Computing Queue-Length Distributions for Power-Law Queues

Matthew Roughan Darryl Veitch Michael Rumsewicz
Software Engineering Research Centre
Level 2, 723 Swanston St, Carlton, Vic 3053, Australia
email: {matt,darryl,mpr}@serc.rmit.edu.au

Abstract: The interest sparked by observations of long-range dependent traffic in real networks has lead to a revival of interest in non-standard queueing systems. One such queueing system is the M/G/1 queue where the service-time distribution has infinite variance. The known results for such systems are asymptotic in nature, typically providing the asymptotic form for the tail of the workload distribution, simulation being required to learn about the rest of the distribution. Simulation however performs very poorly for such systems due to the large impact of rare events. In this paper we provide a method for numerically evaluating the *entire* distribution for the number of customers in the M/G/1 queue with power-law tail service-time. The method is computationally efficient and shown to be accurate through careful simulations. It can be directly extended to other queueing systems and more generally to many problems where the inversion of probability generating functions complicated by power-laws is at issue. Through the use of examples we study the limitations of simulation and show that information on the tail of the queue-length distribution is not always sufficient to answer significant performance questions. We also derive the asymptotic form of the number of *customers* in the system in the case of a service-time distribution with a regularly varying tail (eg infinite variance) and thus illustrate the techniques required to apply the method in other contexts.

I. Introduction

Recent analysis of network data (such as that of Local Area Networks, ISDN and Frame Relay) has demonstrated that traffic arrival processes in real systems have self-similar or long range dependent (LRD) or fractal properties [15]. Erramilli *et al.* [10] showed that such traffic can result in significant degradation of performance in real systems at loads much lower than predicted by traditional models having only short range dependence. In particular, large queues fed by long range dependent traffic may have *heavy tails*, implying that buffers dimensioned using conventional models may be significantly under-provisioned, and more seriously, that increasing buffer size is not a practical method to reduce loss. Furthermore, analysis of World Wide Web data has shown that the distribution of file sizes on the Internet has a heavy tail. In fact the weight in the tail is such that the distribution has a variance so large that it is best modelled as infinite [7].

Although far less tractable than traditional queueing systems, a body of analytical results are now available, a non-exhaustive list of which is [5], [16], [3], [20], [13], [19], [2]. These studies emphasize asymptotic properties, and in particular focus on the form of the *tail* of the workload distribution in infinite buffer queues. They are intrinsically incapable however of providing insight into the behaviour of the 'head' and 'body' of the queue. On the other hand simulation is poorly equiped to fill this gap, as in such systems events of relatively low probability can have an extremely high impact on queueing dynamics. This is especially true when infinite variance is involved, as in practice such distributions must be truncated to finite variance approximations, leading to results which can be markedly truncation level dependent. Simulations must therefore be constructed very carefully, and run over long periods of time before they converge, a fact noted in [9] and [14]. When, in addition, the simulation is intended to study the tail of a queue, additional care must be taken, and even longer runs are required, because the discussed difficulties are exaggerated in the tail.

The main contribution of this paper is the provision of a numerical method for calculating the *entire* distribution of the number of customers in the FIFO M/G/1 queue with service-time distribution having a heavy tail of power-law type. In particular we concentrate on the case with infinite variance. Although we focus on the FIFO M/G/1 system, the method can be extended to other queueing systems, for example the batch arrival M/G/1 queue, or applied to systems with imbedded M/G/1 queues such as the fluid queue fed by independent On/Off constant rate sources with heavy-tailed On times [2], [19]. More generally, the basic elements of the method can be applied to many systems, not necessarily queues, where the inversion of a known Probability Generating Function (PGF), rendered impossible by normal techniques because of the presence of power law behaviour, is required. We show the method to be computationally effective and accurate over a wide range of parameters, and capable of revealing the deficiencies of both the simulation and asymptotic analytic approaches. Indeed, the knowledge of the complete distribution is a vital tool in the testing of simulation methods in such difficult situations. Furthermore, investigations of the queue-length distributions calculated using this method have provided insight indicating that neither the tail nor head of the queue-length distribution is sufficient to answer all of the performance questions of interest, primarily because convergence to the tail behavior of the queue is extremely slow for some parameter values.

A further aim of the paper, and a necessary step in the achievement of the first, is to derive the form of the asymptotic *number* of customers in the M/G/1 queue above with power-law service-times. Our result, although essentially obtained through the application of existing results, illustrates the steps and tools which would be required in the application of our method to new systems which are not as well studied as the FIFO M/G/1 queue. Due to space restrictions, only a sketch of the derivation can be provided (see [17] for full details).

II. Background

We consider the M/GI/1 queue, that is, a queue with Poisson arrivals, independent generally distributed service-times, one server, and an infinite waiting room. This system, and many of its variants, have been carefully studied [4], [6], and many analytical results for these systems are well known. However, numerical solutions are not always easy to obtain from analytic results. Often such results are in the form of PGFs, from which it is easy to derive the first few moments of a distribution, the traditional quantities of interest. In the cases we consider here however where the first or second moments may not exist, the most useful performance measures can only be derived from the distribution itself, and so inversion of the PGF becomes a necessity.

A. Preliminaries

Throughout this paper we refer to several transforms which we now define. The *Probability Generating Function* (PGF) of a discrete probability distribution p_n on the non-negative integers is defined by

$$P^*(z) = \sum_{n=0}^{\infty} p_n z^n, \qquad z \in \mathbb{C} : |z| \leq 1.$$

We will also deal with the generating functions $H^*(z)$ of sequences $\{h_n\}$ which are defined in the same way but which do not correspond to probability distributions.

The *Laplace-Stieltjes transform* [11, p. 432] of a distribution with cumulative probability function $F(t)$, concentrated on $(0, \infty)$ is given by the Riemann-Stieltjes integral

$$\tilde{F}(s) = \int_0^{\infty} e^{-st} dF(t), \qquad s \in \mathbb{C} : \Re s \geq 0.$$

When $F(t)$ is differentiable the above transform becomes the ordinary Laplace transform of the probability density $f(t) = dF/dt$, that is $\tilde{F}(s) = \int_0^\infty e^{-st} f(t) dt$. Finally the *characteristic function* of $F(t)$ is given by the Fourier-Stieltjes transform [11, pp. 499-511]

$$\hat{\Phi}_F(\theta) = \int_0^{\infty} e^{i\theta t} dF(t), \qquad \theta \in \mathbb{R},$$

which reduces to the ordinary Fourier transform of f if it exists, and for a discrete distribution p_n reduces to the sum

$$\hat{\Phi}_p(\theta) = \sum_{n=0}^{\infty} p_n e^{i\theta n}.$$

In the latter case the characteristic function may be inverted through the integral [11, pp. 505-506]

$$p_n = \int_0^1 \hat{\Phi}_p(2\pi\theta) e^{-i2\pi\theta n} d\theta.$$

Apart from transforms, the other theory of interest in this paper is that of *regular variation*, which is that of asymptotic power-law behavior. First recall the notation $h(t) \overset{t_0}{\sim} g(t)$ for asymptotic equivalence, which means $\lim_{t \to t_0} |h(t)|/|g(t)| = 1$. Where t_0^- is used it refers to the limit as t_0 is approached from below. Now define functions as *slowly varying at* t_0 if they satisfy $\lim_{t\to t_0} L(xt)/L(t) = 1$, for every $x > 0$. We can now define a function $h(t)$ as being regularly varying at ∞, with index p, if

$$h(t) \overset{\infty}{\sim} L(t) t^p,$$

where $L(t)$ is slowly varying at ∞. A function $g(t)$ is regularly varying at zero if $g(t) = h(1/t)$ with $h(t)$ regularly varying at ∞. The obvious discrete analogy may be made for a sequence h_n.

Some of the key results we use are Tauberian theorems. Tauberian theorems are relations between the possible power-law behavior of the tail of a function with certain properties to the power-law behavior of its transform near the origin. We have need of the following two.

Theorem II.1: If $\{h_n\}$, $n = 0, 1, 2\ldots$ is an ultimately monotone positive sequence with generating function $H^*(z)$ that converges for $0 \leq z < 1$, and $0 \leq p < \infty$, then

$$h_n \overset{\infty}{\sim} \frac{L(n)}{\Gamma(p)} n^{p-1} \;\Leftrightarrow\; H^*(z) \overset{1^-}{\sim} L\left(\frac{1}{1-z}\right) \frac{1}{(1-z)^p},$$

where L is slowly varying at infinity.

Proof: See Feller [11, p. 447] □

We now extend the Tauberian result of Feller to obtain the following useful lemma.

Lemma II.1.1: Suppose that $\bar{U}_n$ has ultimately monotone difference $u_n = \bar{U}_{n-1} - \bar{U}_n$, for $n > 0$. If $\bar{U}_n \sim L(n) n^p$ with $p < 0$ then

$$u_n \sim -p \frac{\bar{U}_n}{n}.$$

Proof: See the detailed report [17] □

Also used extensively in this paper is the generalised Riemann zeta-function, which can be defined as [12, Equation 9.521]

$$\zeta(\alpha, q) = \sum_{n=0}^{\infty} \frac{1}{(q+n)^\alpha}.$$

When α is real and greater than 1 the function is well approximated by the lower bound

$$\zeta(\alpha, q) \approx \sum_{n=0}^{N} \frac{1}{(q+n)^\alpha} + \frac{(q+N+1)^{1-\alpha}}{\alpha - 1}, \tag{1}$$

where the positive error term $\varepsilon(N)$ satisfies $\varepsilon(N) < (q + N)^{-\alpha}$. For a derivation of this approximation see [17].

B. The M/G/1 queue

We consider the M/G/1 queue with Poisson arrival rate λ, mean service-time $1/\mu$, and traffic intensity $\rho = \lambda/\mu$. It is known [6, pp. 151–174] that the stationary PGF for the number of customers in the stable ($\rho < 1$) M/G/1 queueing system with the First In First Out (FIFO) service discipline is given by the Pollaczek-Khintchine equation

$$P^*(z) \quad = \quad (1-\rho)\frac{(z-1)\tilde{G}\left(\lambda(1-z)\right)}{z-\tilde{G}\left(\lambda(1-z)\right)}, \qquad (2)$$

where $\tilde{G}(s)$ is the Laplace-Stieltjes transform of the service-time distribution.

C. Daigle's method of PGF inversion

Daigle's method [8] of estimating the queue-length distribution relies on the general fact that the PGF of a discrete probability distribution evaluated on the unit circle is precisely the characteristic function of that distribution. That is, given a PGF $P^*(z)$ of a discrete distribution with masses p_n, $n = 0, 1, 2, \ldots$, the characteristic function of the distribution function is

$$\hat{\Phi}_p(2\pi\theta) = P^*(e^{i2\pi\theta}) = \sum_{n=0}^{\infty} p_n e^{i2\pi\theta n}.$$

As noted above, inversion may be performed by

$$p_n = \int_0^1 \hat{\Phi}_p(2\pi\theta)e^{-i2\pi\theta n}\,d\theta.$$

The integral may be numerically approximated by sampling the transform on the unit circle at $K+1$ points and summing. That is

$$c_n^{(K)} = \frac{1}{K+1}\sum_{k=0}^{K} P^*\left(e^{i2\pi k/(K+1)}\right)e^{-i2\pi nk/(K+1)},$$

for $n = 0, 1, \ldots, K$, which is just the inverse discrete Fourier transform, which can be efficiently calculated using the Inverse Fast Fourier Transform (IFFT).

The terms $c_n^{(K)}$ calculated from the PGF are shown by Daigle [8] to obey

$$c_n^{(K)} = p_n + \sum_{m=1}^{\infty} p_{n+m(K+1)}.$$

It is readily seen that the estimate $c_n^{(K)}$ of p_n is contaminated by *alias* terms, a typical effect of discrete sampling.

In the present context p_n is the stationary probability of n customers being in the queueing system. For large K the summation terms aliased into the $c_n^{(K)}$ are drawn from the tail of this queue. The asymptotic tail and hence the aliased terms can be calculated and subsequently subtracted from the IFFT estimates to give p_n. Daigle considers the case when the tail is asymptotically geometric, estimates the parameters of the tail, and subtracts the aliased terms through a standard geometric summation.

We consider the case when there is a power-law governing the tail behavior of service-times, and hence a power-law governing the tail behavior of the number of customers in the system. We then derive a formula for the sum of tail elements aliased into the IFFT terms and then remove them to recover the p_n.

III. The Method

This section contains the main results of the paper. First, we calculate the form of the tail of the queue-length distribution when the service-time distribution is regularly varying. We then present the main contribution: a method for inverting the PGF of the M/G/1 queue to obtain the *entire* queue-length distribution when this distribution has a power-law tail.

A. Calculating the tail of the queue

We assume that the service-time distribution $G(t)$ has a power-law tail, that is

$$1 - G(t) \stackrel{\infty}{\approx} L(t)t^{-\alpha}, \qquad (3)$$

where $L(t)$ is slowly varying at infinity, and $\alpha > 0$. Only the first n moments of such a distribution are finite, where $n = \lfloor \alpha \rfloor$, the largest integer less than α. In particular, when $\alpha \in (0, 1]$ all moments are infinite, and when $\alpha \in (1, 2]$ only the mean is finite. Henceforth only the latter case will be treated, although other cases with $\alpha > 2$ present no difficulties.

We wish to calculate the form of the tail of the queue-length distribution. The answer to the closely related question of the form of the workload distribution in the more general GI/G/1 case was answered some time ago by Cohen [5] (see also [3] for a waiting time calculation in a simple M/G/1 context). It states essentially that the waiting time distribution is regularly varying if and only if the service distribution is regularly varying, with an index which is one less (that is, there exists one less finite moment for the waiting time distribution). Rather than using Cohen's result as a basis for our calculation of the queue-length distribution, we derive it directly from the Tauberian theorems quoted in the previous section. In this way the basic tools needed to apply our inversion approach to other problems where PGF's are known will be illustrated, rather than restricting the first part of the method to a simple queueing context.

To determine the tail behaviour of the queue-length we need to know the behaviour near $z = 1$ of $P^*(z)$, and hence, from Equation (2), the form of the Laplace-Stieltjes distribution $\tilde{G}(s)$ for small s. From Bingham, Goldie and Teugels [1] (see also Brichet *et al.* [3]) this is given by

$$\tilde{G}(s) \quad \stackrel{0}{\sim} \quad 1 - \frac{s}{\mu} + L(1/s)\frac{\Gamma(2-\alpha)}{\alpha-1}s^{\alpha}, \qquad (4)$$

where $1/\mu$ is the mean service-time, and $\alpha \in (1, 2)$.

Denote the complementary distribution function of the queue by $\bar{P}_n = \sum_{k=1}^{\infty} p_{n+k}$. It is not difficult to show that its PGF is given by $\bar{P}^*(z) = (1 - P^*(z))/(1-z)$ and clearly $p_n = \bar{P}_{n-1} - \bar{P}_n$.

The following theorem uses the Laplace-Stieltjes transform above to derive the tail behavior of the distribution of the number of customers in the queue.

Theorem III.1: For the M/G/1 queue where the service-time distribution has a power law tail given by Equation(3) with $\alpha \in (1,2)$, the tail of the stationary queue-length distribution is given by

$$\bar{P}_n \overset{\infty}{\approx} \frac{\beta_n}{\alpha-1} n^{1-\alpha}, \tag{5}$$

and its discrete density by

$$p_n \overset{\infty}{\approx} \beta_n n^{-\alpha}, \tag{6}$$

for large n, where $\beta_n = L(n/\lambda)\lambda^\alpha/(1-\rho)$.

Proof: Setting $s = \lambda(1-z)$ in Equation (4) and then substituting into Equation (2), retaining only the first two terms yields

$$P^*(z) \overset{z \to 1^-}{\sim} 1 - \frac{K_1(z)}{1-\rho}(1-z)^{\alpha-1}.$$

where $K_1(z) = L(1/\lambda(1-z))\Gamma(2-\alpha)\lambda^\alpha/(\alpha-1)$. Now $\bar{P}^*(z) = (1-P^*(z))/(1-z)$ satisfies the conditions of Theorem II.1 with $p = 2-\alpha > 0$ and therefore

$$\bar{P}_n \overset{n \to \infty}{\sim} \frac{L(n/\lambda)\lambda^\alpha}{(\alpha-1)(1-\rho)} n^{1-\alpha}.$$

Applying Lemma II.1.1 to this relation we obtain the stated results for large n. □

Remark 1: The key step in the proof is obtaining a relation with an exponent in the appropriate range so that Theorem II.1 can be applied.

Remark 2: The above result holds for $\alpha \in (1,2)$, where the service-time distribution has finite mean, but infinite variance. Just as for the workload distribution [5], the result implies that the mean queue-length is infinite, and thus that the queue-length distribution has one less finite moment than the service-time distribution in this case.

B. Evaluating the entire distribution

For simplicity in the calculations below we shall consider the simple case where $L(t) = L$. Hence $\beta_n = \beta = L\lambda^\alpha/(1-\rho)$, a constant. More complicated cases can be treated in the same way, at the cost of replacing the Riemann zeta-function below with another, non-standard function expressed as an infinite sum, whose estimation may be slightly more costly and whose error more difficult to control.

It was seen in Section II-C that we can evaluate the $c_n^{(K)}$, which consist of p_n plus tail terms. If K is large enough so that $p_{K+1}, p_{K+2}, \ldots$, are well approximated by the power-law form in Theorem 5, then the sum of aliased terms may be evaluated. Thus

$$\begin{aligned} c_n^{(K)} &= p_n + \sum_{m=1}^{\infty} p_{n+m(K+1)} \\ &\overset{K \to \infty}{\sim} p_n + \beta \sum_{m=1}^{\infty} \Big(n + m(K+1)\Big)^{-\alpha} \\ &= p_n + \beta (K+1)^{-\alpha} \zeta\left(\alpha, \frac{n}{K+1} + 1\right), \end{aligned}$$

where $\zeta(\alpha, q)$ is a generalized Riemann zeta-function. We use the approximation to the Riemann zeta-function given in Equation (1) to write

$$\begin{aligned} c_n^{(K)} \overset{K \to \infty}{\sim} & \; p_n + \beta (K+1)^{-\alpha} \sum_{m=1}^{N} \left(\frac{n}{K+1} + m\right)^{-\alpha} \\ & + \frac{\beta\Big((N+1)(K+1)+n\Big)^{1-\alpha}}{(\alpha-1)(K+1)} + \varepsilon(N), \end{aligned}$$

where $\varepsilon(N) < (n/(K+1)+N)^{-\alpha}$, and thus

$$\begin{aligned} p_n \overset{K \to \infty}{\sim} & \; c_n^{(K)} - \beta \sum_{m=1}^{N} \Big(n + m(K+1)\Big)^{-\alpha} \\ & - \frac{\beta\Big((N+1)(K+1)+n\Big)^{1-\alpha}}{(\alpha-1)(K+1)} - \varepsilon(N). \end{aligned} \tag{7}$$

Although from Theorem III.1 we have an analytic expression for the value of β, it is useful to directly estimate it from Equation (7) with $n = 0$, using the fact that p_0 is known to equal $1-\rho$ [6]. The estimate is

$$\hat{\beta}(K) = \frac{(c_0^{(K)} - p_0)(K+1)^\alpha}{\sum_{m=1}^{N} m^{-\alpha} + \frac{(N+1)^{1-\alpha}}{\alpha-1}}.$$

The discrepancy between β and $\hat{\beta}(K)$ can be used to choose an appropriate value of K. More precisely, note that the absolute error in the calculation of the largest alias term is essentially $|\beta - \hat{\beta}| * (K+1)^{-\alpha}$, and we are therefore interested in minimizing this quantity with respect to K. Although in theory the error will decrease monotonically with K, in practice there will be an optimal K due to numerical errors. The following algorithm is an efficient way of choosing a suitable K.

(1) $i = 3$
(2) $K = 2^i - 1$
(3) `while` $|\beta - \hat{\beta}| * (K+1)^{-\alpha} > \delta$ `do`
    (3a) `evaluate` $c_o^{(K)} = \frac{1}{K+1}\sum_{k=0}^{K} P^*\left(e^{\frac{-2\pi i k}{K+1}}\right)$
    (3b) `determine` $\hat{\beta} = \frac{(c_0^{(K)} - p_0)(K+1)^\alpha}{\sum_{m=1}^{N} m^{-\alpha} + \frac{(N+1)^{1-\alpha}}{\alpha-1}}$
    (3c) $i = i + 1$
    (3d) $K = 2^i - 1$
(4) `Evaluate` $c_n^{(K)}$ `using the IFFT`
(5) `Recover the` p_n `by removing the aliased terms from` $c_n^{(K)}$ `using Equation (10)`

where N is determined by the required accuracy of the finite sum approximation to the generalized Riemann zeta-function. The algorithm may be made even more efficient by using the fact that, although at each iteration of the loop the number $(K+1)$ of sampling points of the PGF P^* increases by a factor of two, half of these have already been computed (and summed) in the previous iteration. It is possible that the optimum value of K can be skipped over, leading to divergence. A test for this possibility was added to the above algorithm.

IV. Two Numerical examples

This section provides some numerical examples of the method described above, both for verification through simulation, and to test its stability. The examples illustrate the calculation procedure, as well as the numerical behaviour of the method, for two service times distributions with very different behaviour around the origin. The examples were chosen to be simple to limit the number of parameters which must be chosen, and to have continuous distribution functions. As we shall see, they are nonetheless rich enough provide the opportunity to demonstrate a number of interesting new observations about power-law tail queues.

A. Example 1

The probability density function of a continuous heavy-tailed distribution suggested in [18] is given below

$$p_G(x) = \begin{cases} \alpha B^{-1} e^{-\frac{\alpha}{B}x}, & \text{for } x \le B, \\ \alpha B^{\alpha} e^{-\alpha} x^{-(\alpha+1)}, & \text{for } x > B, \end{cases}$$

where $B > 0$ marks where the tail 'begins', and $\alpha \in (1,2)$. The lemma of [11, p. 446] shows that the tail of the distribution function is $1 - G(x) \sim B^{\alpha} e^{-\alpha} x^{-\alpha}$, and standard arguments show that the mean is $E[X] = B\left\{1 + e^{-\alpha}/(\alpha - 1)\right\}/\alpha$, that the variance is infinite, and has Laplace-Stieltjes transform

$$G^*(s) \;=\; \alpha \left[\frac{1 - e^{-(sB+\alpha)}}{(sB + \alpha)}\right] + \alpha B^{\alpha} e^{-\alpha} s^{\alpha} \Gamma(-\alpha, sB),$$

where $\Gamma(x, z)$ is the incomplete gamma function [12, 8.350]. Furthermore the β_n in Theorem III.1, are constant and equal to

$$\beta \;=\; \frac{\left(B\lambda e^{-1}\right)^{\alpha}}{1-\rho}. \tag{8}$$

The Laplace-Stieltjes transform above may be used in the Pollaczek-Khintchine formula for the PGF of the number of customers in the system, and thence via the method described above, used to calculate the stationary distribution of the number of customers in the queue.

Figure 1 below shows examples of calculated results compared with simulated results, for three sets of parameters $(\alpha = 1.4, \rho = 0.3)$, $(\alpha = 1.8, \rho = 0.3)$, and $(\alpha = 1.5, \rho = 0.8)$, with $B = 4.0$ in all three cases. Straight lines in the log-log graph correspond to power-law curves, examples of which are the asymptotes of the three distributions calculated using Equations (6) and (8), shown in the figure as dashed lines. The short vertical lines mark the boundary between values calculated by the method (to the left) and those given by the analytic asymptotic tail.

Each simulation of the embedded process was based on 100,000,000 departures. For clarity only a representative sample of simulation points have been plotted on the figure. It is noteworthy that the simulation follows the predicted results until it reaches a probability of approximately 10^{-5}. At roughly this point each simulation diverges up from the

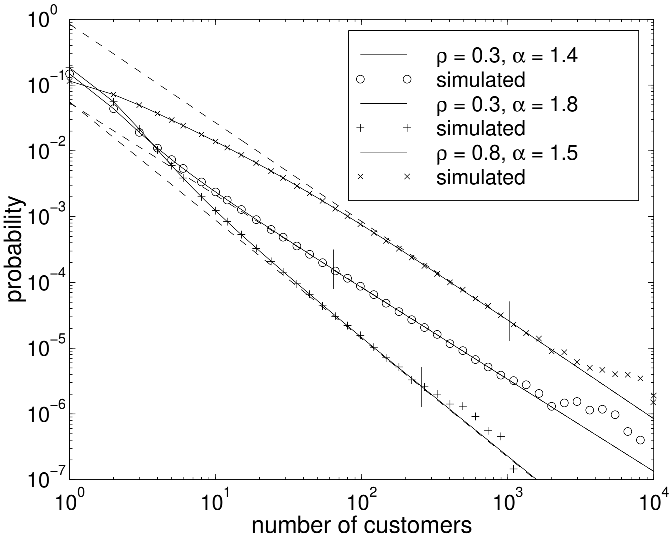

Fig. 1. Three examples of queue-length distributions are shown here, each showing the calculated results (solid line) the corresponding simulations (circles, crosses and x's) and the asymptotic tail (dashed line).

predicted result, and slightly later drops suddenly (not seen on figure). This divergence is a direct effect of the truncation of the service-time distribution to a finite variance distribution, and the point at which the simulation diverges can be altered by changing the truncation point. To shift the truncation point out further than here, much more sophisticated methods of simulation are required.

Simulation confidence intervals have not been shown on the graphs because they are so tight that they would not be visible. In fact the maximum of the standard deviation of the simulations results (based on ten simulations) was 1.7e-04. This is comfortably larger than the maximum absolute difference (taken over the entire queue-length considered before errors in the simulation enter in) between the simulation and calculation based values of p_n, which were 3.2e-05, 3.3e-05, and 5.9e-05 respectively.

An interesting check is provided by the fact that the mean value of this service-time is proportional to B, and thus, setting $B \to B' = cB$ we must take $\lambda' = \lambda/c$ to keep ρ constant. This transformation corresponds to a scaling of time, and therefore the stationary queue distribution should not be affected. We have checked this invariance numerically for $\rho = 0.3$, $\alpha = 1.5$, for the pairs $(B, \lambda) = (0.1, 3.11)$ and $(B, \lambda) = (100.0, 0.00311)$. The results for the two are identical up to numerical accuracy ($\approx 10^{-16}$), demonstrating that our method for calculating the queue-length distribution accurately reflects the invariance with respect to B.

B. Stability

Of interest when investigating any numerical algorithm is its performance over a range of parameters. The performance measures of interest are the accuracy and stability of the method. The accuracy of the method, as far as possible, has been verified through simulation, but the stability of the algorithm is not guaranteed over the entire range of parameters investigated here. There are some parameter values for which the number of required sampling points $K+1$ blows up, to the point where obtaining accurate results is computationally infeasible. This is shown below to occur in the 'heavy traffic' limit, but only in regimes which would be unrealistic for operation of any modern system.

α	$K+1$	time[a]	$p\left(Q > 10^{12}\right)$	$\lvert\beta-\hat{\beta}\rvert(K+1)^{-\alpha}$
1.9	8[b]	1.04	2.18e-11	6.93e-03
1.8	8[b]	0.92	3.57e-10	2.87e-03
1.7	10[b]	1.50	5.91e-09	2.31e-06
1.6	32[b]	1.84	9.91e-08	1.83e-05
1.5	2048	4.40	1.69e-06	3.18e-08
1.4	8192	3.44	2.92e-05	7.63e-08
1.3	32768	11.20	5.18e-04	6.47e-08
1.2	131072	41.10	9.48e-03	6.52e-08
1.1	524288	146.88	1.82e-01	6.60e-08

Table 1: Stability with respect to α, for $\rho = 0.8$.

ρ	$K+1$	time[a]	$p\left(Q > 10^{12}\right)$	$\lvert\beta-\hat{\beta}\rvert(K+1)^{-\alpha}$
0.1	128	0.43	1.66e-08	4.32e-08
0.2	256	0.51	5.27e-08	2.36e-08
0.3	256	0.54	1.11e-07	4.66e-08
0.4	256	0.53	1.99e-07	7.20e-08
0.5	256	0.54	3.33e-07	7.75e-08
0.6	256	0.51	5.48e-07	3.50e-08
0.7	1024	1.00	9.20e-07	2.40e-08
0.8	2048	4.40	1.69e-06	3.18e-08
0.9	4096	7.03	4.02e-06	7.58e-08

Table 2: Stability with respect to ρ, for $\alpha = 1.5$.

[a] The time indicates CPU time (in seconds) required by the algorithm, coded in MATLAB, running on a Sun Ultra Server 2.

[b] The cases where divergence forced the iteration to stop before the correct δ was reached. A search then found the best number of sampling points.

Table I illustrates the stability of the method with respect to α while $\rho = 0.8$ is kept constant. The table shows the number of sampling points chosen by the algorithm for a range of parameters. As $\alpha \to 1^+$ the number of sampling points grows exponentially, and the computation time required by the algorithm grows at a similar pace. As shown in *Table II* the number of sampling points also grows as $\rho \to 1^-$, but not nearly as quickly.

The instabilities arise because in the heavy traffic regime the queue-length distribution converges only slowly to its asymptotic tail behavior. An example of the slow convergence can be seen in *Figure 2*, which shows the calculated and simulated queue-length distribution for $\alpha = 1.2$ and $\rho = 0.8$ next to the exact asymptotic tail. The curves are converging, but very slowly, particularly in view of the log scale on the abcissa.

In order for the algorithm to work, K must be large enough so that the $p_K, p_{K+1}, \ldots,$ are well approximated by their asymptotic power-law tail. Obviously if the distribution converges only slowly to its asymptotic tail, which is the case here for small α, K must be very large as in *Figure 2*.

These instabilities do not negate the usefulness of the method because the heavy-traffic regimes in which they occur have unacceptably high loss rates even for *very* long buffers. *Table II* illustrates this fact by also giving the probability that the queue-length exceeds 10^{12} customers $p\left(Q > 10^{12}\right)$ derived by summing over the asymptotic tail of the distribution. The probability grows far more quickly than the number of sampling points, until for $\alpha = 1.1$, and $\rho = 0.8$ the probability is greater than 0.1. Furthermore the average waiting time in a such a queue (truncated at 10^{12} customers) would result in an unacceptably large delay.

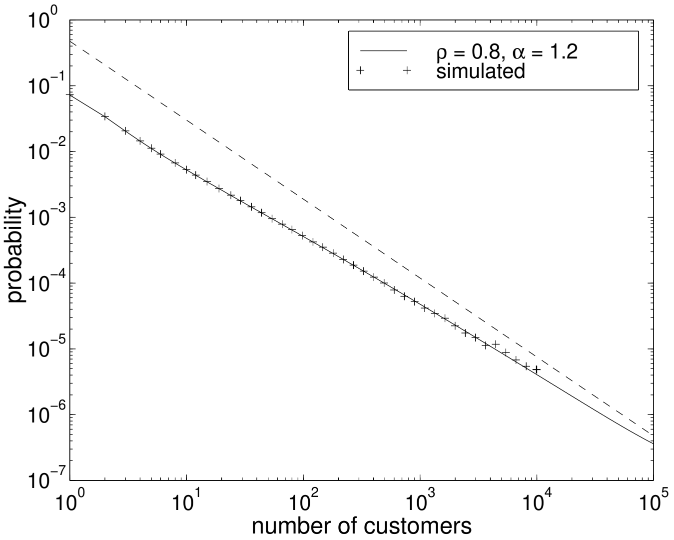

Fig. 2. An example which shows the queue-length distribution $\{p_n\}$ converging only slowly to its asymptotic behavior. The calculated results (solid line), the corresponding simulations (points), and the asymptotic tail (dashed line) are shown. The number of sampling points in this example $K + 1 = 131{,}072$.

Furthermore it is worth noting that the computation times in *Table II* (measured in seconds) are still orders of magnitude less than the time required by the simulations which, although written in C (rather than MATLAB) still took of the order of hours to complete.

The instability is in fact instructive because it illustrates well the limitations of both simulation and asymptotic results. Neither method provides the insight that the distribution may converge very slowly: asymptotic results are intrinsically unsuitable for such a deduction unless they include some idea of their rate of convergence, whilst simulation of the tail behavior is inherently untrustworthy due to the difficulty of performing such simulations correctly. Only a method which allows computation of the whole queue-length distribution could provide the valuable insight that the queue converges only slowly to its tail behavior.

In illustrating the slow convergence of the tail to its asymptotic behavior, *Figure 2* also demonstrates that measurements of tail behavior, based on real data or simulated, might be misleading. For instance, a linear least-squared-error estimate of the slope the simulated data in *Figure 2* results in an estimate for α of 1.03, whereas the actual value is 1.2.

A second problem with the method occurs when the divergence test stops the algorithm before the error in $\hat{\beta}$ has become small enough. Cases where this has happened include the examples in *Table I* with $\rho = 0.8$ and $\alpha = 1.6, \ldots, 1.9$. *Figure 3* compares the predictions with the simulations for the case $\rho = 0.8$ and $\alpha = 1.9$. The errors in this case are noticeable but not large compared with typical modeling errors such as uncertainties in parameter estimates. The reason(s) for this problem are not fully understood.

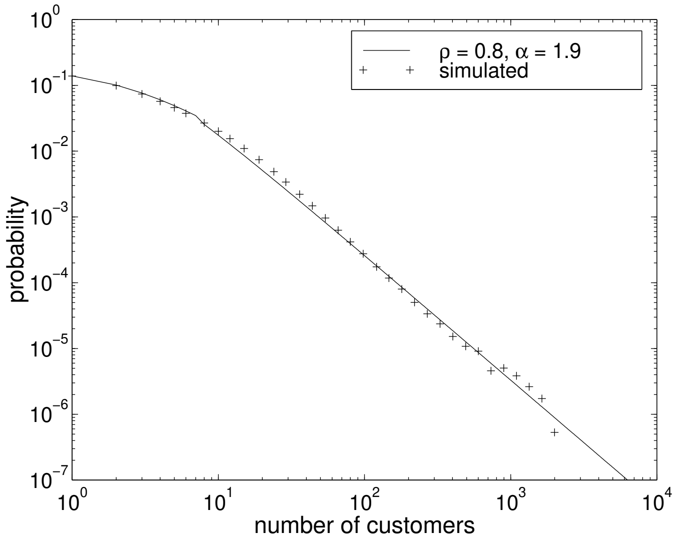

Fig. 3. The calculated values of $\{p_n\}$ (solid line) and the corresponding simulations (crosses) for $\alpha = 1.9$, $\rho = 0.8$, $B = 4.0$, $K + 1 = 8$.

C. Example 2

In this example we consider a case with $L(t)$ non-constant. Specifically the service-time distribution is given by

$$g(x) = \frac{b^\alpha e^{-b/x}}{\Gamma(\alpha) x^{\alpha+1}},$$

where $b > 0$ and $\alpha \in (1, 2)$. This density drops to zero at the origin and therefore has a maximum at some positive value, in contrast to the first example where the density was monotonically decreasing. The complementary distribution function [11, p. 446] is given by $1 - G(x) \sim b^\alpha e^{-b/x}/\Gamma(\alpha+1) \cdot x^{-\alpha}$, the mean by $b/(\alpha - 1)$, and it has infinite variance. The Laplace-Stieltjes distribution is given by [12, 3.471]

$$G^*(s) = \frac{2(bs)^{\alpha/2}}{\Gamma(\alpha)} K_\alpha(2\sqrt{bs}),$$

for $\Re s > 0$, where K_α is a Bessel function of an imaginary argument [12, Equation 8.407]. From Theorem III.1, and the fact that in this example $L(x) = b^\alpha e^{-b/x}/\Gamma(\alpha+1)$,

$$\begin{aligned} p_n &\overset{\infty}{\sim} \beta_n n^{-\alpha}, \\ \beta_n &= \frac{(b\lambda)^\alpha e^{-b\lambda/n}}{(1-\rho)\Gamma(\alpha+1)}. \end{aligned}$$

Although β_n is no longer constant, the summation terms in $c_n^{(K)}$ remain the same because $e^{-b\lambda/(n+m(K+1))} \overset{\infty}{\approx} 1$, and therefore the algorithm above may be used unmodified with

$$\beta = \frac{(b\lambda)^\alpha}{(1-\rho)\Gamma(\alpha+1)}. \tag{9}$$

Similar numerical results in terms of accuracy, stability, and the slow approach to tail behaviour for small α, are obtained for this second example. In addition it was found that queue-length distribution depends only on α and ρ, not upon b. An interesting feature is that when comparing the queueing distributions across the two examples with α, ρ fixed, very similar results are found despite the qualitatively different behaviours of the service distributions at the origin. Two examples of this are given in *Figure 4(a)* and *(b)*, where in each B was chosen to match the mean service-times. Note that the range of the figures has been deliberately truncated to highlight the differences between the distributions, which are not easily discernible on a figure with the same range as *Figure 1*.

V. Conclusion

A method has been presented for numerically calculating the entire stationary queue-length distribution for the M/G/1 queue when the service-time distribution has a regularly varying tail. Although we gave explicit results for the infinite variance case with a slowly varying function which was simply a constant, the method is essentially the same in the case of an arbitrary regularly varying tail with exponent $\alpha > 1$. The method is not restricted to the M/G/1 queue, but with minor modifications could be applied to any other queueing system for which the PGF of the distribution is known, for instance the batch arrival M/G/1 queue with power-law batch sizes. Even more generally, if a PGF is known but can not be inverted because of infinite moments, then it is likely that the approach illustrated here: first the calculation of the asymptotic behaviour of the PGF via Tauberian theorems, followed by the application of Daigle's method, can be followed through.

The method inverts the transform efficiently for most parameter values of interest, though instability of the method in the heavy-queueing regime reduces its effectiveness there. The heavy queueing regime, by which is meant a region in (α, ρ) space where the queueing tail has high mass, is of little interest for these systems in any case as it implies very high model loss probabilities, translating to high loss and/or long delays in real systems of interest.

A numerical method which gives the whole distribution is especially important in the case of power-law tail queues because

1. simulation of such systems is difficult and slow,

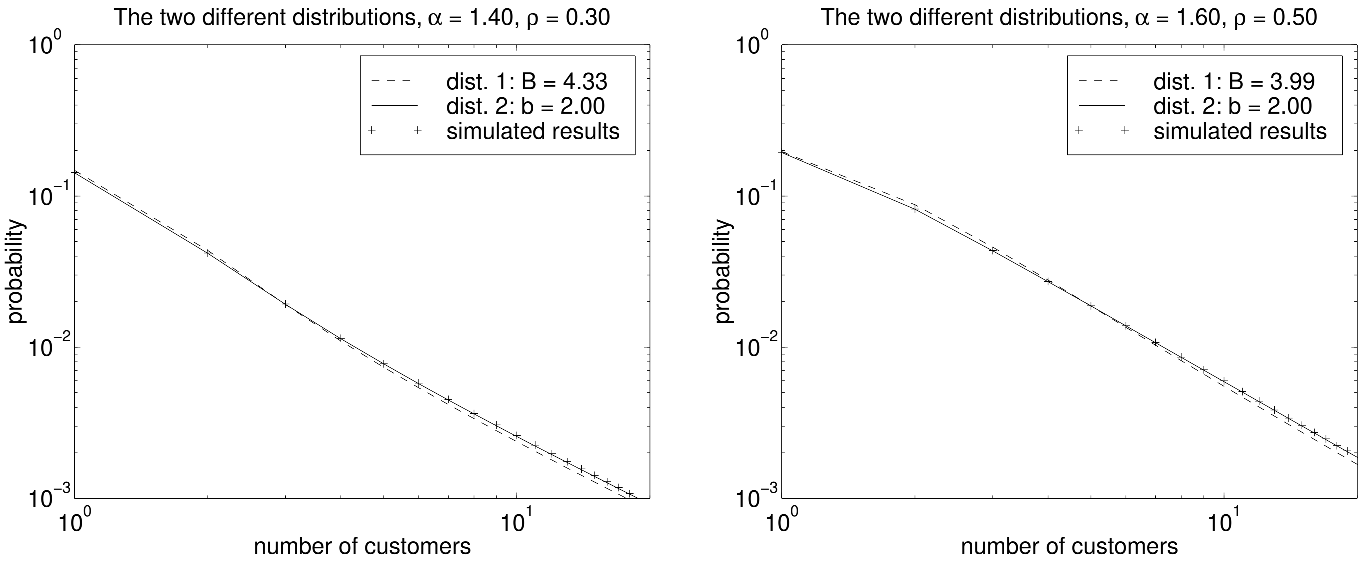

Fig. 4. A comparison of the results for the second distribution: calculated (solid line) and simulated (crosses), with parameters shown in the figures. The dashed line shows the results for the first distribution with the parameters chosen so that α and ρ are the same for both distributions.

2. the tail of the queue can contain a significant part of the mass of the queue-length distribution,
3. whilst the tail may be estimated using asymptotic results, slow convergence of the distribution to its asymptotic tail limits the practical applicability of the latter, and
4. understanding the pitfalls of simulations in such a difficult context requires knowledge of the real answer.

The second last point is of particular interest, as it may be a general feature of such systems which has not previously been discovered because of the reliance on asymptotic results or simulations to provide insight into power-law tail queues.

Another interesting feature noted in this study is that the head of the service-time distribution, including the point at which the power-law tail begins to dominate, has very little effect on both the head and tail of the queue-length distribution. The insensitivity to such short-range effects has been noted before in other contexts, for example in [14].

References

[1] N.H. Bingham, C.M. Goldie, and J.L. Teugels. *Regular Variation.* Cambridge University Press, Cambridge England, 1987.

[2] O.J. Boxma. Regular variation in a multi-source fluid queue. In V.Ramaswami and P.E.Wirth, editors, *Proceedings of the 15th International Teletraffic Congress - ITC 15; Teletraffic Contributions for the Information Age*, volume 2b, pages 391–402, Washinton, D.C., USA, June 1997. Elsevier, Amsterdam.

[3] F. Brichet, J. Roberts, A. Simonian, and D. Veitch. Heavy traffic analysis of a storage model with long range dependent on/off sources. *Queueing Systems*, 23:197–225, 1996.

[4] J.W. Cohen. *The Single Server Queue.* North-Holland, Amsterdam, 1969.

[5] J.W. Cohen. Some results on regular variation for the distributions in queueing and fluctuation theory. *Journal of Applied Probability*, 10:343–353, 1973.

[6] R.B. Cooper. *Introduction to Queing Theory.* The Macmillian Company, 1972.

[7] Mark E. Crovella and Azer Bestavros. Self-similarity in World Wide Web traffic: Evidence and possible causes. *IEEE/ACM Transactions on Networking*, 5(6), December 1997.

[8] John N. Daigle. Queue length distributions from probability generating functions via discrete fourier transforms. *Operations Research Letters*, 8:229–236, August 1989.

[9] Ashok Erramilli, James Gordon, and Walter Willinger. Applications of fractals in engineering for realistic traffic processes. In Jaques Labetoulle and James W. Roberts, editors, *Proceedings of the 14th International Teletraffic Congress - ITC 14*, volume 1a, pages 35–44. Elsevier, Amsterdam, 1994.

[10] Ashok Erramilli, Onuttom Narayan, and Walter Willinger. Experimental queueing analysis with long-range dependent packet traffic. *IEEE/ACM Transactions on Networking*, 4(2):209–223, April 1996.

[11] William Feller. *An Introducttion to Probability Theory and its Applications*, volume II. John Wiley and Sons, Brisbane, second edition, 1966.

[12] I.S. Gradshteyn and I.M. Ryzhik. *Table of integrals, series and products.* Academic Press, corrected and enlarged edition, 1980.

[13] Predrag R. Jelenkovic and Aurel A. Lazar. Asymptotic results for multiplexing subexponential on-off sources. *submitted to Advances in Applied Probability*, 1997.

[14] Predrag R. Jelenkovic and Aurel A. Lazar. Multiplexing on-off sources with subexponential on periods: Part ii. In V.Ramaswami and P.E.Wirth, editors, *Proceedings of the 15th International Teletraffic Congress - ITC 15; Teletraffic Contributions for the Information Age*, volume 2b, pages 965–974, Washinton, D.C., USA, June 1997. Elsevier, Amsterdam.

[15] Will E. Leland, Murad S. Taqqu, Walter Willinger, and Daniel V. Wilson. On the self-similar nature of ethernet traffic (extended version). *IEEE/ACM Transactions on Networking*, 2(1):1–15, Feb 1994.

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

[17] Matthew Roughan, Darryl Veitch, and Michael P. Rumsewicz. Numerical inversion of probability generating functions of power-law tail queues. Technical Report 0040, SERC, Software Engineering Research Centre, Level 2, 723 Swanston St, Carlton Vic, 3053, Australia, 1997.

[18] Bong K. Ryu and Steven B. Lowen. Point process approaches to the modelling and analysis of self-similar traffic - part i: Model construction. In *IEEE INFOCOM'96: The Conference on Computer Communications*, volume 3, pages 1468–1475, San Francisco, California, March 1996. IEEE Computer Society Press, Los Alamitos, California.

[19] A. Simonian and D. Veitch. A storage model with high rate and long range dependent on/off sources. preprint, 1997.

[20] Konstantinos P. Toukatos and Armand M. Makowski. Heavy traffic limits associated with M/G/1 input processes. preprint, 1997.