271–296.

[23] S. Resnick, G. Samorodnitsky, Steady-state distribution of the buffer content for M/G/∞ input fluid queues, Bernoulli 7 (2) (2001) 191–210.

[24] B. K. Ryu, A. Elwalid, The importance of long-range dependence of VBR video traffic in ATM traffic engineering: myths and realities, ACM SIGCOMM Computer Communication Review 26 (4) (1996) 3–14.

[25] B. Tsybakov, N. D. Georganas, Self-similar traffic and upper bounds to buffer-overflow probability in an ATM queue, Performance Evaluation 32 (1) (1998) 57–80.

[26] B. Tsybakov, N. D. Georganas, Overflow and losses in a network queue with a self-similar input, Queueing Syst. 35 (1-4) (2000) 201–235.

[27] K. P. Tsoukatos, A. M. Makowski, Power-law vs exponential queueing in a network traffic model, Performance Evaluation 65 (1) (2008) 349–362.

[28] B. Zwart, S. Borst, M. Mandjes, Exact asymptotics for fluid queues fed by multiple heavy-tailed on-off flows, Ann. Appl. Probab. 14 (2004) 903–957.
[29] J. Roberts, U. Mocci, J. Virtamo, *Broadband Network Teletraffic*, Final Report of Action COST 242, Lecture Notes in Computer Science, Springer, 1996.
[30] R. G. Addie, P. Mannersalo, I. Norros, Most probable paths and performance formulae for buffers with Gaussian input traffic, European Transactions on Telecommunications 13 (3) (2002) 183–196.
[31] D. T. Z. Liu, P. Nain, Z.-L. Zhang, Asymptotic behavior of a multiplexer fed by a long-range dependent process, J. Appl. Probab. 36 (1) (1999) 105–118.
[32] C. Courcoubetis, R. R. Weber, Buffer overflow asymptotics for a buffer handling many traffic sources, J. Appl. Probab. 33 (1996) 886–903.
[33] D. D. Botvich, N. G. Duffield, Large deviations, economies of scale, and the shape of the loss curve in large multiplexers, Queueing Syst. 20 (1995) 293–320.
[34] N. Likhanov, R. R. Mazumdar, Cell loss asymptotics for buffers fed with a large number of independent stationary sources, J. Appl. Probab. 36 (1) (1999) 86–96.
[35] G. Appenzeller, I. Keslassy, N. McKeown, Sizing router buffers, in: Proc. ACM SIGCOMM, 2004, pp. 281–292.
[36] L. L. H. Andrew, T. Cui, J. Sun, M. Zukerman, K.-T. Ko, S. Chan, Buffer sizing for nonhomogeneous TCP sources, IEEE Commun. Lett. 9 (6) (2005) 567–569.
[37] D. Wischik, N. McKeown, Part I: Buffer sizes for core routers, ACM/SIGCOMM Computer Communication Review 35 (3) (2005) 75–78.
[38] J. Cao, K. Ramanan, A Poisson limit for buffer overflow probabilities, in: Proc. IEEE INFOCOM 2002, Vol. 2, 2002, pp. 994–1003.
[39] R. G. Addie, On weak convergence of long-range-dependent traffic processes, Journal of Statistical Planning and Inference 80 (1-2) (1999) 155–171.
[40] T. Mikosch, S. Resnick, H. Rootzen, A. Stegeman, Is network traffic approximated by stable levy motion or fractional brownian motion?, Ann. Appl. Probab. 12 (1) (2002) 23–68.
[41] R. G. Addie, T. D. Neame, M. Zukerman, On asymptotic accuracy in queueing theory, in: Proc. Australian Telecommunication Networks and Applications Conference (ATNAC), 2003.
[42] G. L. Choudhury, D. M. Lucantoni, W. Whitt, On the effectiveness of effective bandwidth for admission control in ATM networks, in: The Fundamental Role of Teletraffic in the Evolution of Telecommunications Networks, Proc. 14th International Teletraffic Congress, Vol. 1a, Elsevier, 1994, pp. 411–420.
[43] G. L. Choudhury, D. M. Lucantoni, W. Whitt, Squeezing the most out of ATM, IEEE Trans. Commun. 44 (2) (1996) 203–217.
[44] J. Choe, N. B. Shroff, On the supremum distribution of integrated stationary Gaussian processes with negative linear drift, Adv. in Appl. Probab. 31 (1999) 134–156.
[45] D. J. Wischik, Sample path large deviation for queues with many inputs, Ann. Appl. Probab. 11 (2) (2001) 379–404.
[46] A. Shwartz, A. Weiss, Large Deviations for Performance Analysis, Chapman & Hall, 1995.
[47] M. Abramowitz, I. A. Stegun, Handbook of Mathematical Functions, Dover, New York, 1970.
[48] R. G. Addie, M. Zukerman, T. D. Neame, Performance analysis of a poisson-pareto traffic model over the entire range of buffer sizes – extended version, Tech. Rep. SC-MC-0822, University of Southern Queensland, http://www.sci.usq.edu.au/research/workingpapers.php (November 2008).
[49] R. G. Addie, Quasi-stationary approximation for PPBP queues – Mathematica source code, Tech. Rep. SC-MC-0819, University of Southern Queensland, http://www.sci.usq.edu.au/research/workingpapers.php (November 2008).

[50] A. K. Wong, Queueing analysis for ATM switching of continuous-bit-rate traffic – a recursion computation method, in: Proc. IEEE GLOBECOM '90, Vol. 3, 1990, pp. 1438–1444.
[51] M. A. Lifshits, Gaussian Random Functions, Kluwer Academic Publishers, 1995.