On asymptotic accuracy in queueing theory - the tale of the misleading tail

Ronald G. Addie* Timothy D. Neame+ Moshe Zukerman+

* Department of Mathematics and Computing, University of Southern Queensland, Australia

+ ARC Special Research Centre for Ultra-Broadband Information Networks, EEE Dept, The University of Melbourne, Australia.

Abstract— Recently results have shown that a single server queue Poisson Pareto Burst Process input has a tail which is bounded by hyperbolic functions. We show that the hyperbolic upper and lower bounds for this system can be very misleading, that this hyperbolic tail result is relevant only from a certain threshold onwards, and the magnitude of this threshold may be very large. We also show that any hyperbolic upper and lower bounds for a tail of the stationary waiting time complementary distribution necessarily become further apart as the rate of the process increases.

I. Introduction

When we tackle difficult problems we often need to compromise in the accuracy of our answers or the rigor with which they are derived. It is important that neither of these is compromised excessively. A purportedly very accurate answer which might be wrong because it is not rigorously supported may not be useful, and neither is a very inaccurate answer proved beyond doubt.

In practice there is always a degree of uncertainty and a degree of inaccuracy. In this paper, we investigate a problem where an asymptotically accurate answer obtained rigorously appears to be insufficiently accurate for practical use.

So, two approaches have been used. One of these approaches is rigorous, but merely demonstrates that in a certain asymptotic region of the parameter space, the hyperbolic approximations under consideration necessarily become worse and worse. The second approach requires more indulgence of the reader. We ask the reader to accept a new approximation [3] for the problem under study in order to quantify the issue of accuracy and the region of parameter space where accuracy of the hyperbolic bounds becomes unacceptable.

Large Deviations Theory is a particularly powerful and elegant approach to deriving asymptotically accurate approximations which has evolved rapidly over recent years and been applied to many problems of analysis of communication systems. The large deviations approach provides an asymptotic approximation. Furthermore, the type of accuracy sought is usually somewhat more relaxed than the obvious measure on account of being expressed in terms of a logarithm of the function under study.

The concept of seeking asymptotic accuracy rather than accuracy in its own right has led, in some cases, to the use of the term *exact* for approximations which vary from the quantity of interest by a ratio which tends, in the limit, to one. However, there are significant risks in accepting such approximations. We need to have some assurance that the operating region of real interest is not so far from the limiting region where the approximation is valid. Otherwise, the approximation will not be of any value.

In many cases, the large-deviations style of approximation seems to be well suited to the task in hand. By relaxing the accuracy required in the sought after system characterization it becomes possible to see more easily, more clearly, the truly important features of the system.

Recently, an example of considerable interest and importance in which the asymptotic shape of the tail appears to be *necessarily* a poor approximation of the function as a whole has arisen naturally. Use of the asymptotic shape of the tail can produce results which can be misleading from a practical point of view because the parameter region where accuracy becomes satisfactory is very remote from realistic values.

In the next section, the Poisson Pareto Burst Process (PPBP) (also referred to as the M/G/∞ traffic process by many authors) [9], [11], [12], [15], [17], [18] is introduced and its analysis by two different approaches is presented: the first approach uses asymptotic shape analysis whereas the second, introduced in [3],where it was termed *quasi-stationary approximation*, relies on a separation of time scales.

In the third section two arguments are used to show that the hyperbolic upper and lower bounds for the stationary waiting time in a PPBP queue are misleading. First it is shown that any hyperbolic upper and lower bounds for a tail of the stationary waiting time complementary distribution function (CDF) necessarily diverge from each other as the rate of the PPBP input increases. Then it is shown that for any value of λ there is a buffer level, x_g say, only beyond which this hyperbolic tail can be expected to show itself. An estimate of x_g as a function of the other parameters is presented.

It should be noted that the problem of asymptotic slope analysis methods providing misleading guidance has been raised previously in [6], [7]. The present instance is somewhat more complicated because here we consider models which exhibit long range dependence.

II. A Queue subjected to PPBP traffic

The traffic in a PPBP is made up of bursts, the burst starts forming a Poisson process, the burst lengths following a

Pareto distribution. Except in [11], [17] the rate at which packets are generated *during* each burst is constant.

We form a *discrete time* PPBP by dividing time into fixed length intervals, and for each such time interval, we consider the total amount of work contributed during this period of time by all the bursts. This discrete time PPBP is the process studied in this paper.

Denoting a burst duration by d, the CDF of the Pareto distribution takes the form:

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

$\delta > 0$. For $1 < \gamma < 2$, we have that $\mathrm{E}(d) = \frac{\delta\gamma}{(\gamma-1)}$ and the variance of d is infinite.

A. Hyperbolic Bounds

Upper and lower hyperbolic bounds on the stationary CDF for an SSQ with M/G/∞ input process were obtained in [18]. Large Deviations Theory was used to obtain a consistent result in [15]. Related results have been obtained in [11], [13]. In [18] estimates of both finite buffer time congestion probabilities and finite queue loss probabilities are presented. Only the expressions for time congestion probabilities are presented here.

The upper bound for the CDF of the workload, Q, of the SSQ is given by

$$\Pr(Q > x) \leq \frac{\left(\lambda\gamma\delta^{\gamma}(\gamma-1)^{-\gamma}\left(\frac{C}{r}+2\right)^{\gamma-1} r^{\gamma-1}\right)^{k} x^{(-\gamma+1)k}}{k!} \tag{2}$$

and the lower bound by

$$\Pr(Q > x) \geq \frac{\gamma^{k}\delta^{\gamma k} r^{(\gamma-1)k} x^{(-\gamma+1)k}}{\gamma(\gamma-1)^{k}\left(\mathrm{E(d)} + (1-\mathrm{e}^{-\rho/\mathrm{E(d)}})^{-1} - 1\right)^{\gamma+k}}, \tag{3}$$

where $\mathrm{E(d)}$ is the mean burst duration. For Pareto distributed burst lengths, as in the PPBP, $\mathrm{E(d)} = \frac{\delta\gamma}{(\gamma-1)}$. and the parameter k is given by $k = 1 + \left\lfloor \frac{C}{r} - \lambda\mathrm{E(d)} \right\rfloor$. Finally, the value of ρ depends upon $\lambda\mathrm{E(d)}$. If $\lambda\mathrm{E(d)} \leq 1$ then $\rho = \lambda\mathrm{E(d)}$, otherwise, ρ may be any value in the range

$$0 \leq \rho < \begin{cases} 1 + \delta_p - \Delta, & \text{if } \Delta \geq \delta_p, \\ \delta_p - \Delta, & \text{if } \Delta < \delta_p. \end{cases} \tag{4}$$

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

The upper and lower bounds at (2) and (3) decay at the same rate. We would therefore naturaly predict that for large buffer sizes our PPBP should show queueing performance in which the probability of loss decays as $x^{(-\gamma+1)k}$.

The fact that the upper and lower bounds decay at the same hyperbolic rate which gives us some confidence that the true "asymptotic shape" of the CDF has been identified. A result which includes both shape *and weight* for the asymptotic form of a stationary PPBP queue buffer CDF, is given in [11]. That is to say, in that paper a function is given explicitly whose ratio to the CDF tends to one as the buffer level tends to infinity. This asymptotic form is neither an upper nor a lower bound. However, even when the asymptotic form of a CDF is known in this sense, there are considerable risks that the result may provide a very poor approximation to the CDF if the region where the approximation becomes accurate is very remote. Results in the next section demonstrate that this problem appears with full force in the PPBP queue.

B. The Decoupling Approach

A different approximation for the queueing performance of the PPBP SSQ was introduced in [3]. This approach is based on decoupling 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 of the initial bursts which last for the entire time period, we label as *long bursts*. All other bursts are called *short bursts*. The short bursts include: (1) those bursts that start at or before t and end before $t+W$, (2) those bursts that start after t and finish at or after $t+W$ and (3) those bursts that start after t and finish before $t+W$. Considering these long and short bursts, we divide the PPBP into two independent processes: (1) the *long bursts process* and (2) the *short bursts process*. The long bursts process is a stationary but non-ergodic process containing only the long bursts. The short bursts process contains all the remaining bursts, and is stationary on the interval $[0, W]$ [3].

Denote the service rate by C. For a given W, supposing that there are n simultaneous long (length W) bursts, we can use known techniques for SRD processes (e.g. the technique given in [1]) to calculate the performance of the short bursts process in a queue with service rate $C - nr$. We then calculate an estimate of the performance of the PPBP in a queue with service rate C by summing these estimates, weighted by the probability that the long bursts process will contain n bursts.

There are many possible ways to separate long and short bursts. We have chosen the above to guarantee stationarity of the two processes. Clearly, we allow some "short" bursts to be longer than some "long" bursts, but this does not affect the consistency or usefulness of the model. Notice that SRD process we use to model the short bursts does allow for very long bursts, although none of them attain this length inside the interval of length W.

III. The Misleading Tail in the PPBP Queue

The large deviations method and related asymptotic approximations, as applied in [13], [15], [17], [18], focus on one particular method by means of which large buffer levels can occur – the simultaneous occurrence of sufficiently many long bursts for the server to be overloaded while all these long bursts are active. However, there is another way in which a large buffer level may develop [3]. It is possible that many simultaneous *long bursts* could lead to levels of server utilization *moderately close* to capacity, although not sufficient to cause overload all by themselves. At the same time that these long bursts take the server moderately close to capacity, a fluctuating load of short bursts could add sufficient load to lead to steadily increasing buffer levels.

The model of queue behaviour in which long bursts and short bursts combine to produce the overflow includes as a special case the situation where the long bursts produce an overflow all by themselves, which is the dominant cause for reaching very high buffer levels. As a consequence, the tail behaviour of the CDF of stationary buffer levels of [3] is consistent with that of [17], i.e. it is hyperbolic.

The next subsection reveals a problem with models where only long bursts cause congestion and the joint effect of long and short bursts is neglected, i.e. purely hyperbolic approximations.

A. Unbounded separation of bounds for large λ

From [2], we know that the stationary waiting time CDF for a single server queue fed by a Poisson-Pareto Burst process converges weakly, as the intensity of the Poisson process, λ, increases, to the waiting time CDF of the corresponding Gaussian queueing system. The fact that the Gaussian system has a Weibull tail appears to contradict the hyperbolic tails of individual functions making up the limit, however this is not necessarily a contradiction because as $\lambda \longrightarrow \infty$, the remoteness of the hyperbolic tails may increase, as we shall see.

Proposition III.1 confirms this explanation. In preparation we need several lemmas.

Lemma 1: Given arbitrary B, $D_1 > D_2 > 0$, $\beta > 0$, for any $K > 0$, $x_0 > 0$, there exists $x_L > 0$ such that

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

Proof

Set $f(x) = \alpha x^{-\lambda}$ and $g_i(x) = Be^{-D_i x^\beta}$, $i = 1, 2$. We want to show that for any $K > 0$, we can find x_L such that the best possible result obtainable by varying α and λ will still produce, for some $x \in [x_0, x_L]$, either $\frac{f(x)}{g_2(x)} > K$ or $\frac{g_1(x)}{f(x)} > K$.

Set $f'(u) = \log f(u^{1/\beta}) = \log\alpha + \left(-\frac{\lambda}{\beta}\right)\log u$ and $g_i'(u) = \log g_i(u^{1/\beta}) = \log B - D_i u$. Thus, $f'(u) - g_i'(u) = \left(-\frac{\lambda}{\beta}\right)\log u + D_i u - (\log B - \log\alpha)$, and

$$|f'(u) - g_i'(u)| \le \theta \Longleftrightarrow e^{-\theta} < \frac{f(u^{1/\beta})}{g_i(u^{1/\beta})} \le e^{\theta}.$$

The maximum of $\log u - a'u + b'$, for $u > 0$, occurs when $u = 1/a'$ and the value of this maximum is $-1 - \log a' + b'$. The minimum value, in the interval (u_0, u_L), is $\min(\log u_0 - a'u_0 + b', \log u_L - a'u_L + b')$ and occurs when $u = u_0$ or u_L. The value of b' which minimises the maximum of $|\log u - a'u + b'|$ on $[u_0, u_L]$ is $-\frac{1}{2}(1 + \log a' + \max(\log u_0 - a'u_0, \log u_L - a'u_L))$, which is also the value of $|\log u - a'u + b'|$ at this minimax choice of b', and u. The minimum of this value over positive a' occurs when $\log u_0 - a'u_0 = \log u_L - a'u_L$ hence $a' = \frac{\log u_0 - \log u_L}{u_0 - u_L}$ and so the minimum over all positive a' is

$$-\frac{1}{2}\left(1 + \log\left(\frac{\log u_0 - \log u_L}{u_0 - u_L}\right) + \frac{u_L \log u_0 - u_0 \log u_L}{u_L - u_0}\right).$$

Applying this result to $f_1(u) - g_1(u)$, which is a multiple of $\log u - a'u + b'$, for suitable a', b', we find that $\min_{u\in[u_0,u_l]} |f_1(u) - g_1(u)|$ is unbounded as $u_L = x_L^{\frac{1}{\beta}}$ increases. □

Lemma 2: If Q denotes the buffer level in a stationary Gaussian queue with variance function σ_b^2 and net mean input $-\mu$, and the limit $g(c) = \lim_{t\to\infty} \frac{\sigma_t^2}{c^2\sigma_{t/c}^2}$ exists, for $c > 0$.

$$\lim_{b\to\infty} \frac{\sigma_b^2}{b^2} \log P(Q > b) = -\inf_{c>0} g(c)(c+\mu)^2/2.$$

Proof

See [8]. □

Lemma 3: Suppose now that a Gaussian process having the same autocovariance as the PPBP process and net mean $-\mu$ and this process supplies input to a stationary queue. Then, if Q is the buffer contents in this system, for any $\varepsilon > 0$ there exists $x_0 > 0$ such that for $x > x_0$,

$$Ce^{-(D+\varepsilon)x^\beta} \le P(Q > x) \le Ce^{-(D-\varepsilon)x^\beta}, \tag{5}$$

where C is 1 (or any other number), $\beta = \gamma - 1$, and $D = \frac{4\lambda r^2 \delta^\gamma (3-\gamma)^{\gamma-1}}{(2-\gamma)(3-\gamma)^3(\gamma-1)^\gamma}\mu^{3-\gamma}$.

Proof

As presented in [3],

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

The term in $t^{3-\gamma}$ is dominant even for relatively small t. In particular, setting $H = \frac{2\lambda r^2\delta^\gamma}{(\gamma-1)(2-\gamma)(3-\gamma)}$, for any $\delta > 0$ there exist $t_0 > 0$ such that for all $t > t_0$,

$$(H - \delta)t^{3-\gamma} < \sigma_t^2 < (H + \delta)t^{3-\gamma}. \tag{7}$$

Clearly, in this instance, the limit defining $g(c)$ exists and $g(c) = c^{1-\gamma}$, $c > 0$. It follows that $\inf_{c>0} g(c)(c+\mu)^2/2 = \frac{2(\gamma-1)(3-\gamma)^\gamma}{(3-\gamma)^3(\gamma-1)^\gamma}\mu^{3-\gamma}$. So, applying the previous lemma,

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

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

$$\left(-\frac{2(\gamma-1)(3-\gamma)^\gamma}{(3-\gamma)^3(\gamma-1)^\gamma}\mu^{3-\gamma} - \delta'\right)\frac{b^2}{\sigma_b^2} < \log P(Q > b) < \left(-\frac{2(\gamma-1)(3-\gamma)^\gamma}{(3-\gamma)^3(\gamma-1)^\gamma}\mu^{3-\gamma} + \delta'\right)\frac{b^2}{\sigma_b^2} \tag{8}$$

Using (7), we find, therefore, for any $\varepsilon > 0$, there is an $x_0 > 0$ such that for $x > x_0$,

$$\left(-\frac{2(\gamma-1)(3-\gamma)^\gamma}{(3-\gamma)^3(\gamma-1)^\gamma}\mu^{3-\gamma}H - \varepsilon\right)b^{\gamma-1} < \log P(Q > b) < \left(-\frac{2(\gamma-1)(3-\gamma)^\gamma}{(3-\gamma)^3(\gamma-1)^\gamma}\mu^{3-\gamma}H + \varepsilon\right)b^{\gamma-1} \tag{9}$$

which is as we set out to show. □

Lemma 4: The CDF of a Gaussian queueing system is continuous on $(0,\infty)$.

Proof

Let us denote the workload arriving up to time t (starting at time zero) by X_t. The queued workload in this Gaussian queue at time t can be expressed as

$$Q_t = \sup_{s \le t} X_t - X_s = X_t - X_{\mathcal{S}_t}$$

where $\mathcal{S}_t$ denotes the time, s, previous to t where $X_t - X_s$ achieves its maximum value. The process $\mathcal{S}_t$ is an increasing process which alternates between periods when it remains fixed, while Q_t fluctuates with the same relative movements as X_t, and periods when it takes the value t, during which Q_t takes the value 0. The periods when $\mathcal{S}_t$ remains fixed are known as *busy periods* and the intervals between the busy periods are known as *idle periods*. We can confine out attention to the busy periods, since $Q_t \equiv 0$ in the idle periods.

We desire to prove that $P(Q_t = x) = 0$ for any x other than 0. During an individual busy period, $X_{\mathcal{S}_t}$ remains fixed.

For a moment let us confine our attention to the behaviour of Q_t in one busy period. Suppose X_t ranges from a to b during this busy period, i.e. the smallest value taken by X_t is a and the largest is b. It follows that Q_t will range from 0 to $b-a$. For any $y \in ([a,b]$, the Lesbesgue measure of the set $\{t : X_t = y\}$, which we shall denote by $m(\{t : X_t = y\})$, is zero, with probability one, so it is also the case that if $x > 0$, $m(\{t : Q_t = x\})$ must be zero with probability one.

It follows, since in any finite interval there can only be a countable number of busy periods, that for any $T > 0$, $x > 0$, $m(\{t : Q_t = x, t \in [-T,T]\}) = 0$ with probability one.

By the ergodicity of the process Q_t,

$$P(Q_t \in (x,y]) = \lim_{T \longrightarrow \infty} \frac{1}{2T} m(\{t : Q_t \in (x,y], t \in [-T,T]\})$$

with probability one, where $m(\cdot)$ denotes Lesbesgue measure. In particular, if we set $x = y$, since

$$m(\{t : Q_t = x, t \in [-T,T]\}) \equiv 0$$

for all $T > 0$, $P(Q_t = x) = 0$, as we set out to show.

□

Note that we can now deduce, by means of Theorem 11 from [10], the stronger result that the stationary CDF of the Gaussian queue contents is *absolutely continuous*. The result in [10] is more general and has a conclusion which is stronger, in this respect, but does not exclude the possibility that the CDF has one discontinuity at a location r_0 other than 0, which is specifically excluded by the result just proved. Absolute continuity of the CDF will not be used in the sequel.

Lemma 5: The sequence $\phi_\lambda(x)$ converges to the CDF of the Gaussian queueing system with the same first and second order statistics uniformly in x on any finite interval in $[0,\infty)$.

Proof

Point-wise convergence follows from the CLT [2] and Lemma 4. 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). Choose $\varepsilon > 0$. We seek Λ such that for all $\lambda > \Lambda$, for all $x \in [a,b]$, $|\phi_\lambda(x) - \phi(x)| < \varepsilon$.

By uniform continuity of $\phi(x)$ we can find $\delta > 0$ such that whenever $|x-y| < \delta$, $|\phi(x) - \phi(y)| < \varepsilon/2$. Divide $[a,b]$ into $M = \lceil (b-a)/\delta \rceil$ intervals of length at most δ, with endpoints $a_0, \ldots, a_M$ and midpoints $x_0, \ldots, x_{M-1}$. Notice that every point in $[a,b]$ lies between two adjacent points in the set $\{x_0, \ldots, x_{M-1}\}$ and is no more than δ distant from each of these points.

Choose Λ sufficiently large that $|\phi_\lambda(x_i) - \phi(x_i)| < \varepsilon/2$ for all $i < M$ and for all $\lambda > \Lambda$.

Thus, using the fact that all the $\phi_\lambda(x)$ are non-increasing functions, when $\lambda > \Lambda$, supposing that $x_i \le x < x_{i+1}$,

$$\begin{aligned} \phi_\lambda(x) &< \phi_\lambda(x_i) < \phi(x_i) + \varepsilon/2 \\ &< \phi(x) + \varepsilon. \end{aligned} \tag{10}$$

A similar argument shows that $\phi_\lambda(x) > \phi(x) - \varepsilon$. □

The next proposition applies to the situation where the traffic on a link grows steadily and the size of the link carrying this traffic also grows steadily in a manner which asymptotically provides a consistent quality of service. For convenience, we assume that the traffic is rescaled so that the system converges weakly to a specific Gaussian limit, as discussed in [2], [4].

Proposition III.1: Suppose, for $\lambda = \lambda_1, \lambda_2, \ldots \to \infty$, $\phi_\lambda(x)$ is the CDF of a stationary PPBD queueing system S_λ, such that the service capacity, C_λ, and the rate, r_λ, of each system is chosen (in particular, by changing the scale used for work) so that the first and second order statistics of the *net* input process (the input process minus the service process) are the same for all systems. Thus, $r_\lambda = r_1 \sqrt{\frac{\lambda_1}{\lambda}}$ for all λ, and the service rate of system S_λ is $C_\lambda = m + r_1 E(d) \sqrt{\lambda \lambda_1}$ for all λ for a certain $m > 0$.

Then, for any numbers, A_λ, B_λ and increasing function, $f(\lambda)$, defined on $[0,\infty)$, such that

$$A_\lambda x^{-f(\lambda)} \le \phi_\lambda(x) \le B_\lambda x^{-f(\lambda)}, \tag{11}$$

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

Proof

Suppose, to the contrary, that $\frac{B_\lambda}{A_\lambda} \le K$ for all $\lambda > 0$. By the Central Limit theorem [2], $\phi_\lambda(x)$ converges to the complementary waiting time distribution of a Gaussian queueing system ($\psi(x)$ say) with the same first and second order statistics, which is continuous, by Lemma 4.

By Lemma 3, for any $\varepsilon > 0$, for some $x_1 > x_0$, (5) 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 (5), for all $x > x_1$,

$$C e^{-D_1 x^\beta} \le \psi(x) \le C e^{-D_2 x^\beta}. \tag{12}$$

Now, by Lemma 1, we can find $x_L > x_1$, such that over the range of x values, from x_1 to x_L, the best we can do in approximating $Ce^{-D_1 x^\beta}$, $Ce^{-D_2 x^\beta}$ by $\alpha x^{-\kappa}$, by varying κ and α, produces a ratio of at least K^2 for some $x \in (x_1, x_L)$, in the sense that either $\frac{\alpha x^{-\kappa}}{Ce^{-D_2 x^\beta}} > K^2$ or $\frac{Ce^{-D_1 x^\beta}}{\alpha x^{-\kappa}} > K^2$. 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 $Ce^{-D_1 x^\beta}$ and $Ce^{-D_2 x^\beta}$ 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

$$\frac{B_\lambda x_\lambda^{-f(\lambda)}}{Ce^{-D_2 x_\lambda^\beta}} > K^2 \tag{13}$$

or

$$\frac{Ce^{-D_1 x_\lambda^\beta}}{B_\lambda x_\lambda^{-f(\lambda)}} > K^2. \tag{14}$$

By Lemma 5, $\phi_\lambda(\cdot)$ converges to its limit $\psi(\cdot)$ as $\lambda \to \infty$ uniformly on any finite interval of x values. So we can choose λ_L sufficiently large that for all $\lambda > \lambda_L$ the ratio $\phi_\lambda(x)/\psi(x)$ is more than $1/\sqrt{K}$ and less than $\sqrt{K}$ over (x_1, x_L). It follows that for all $\lambda > \lambda_L$

$$\phi_\lambda(x)/Ce^{-D_1 x^\beta} < K \tag{15}$$

and

$$Ce^{-D_2 x^\beta}/\phi_\lambda(x) < K \tag{16}$$

over (x_1, x_L).

So, if (13) holds, combine it with (15) to give

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

or if (14) holds, combine it with (16) to give

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

either of which contradict our assumption that $\frac{B_\lambda}{A_\lambda} < K$.

This completes the proof. □

The ratio between the upper and lower hyperbolic bounds formulated in [18] is plotted as a function of λ in Figure 1. The parameters in this example are $\gamma = 1.4$, $\delta = 1$, $r = 6.32$. The ratio increases extremely quickly as a function of λ, confirming Proposition III.1.

B. The level where the tail becomes hyperbolic

In the last subsection, an inherent problem with upper and lower bounds on a hyperbolic tail approximation for stationary buffer distributions for a PPBP queueing system was identified. What about asymptotic forms for the tail which do not attempt to provide upper or lower bounds, e.g. as provided in [11]?

From [11], for any PPBP queueing system, S_λ, in which λ denotes the intensity of the Poisson arrival process of bursts, the stationary CDF of the buffer contents, $\phi_{\lambda_L}(x)$, satisfies the limit

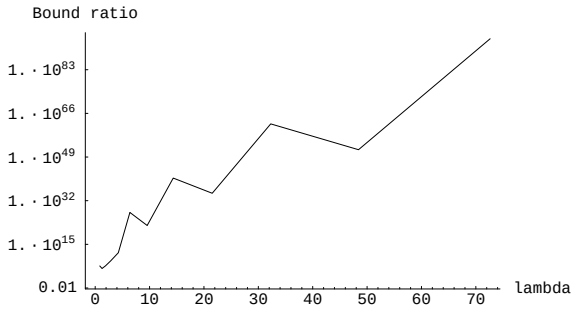

Fig. 1
THE RATIO BETWEEN UPPER AND LOWER BOUNDS OBTAINED BY TSYBAKOV AND GEORGANAS

$$\phi_{\lambda_L}(x)/x^{-\kappa_\lambda} \longrightarrow C_\lambda,$$

where κ_λ is the exponent corresponding to the minimal configuration of bursts which leads to overload and C_λ is a certain constant. Hence, for each relative accuracy, $\rho > 0$, there exists a buffer level, x_ρ, beyond which the approximation $C_\lambda x^{-\kappa_\lambda}$ for $\phi_{\lambda_L}(x)$ achieves this relative accuracy. Now consider the sequence of systems S_λ of Proposition III.1 and let $x_\rho^{\{\lambda\}}$ denote the buffer level at which relative accuracy ρ is achieved for system S_λ. Proposition III.1 implies that for any $\rho > 0$, as $\lambda \to \infty$, $x_\rho^{\{\lambda\}} \to \infty$. For, if $\{x_\rho^{\{\lambda\}} : \lambda > 0\}$ was bounded, we could choose any such bound as the value of x_0 in Proposition III.1, thereby contradicting its conclusion.

Thus, as λ increases, the point where the hyperbolic tail becomes accurate becomes more and more remote.

To estimate the point where the hyperbolic tail *starts*, let us denote, for a given buffer level, x, the number of simultaneous long bursts which is most likely to have occurred in order to give rise the given level by $\kappa(x)$. We expect this number to approach the number, $k(x)$ say, required to cause system overload, for large x. But how large does x have to be in order that $\kappa(x) = k(x)$?

To answer this question let us use the decoupling approach of [3], discussed in §II-B, using a Gaussian approximation for the process of short bursts to estimate the probability that the level x is exceeded, given a certain background of long bursts. Since the Poisson Pareto Burst Process has autocovariance very similar to Fractional Brownian Noise, it will be reasonable to use the well known [14] formula for the probability of exceeding the level x:

$$P\{V_\infty > x\} \approx e^{\left(-\frac{(C-m)^2 x^{2-2H}}{2(H)^{2H}(1-H)^{2-2H}\sigma^2}\right)}. \tag{17}$$

The probability that the level x is exceeded *and* this is achieved by precisely $k = \left\lfloor \frac{C-m}{r} \right\rfloor$ long bursts is therefore approximately

$$\frac{\left(\frac{\mathrm{E(d)}\lambda(t^*/\delta)^{1-\gamma}}{\gamma}\right)^k e^{\frac{-\mathrm{E(d)}\lambda(t^*/\delta)^{1-\gamma}}{\gamma}}}{k!} \times e^{\left(-\frac{(C-m-kr)^2 x^{2-2H}}{2(H)^{2H}(1-H)^{2-2H}\sigma^2}\right)}$$

while the probability that level x is exceeded and this is achieved by precisely $k-1$ simultaneous long bursts is approximately

$$\frac{\left(\frac{\mathrm{E(d)}\lambda(\mathrm{t}^*/\delta)^{1-\gamma}}{\gamma}\right)^{k-1} e^{-\frac{\mathrm{E(d)}\lambda(\mathrm{t}^*/\delta)^{1-\gamma}}{\gamma}}}{(k-1)!} \times e^{\left(-\frac{(C-m-(k-1)r)^2 x^{2-2H}}{2(H)^{2H}(1-H)^{2-2H}\sigma^2}\right)},$$

in which m denotes the normal mean load generated by the PPBP, $H=\frac{3-\gamma}{2}$, and t^* is the estimated length of the long bursts, which must be at least

$$t_x^* = \frac{Hx}{(1-H)(C-m-kr)}, \qquad x \geq 0, \tag{18}$$

because otherwise the FBM approximation could not be used. In order to select the maximum probability for the level x to be exceeded, we choose the minimum possible length for the long bursts, hence t^*.

It follows that the buffer level at which the hyperbolic behaviour of the tail first sets in, which we denoted by x_g earlier, is the solution (for x) of the following equation

$$\frac{\mathrm{E(d)}\lambda(\mathrm{t}^*/\delta)^{1-\gamma}}{\gamma \mathrm{k}} = e^{\left(\frac{\left((C-m-kr)^2-(C-m-(k-1)r)^2\right)x^{2-2H}}{2(H)^{2H}(1-H)^{2-2H}\sigma^2}\right)}. \tag{19}$$

A plot of the solution of (19) as a function of λ is shown in Figure 2. In this particular case ($\delta=1$, $\gamma=1.5$, $r=0.05$), the relationship between λ and the hyperbolic threshhold, x_g, is almost exactly $x_g = 33\lambda^2$. Although x_g, as revealed in this figure, is not *impossibly remote* in all cases investigated so far, it is sufficiently remote that the CDF takes negligable values by the time buffer levels are this high.

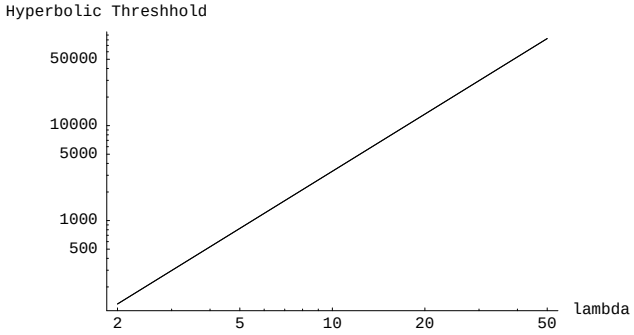

Fig. 2
THE THRESHOLD FOR HYPERBOLIC BEHAVIOR OF THE TAIL

C. Asymptotic Forms of the PPBP CDF in both Senses

The heuristic formula of the last subsection can also be viewed in another way which provides a more rigorous foundation. This formula is consistent with the known large buffer asymptotics of the PPBP queueing system and is also consistent with the known many sources asymptotics of this system. This might seem impossible, since one interpretation of the previous results is that they show the large buffer asymptotics to be inconsistent with the many sources asymptotic form of the CDF. However, asymptotic forms cannot be characterized by a single function. Both asymptotic forms should be properly thought of as an equivalence class of functions, and no matter how simple a function found in one of these classes might be, it is still only a representative from the class. The class of functions with the same large buffer asymptotics as the stationary CDF of the PPBP queue and the class of functions with the same many sources asymptotic form of this CDF have a non-empty intersection (for the true CDF is in both classes). Any simple function in both classes has a much better chance of characterising the true behaviour of the PPBP queueing system than a function which is only in one class or the other. The approximation obtained in Subsection II-B is in both classes.

IV. Concluding Remarks

We have shown that the widely known hyperbolic form queueing formulae for PPBP service systems, in many cases, will not give us an acceptable answer for the practical problems we seek answers for. What approach should we use instead? If the traffic is sufficiently multiplexed such that it is approximately Gaussian, it will be acceptable to use the results of [1] as the results there are appropriate for Gaussian SRD as well as LRD SSQ. Otherwise we can use the decoupling approach of [3], as discussed in Subsection II-B, where we classify the bursts into long and short bursts (equivalent to the so-called "mice and elepants" concept [5], [16]) and use the existing result for FBN queues for the SRD process. The accuracy of this approach appears to be good and it also appears to be quite straightforward to use it to calculate probabilities of interest.

References

[1] R. Addie, P. Mannersalo, and I. Norros. Performance formulae for queues with Gaussian input. In *Proceedings of ITC 16*, pages 1169–1178, June 1999.

[2] R. G. Addie. On weak convergence of long-range-dependent traffic processes. *journal of statistical planning and inference*, 80:155–171, 1999.

[3] R. G. Addie, T. D. Neame, and M. Zukerman. Performance evaluation of a queue fed by a Poisson Pareto burst process. *Computer Networks*, 40(3):377–397, Oct. 2002.

[4] R. G. Addie, M. Zukerman, and T. Neame. Broadband traffic modeling: simple solutions to hard problems. *IEEE Communications Magazine*, 1998.

[5] J. Boyer, F. Guillemin, P. Robert, and B. Zwart. Heavy tailed M/G/1-PS queues with impatience and admission control in packet networks. In *Proceedings of INFOCOM 2003*. IEEE, 2003.

[6] G. L. Choudhury, D. M. Lucantoni, and W. Whitt. On the effectiveness of effective bandwidth for admission control in ATM networks. In *Proceedings of ITC 14*. North Holland, 1994.

[7] G. L. Choudhury, D. M. Lucantoni, and W. Whitt. Squeezing the most our of ATM. *IEEE Transactions on Communications*, 44, 1996.

[8] N. G. Duffield and N. O;Connell. Large deviations and overflow probabilities for the general single-server queue, with applications. *Mathematical Pro-*

ceedings of the Cambridge Philosophical Society, 118:363–374, 1995.

[9] M. M. Krunz and A. M. Makowski. Modeling video traffic using M/G/∞ input processes: A compromise between Markovian and LRD models. *IEEE Journal on Selected Areas in Communications*, 16(5):733–748, June 1998.

[10] M. A. Lifshits. *Gaussian Random Functions*. Kluwer Academic Publishers, 1995.

[11] N. Likhanov and R. R. Mazumdar. Loss asymptotics in large buffers fed by heterogeneous long-tailed sources. *Advances in Applied Probability*, 32:1168–1189, 2000.

[12] N. Likhanov, B. Tsybakov, and N. D. Georganas. Analysis of an ATM buffer with self-similar ("fractal") input traffic. In *Proceedings, IEEE Infocom 1995*, pages 1–15. IEEE, April 1995.

[13] M. Mandjes. A note on queues with M/G/∞ input. *Operations Research Letters,*, 28:233–242, 2001.

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

[15] M. Parulekar and A. M. Makowski. Tail probabilities for M/G/∞ processes (I): Preliminary asymptotics. *Queueing Systems - Theory and Applications*, 27:271–296, 1997.

[16] V. Paxson and S. Floyd. Wide-area traffic: The failure of Poisson modeling. *IEEE/ACM Transactions on Networking*, 3(3):226–244, 1995.

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

[18] B. Tsybakov and N. D. Georganas. Overflow and losses in a network queue with a self-similar input. *Queueing Systems*, 35, 2000.