Queueing Systems 16 (1994) 311–338

Waiting-time tail probabilities in queues with long-tail service-time distributions

Joseph Abate

900 Hammond Road, Ridgewood, NJ 07450-2908, USA

Gagan L. Choudhury

AT&T Bell Laboratories, Room 1L-238, Holmdel, NJ 07733-3030, USA

Ward Whitt

AT&T Bell Laboratories, Room 2C-178, Murray Hill, NJ 07974-0636, USA

Received 26 February 1993; revised 13 October 1993

We consider the standard $GI/G/1$ queue with unlimited waiting room and the first-in first-out service discipline. We investigate the steady-state waiting-time tail probabilities $P(W > x)$ when the service-time distribution has a long-tail distribution, i.e., when the service-time distribution fails to have a finite moment generating function. We have developed algorithms for computing the waiting-time distribution by Laplace transform inversion when the Laplace transforms of the interarrival-time and service-time distributions are known. One algorithm, exploiting Pollaczek's classical contour-integral representation of the Laplace transform, does not require that either of these transforms be rational. To facilitate such calculations, we introduce a convenient two-parameter family of long-tail distributions on the positive half line with explicit Laplace transforms. This family is a Pareto mixture of exponential (PME) distributions. These PME distributions have monotone densities and Pareto-like tails, i.e., are of order x^{-r} for $r > 1$. We use this family of long-tail distributions to investigate the quality of approximations based on asymptotics for $P(W > x)$ as $x \to \infty$. We show that the asymptotic approximations with these long-tail service-time distributions can be remarkably inaccurate for typical x values of interest. We also derive multi-term asymptotic expansions for the waiting-time tail probabilities in the $M/G/1$ queue. Even three terms of this expansion can be remarkably inaccurate for typical x values of interest. Thus, we evidently must rely on numerical algorithms for determining the waiting-time tail probabilities in this case. When working with service-time data, we suggest using empirical Laplace transforms.

Keywords: Long-tail distributions, $GI/G/1$ queue, waiting time, tail probabilities, numerical transform inversion, computational probability, Pollaczek contour integrals, Pareto distributions, asymptotics.

1. Introduction and summary

In recent work we have focused on approximations for the steady-state waiting time tail probabilities in a single-server queue based on exponential

asymptotics; i.e.,

$$P(W > x) \sim \alpha e^{-\eta x} \quad \text{as } x \to \infty, \tag{1.1}$$

where W is the steady-state waiting time, η is a positive constant called the *asymptotic decay rate*, α is a positive constant called the *asymptotic constant*, and $f(x) \sim g(x)$ as $x \to \infty$ means that $f(x)/g(x) \to 1$ as $x \to \infty$; e.g., see Abate et al. [1, 2]. When (1.1) holds, it can serve as the basis for remarkably good approximations. We can approximate $P(W > x)$ directly by $\alpha e^{-\eta x}$ or we can obtain further approximations by approximating α and η. We can also use (1.1) as a basis for more refined approximations.

Our concern here is with the case in which (1.1) does *not* hold. Overall, it seems useful to identify *three cases*: The first case is when (1.1) holds. The second case is when W has a finite moment generating function, i.e., when $E e^{sW} < \infty$ for some positive s, but (1.1) does *not* hold. The third case is when W fails to have a finite moment generating function, i.e., when $E e^{sW} = \infty$ for all $s > 0$.

The second case may seem closely related to the first case, because the waiting-time tail probabilities are dominated by an exponential for large t, but we contend that the second case is actually closer to the third; see section 5 for more discussion. It may come as a surprise that there even is a second case. The second case is part of established theory in Borovkov [9] and Pakes [25]; it is illustrated by the $M/G/1$ model in example 5 of [1] for suitably small traffic intensities.

Here we primarily focus on the third case in which W does not have a finite moment generating function. We consider the $GI/G/1$ model; i.e., a single-server queue with unlimited waiting space, the first-in first-out (FIFO) service discipline and i.i.d. (independent and identically distributed) service times that are independent of i.i.d. interarrival times. In the $GI/G/1$ model we get the third case for W by having a long-tail service-time distribution, i.e., if we have a service time V for which $E e^{sV} = \infty$ for all $s > 0$. We are interested in this case because we believe that (i) long-tail service-time distributions can occur in applications, (ii) they might not be recognized, and (iii) they can have a great impact on queueing performance. It appears that long-tail distributions arise naturally when we encounter mixtures of distributions in very different time scales. For example, Duffy et al. [16, 17] report that statistical analysis of holding times of ordinary telephone calls has revealed a long-tail distribution, consistent with a Pareto-like tail, i.e., $P(V > x) - \alpha_r x^{-r}$ as $x \to \infty$ for r near 1. This is understandable when we recall that there now are modems for data traffic connected for many hours together with ordinary voice calls of average duration only a few minutes.

Meier-Hellstern et al. [24] also found that the distribution of *interarrival times* in one state of an ISDN traffic model also has a long-tail distribution. Our analysis indicates that long-tail interarrival times have much less impact upon waiting-time tail probabilities than long-tail service times. However, as pointed out in [24], there

are important practical consequences of long-tail interarrival-time distributions; e.g., sample moments can become poor descriptors.

A long-tail service-time distribution suggests that it might be better to use a different service discipline than FIFO, such as processor sharing. It also suggests that transient analysis may be more appropriate than steady-state analysis, because steady-state may be approached slowly. Nevertheless, in this paper we focus on the impact of a long-tail service-time distribution upon the distribution of the steady-state waiting time in the standard $GI/G/1$ model with the FIFO discipline.

It is significant that the long-tail phenomenon can easily be missed, e.g., if we consider only the first two moments of the distribution. Moreover, if we are interested in small steady-state waiting-time tail probabilities $P(W > x)$, then we might seriously underestimate them if we actually have a long-tail service-time distribution. To quickly illustrate, we display the steady-state waiting-time tail probabilities $P(W > x)$ for two $M/G/1$ models in table 1. (Asymptotic approximations also are displayed; they will be discussed later.) The exact values are computed using numerical transform inversion, as described in Abate and Whitt [4]. The arrival rate (and traffic intensity) is $\rho = 0.8$. Both models have two-parameter service-time distributions which we specify by stipulating that the first two moments are $m_1 = 1.0$ and $m_2 = 2.67$. One has a gamma service-time distribution, while the second has a long-tail distribution, in particular, the distribution in (2.5) below with parameter $r = 3.0$; also see (2.14). Its third and higher moments are infinite. From table 1, we see striking differences in the tail probabilities $P(W > x)$ for the

Table 1
The steady-state waiting-time tail probabilities $P(W > x)$ in two $M/G/1$ queues having service-time distributions with common first two moments $m_1 = 1$ and $m_2 = 2.67$ and common arrival rate $\rho = 0.8$. The approximation based on the asymptotics is also shown for the long-tail service-time distribution.

	Gamma service time	Long-tail service time with $r = 3.0$	
x	exact by [4] and asymptotics in (1.1)	exact by [4]	approximation from asymptotics in (1.5)
10	1.797×10^{-1}	1.675×10^{-1}	8.296×10^{-2}
20	4.122×10^{-3}	4.673×10^{-2}	1.481×10^{-2}
30	9.454×10^{-3}	1.573×10^{-2}	5.706×10^{-3}
40	2.168×10^{-3}	6.407×10^{-3}	2.963×10^{-3}
50	4.974×10^{-4}	3.143×10^{-3}	1.801×10^{-3}
60	1.141×10^{-4}	1.804×10^{-3}	1.207×10^{-3}
70	2.617×10^{-5}	1.165×10^{-3}	8.638×10^{-4}
80	6.002×10^{-6}	8.171×10^{-4}	6.481×10^{-4}
90	1.377×10^{-6}	6.075×10^{-4}	5.040×10^{-4}
100	3.158×10^{-7}	4.708×10^{-4}	4.030×10^{-4}

two service distributions for x values typically of interest, e.g., from $x = 10$ to $x = 100$.

The presence of long-tail distributions has other significant side effects. For example, if the service times fail to have a finite fourth moment, then the sequence of waiting times will have *long-range dependence* in the sense that

$$n \operatorname{Var}(\bar{W}_n) \to \infty \quad \text{as } n \to \infty,$$

where $\bar{W}_n$ is the sample mean of the first n waiting times and Var is the variance, e.g., see §3 of Whitt [30] and references cited there. This long-range dependence can have a number of consequences, one of which is that it invalidates standard (normal-theory based) procedures of statistical analysis, e.g., confidence intervals. Indeed, the steady-state mean might not even be finite. In applications, this phenomenon is manifested by erratic unreliable estimates, and sample means growing as the size of the data set increases when the mean is infinite. Of course, when the possibility of long-tail service-time distributions is recognized, appropriate statistical procedures can be devised. See [16, 17] for further discussion.

We have recently developed algorithms for calculating the waiting-time tail probabilities in a large class of single-server queues. For the $M/G/1$ queue, we use direct numerical transform inversion, as described in Abate and Whitt [4]. For the $GI/G/1$ queue, we again use numerical transform inversion as in [4], after obtaining the transform values by numerically integrating the classical Pollaczek [27, 28] contour integrals; see Abate et al. [3]. For $BMAP/G/1$ queues with batch Markovian arrival processes (BMAPs), we combine matrix analytical methods with numerical transform inversion, as described in Abate et al. [1, 2] and Choudhury et al. [11, 12]. Hence, we can compute the waiting-time tail probabilities provided that we can compute values of the Laplace transform of the service-time distribution.

An alternative to detailed computation is to exploit asymptotics, as in (1.1). For the $GI/G/1$ queue, asymptotics for this long-tail case is given in Cohen [14], theorem 1 of Pakes [25] and theorem 12 on p. 132 of Borovkov [9]. To state one such result, let U and V be generic interarrival and service times, and let $G_a(x) = P(U \le x)$ and $G_s(x) = P(V \le x)$, $x \ge 0$, be their cdf's (cumulative distribution functions). For any cdf F with finite mean m let the *complementary cdf* be $F^c(x) = 1 - F(x)$ and let the associated *stationary-excess cdf* (or equilibrium residual-lifetime cdf) be

$$F_e(x) = \frac{1}{m} \int_x^\infty F^c(y)\, dy, \quad x \ge 0. \tag{1.2}$$

Throughout this paper we will assume that $EV = 1 < EU < \infty$, so that $\rho \equiv EV/EU < 1$. We also need a technical regularity condition on the service-time cdf (consistent with $E\, e^{sV} = \infty$ for all $s > 0$). Following Pakes [25], we assume that G_s is a *subexponential* cdf; i.e., if $G_{s2}(x)$ is the convolution of $G_s(x)$ with itself,

then

$$G_{s2}^c(x)/G_s^c(x) \to 2 \quad \text{as } x \to \infty. \tag{1.3}$$

For practical purposes, the subexponential property can be regarded as equivalent to $E\,e^{sV} = \infty$ for all $s > 0$. The subexponential property implies that $E\,e^{sV} = \infty$; see appendix 4 of Bingham et al. [8].

THEOREM 1 (Pakes)

Consider a $GI/G/1$ queue, where $\rho < 1$, $EV = 1$, $E\,e^{sV} = \infty$ for all $s > 0$ and $G_s(x)$ has the subexponential property. Then

$$P(W > x) \sim \frac{\rho}{1-\rho} G_{se}^c(x) \equiv \frac{\rho}{1-\rho} \int_x^\infty G_s^c(y)\mathrm{d}y \quad \text{as } x \to \infty. \tag{1.4}$$

Given the great success with (1.1) as an approximation in the first case, it is natural to consider (1.4) as a basis for simple approximations in the third case. For further simplification, we exploit the following lemma; see p. 18 of Erdélyi [17].

LEMMA 1

If $f(x) \sim g(x)$ as $x \to \infty$, then $\int_x^\infty f(y)\,\mathrm{d}y \sim \int_x^\infty g(y)\,\mathrm{d}y$ as $x \to \infty$.

Lemma 1 implies that we can obtain the asymptotics for the service-time stationary-excess cdf G_{se} in (1.4) directly from the asymptotics for the service-time cdf G_s itself. We can also go one step further and start with the density if that is convenient.

We shall focus on the case in which $P(V > x) \sim \alpha_r x^{-r}$ as $x \to \infty$ for $r > 1$. This is a case in which the departure from (1.1) is dramatic (e.g., in contrast to $P(V > x) \sim e^{-\alpha x^b}$ as $x \to \infty$ for b less than but near 1, as with some Weibull distributions). From theorem 1 and lemma 1 we obtain the following result.

COROLLARY

If $EV = 1$ and $P(V > x) \sim \alpha_r x^{-r}$ as $x \to \infty$ for $r > 1$, then

$$P(W > x) \sim \left(\frac{\rho}{1-\rho}\right)\left(\frac{\alpha_r}{r-1}\right) x^{-(r-1)} \quad \text{as } x \to \infty. \tag{1.5}$$

A remarkable feature of (1.4) and (1.5) is that, unlike (1.1), *the asymptotic behavior depends on the interarrival-time distribution only through its mean.* However, we know that, with case-1 service-time distributions, the waiting-time tail

probabilities depend strongly on the interarrival-time distribution beyond its mean (e.g., on the second moment or, equivalently, the SCV (squared coefficient of variation), which we denote by c_a^2); e.g., for the $GI/M/1$ queue, see Whitt [29]. Hence, from the outset, we should be highly suspicious of (1.4) and (1.5) as approximations.

Indeed, *we show that (1.4) and (1.5) can be very poor approximations for typical x values of interest.* In fact, our experience indicates that the approximation provided by (1.1) is *typically good*, while the approximation provided by (1.5) is *typically bad.* Moreover, we show that the approximation for $P(W > x)$ provided by (1.5) can be *too small*, so that it is *not conservative.* This is illustrated by table 1. The asymptotic approximation based on (1.1) with the gamma service-time distribution is not displayed, because it is *exact in the stated four-digit precision*, whereas the asymptotic approximation based on (1.5) is not yet very close by $x = 100$. Moreover, as stated above, the approximation based on (1.5) *underestimates* the exact value.

One approach to this difficulty with the asymptotics in this case is to develop refined asymptotics with more terms, as has been done for the $M/G/1$ queue by Willekens and Teugels [31]. Indeed, we show that such refined asymptotics for the $M/G/1$ queue can also be developed using Laplace transforms. Unfortunately, however, we also show that even three terms can be remarkably inaccurate for typical x values of interest. Hence, it appears that it is preferable to directly calculate the tail probabilities.

In order to compute the steady-state waiting-time tail probabilities with a long-tail service-time distribution by our algorithms, we need to be able to compute the Laplace transform values for the service-time distribution. Unfortunately, however, explicit Laplace transforms do not seem available for familiar long-tail distributions, such as lognormal, Pareto or Weibull, see chapters 14, 19 and 30 of Johnson and Kotz [23]. This motivated us to introduce (what appears to be) a new family of long-tail distributions with explicit Laplace transforms. These are Pareto mixtures of exponential (PME) distributions. We believe that this family can be useful for queueing applications, and perhaps elsewhere as well.

We also suggest an alternative approach. Given any service-time distribution, we suggest constructing an approximating distribution with finite support, and then directly calculating its Laplace transform. For example, given any cdf G on $[0, \infty)$ and any n points x_i, $0 \le i \le n$ with $x_0 = 0$, $x_i < x_{i+1}$ and $x_n = \infty$, we can construct probability mass functions p_l, p_m and p_u defined by

$$p_l(x_i) = G(x_{i+1}) - G(x_i) \qquad 0 \le i \le n-1, \tag{1.6}$$

$$p_m((x_i + x_{i+1})/2) = G(x_{i+1}) - G(x_i), \qquad 0 \le i \le n-2,$$

$$p_m(x_{n-1}) = 1 - G(x_{n-1}), \tag{1.7}$$

$$p_u(x_i) = G(x_i) - G(x_{i-1}), \qquad 1 \le i \le n-2,$$

$$p_u(x_{n-1}) = 1 - G(x_{n-2}). \tag{1.8}$$

Note that the cdf G_l associated with p_l in (1.6) is a *stochastic lower bound* for G, i.e., $G_l^c(x) \leq G^c(x)$ for all x, while the cdf G_u is *almost* a stochastic upper bound for G. (It would be except for the upper tail above x_{n-1}.) If we consider successive refinements of the set of points $\{x_i\}$, then these three cdf's G_l, G_m and G_u will all converge to the original cdf G, while the cdf's G_l will converge from below (in stochastic order) to the cdf G. The associated waiting-time cdf's will also converge to the true waiting time cdf (provided that G has a finite mean), and also from below for G_l; see §§11, 21, 24 of Borovkov [9].

Given a probability mass function with finite support $\{y_i : 1 \leq i \leq n\}$, we can directly calculate the Laplace transform as

$$\sum_{i=1}^{n} p(y_i)\, e^{-sy_i}. \tag{1.9}$$

Given data, e.g., a sample of n service times $\{y_i\}$, we can directly construct the *empirical Laplace transform*, which is also given by (1.9). This approach is related to Gaver and Jacobs [20], but they assume and exploit (1.1). With data, we might choose to do some smoothing or approximate further to reduce the number of points.

We show that this finite approximation scheme is effective by applying it with the long-tail distributions for which we have explicit Laplace transforms. From our analysis with finite approximations, we infer that it is also possible to suitably approximate long-tail distributions with other classes of distributions, such as phase-type distributions. The important point is to match the true service-time distribution in the region where it has a significant effect on the waiting-time probability of interest. The $GI/G/1$ queueing theory suggests that, if we want to predict $P(W > x)$ for some given x, then we might try to match the complementary stationary-excess service time distribution $F_e^c(y)$ in (1.2) for $0 \leq y \leq x$. (See (1.4) and (4.2), for example.) From (1.2), we see that it is critical to get the correct service-time mean, but that we might otherwise aim for appropriate percentiles rather than higher moments.

Here is how the rest of the paper is organized. In section 2 we introduce the new two-parameter family of long-tail distributions, and discuss its properties. In section 3 we develop multi-term asymptotics for the $M/G/1$ waiting-time distribution with long-tailed service-time distribution. Our approach exploits Laplace transforms, and so is different from the previous procedures, in particular, Willekens and Teugels [31]. In section 4 we discuss numerical examples. In section 5 we discuss the second case in which $E\, e^{sW} < \infty$ for some $s > 0$, but (1.1) does not hold. In section 5 we show that any case-3 service time distribution can be converted to a case-2 service-time distribution simply by exponential damping. As illustrated by example 5 of [1], our experience is that the asymptotic approximations for the second case also perform poorly. Finally, in section 6 we draw conclusions.

2. A convenient family of long-tail distributions: Pareto mixtures of exponentials

We seek a family of cdf's $F_r(x)$ with tail behavior

$$F_r^c(x) \sim \alpha_r x^{-r} \quad \text{as } x \to \infty, \tag{2.1}$$

mean 1 and tractable Laplace transform. To have finite mean, we need $r > 1$. If we would be content with distributions without moments, then we could use the non-negative stable laws with transforms e^{-s^α}, $0 < \alpha < 1$, as on p. 448 of Feller [19], but we are primarily interested in distributions with moments. We want the service-time distribution to have a finite mean, so that W has a proper distribution.

A familiar family satisfying (2.1) is the *Pareto distribution*; see chapter 22 of Johnson and Kotz [23]. The Pareto complementary cdf is

$$F_r^c(x) = \left(\frac{r-1}{r}\right)^r x^{-r}, \quad x \geq (r-1)/r. \tag{2.2}$$

and its density is

$$f_r(x) = r\left(\frac{r-1}{r}\right)^r x^{-(r+1)}, \quad x \geq (r-1)/r. \tag{2.3}$$

The Pareto distribution has moments

$$m_n' = \frac{r}{(r-n)}\left(\frac{r-1}{r}\right)^n, \quad 1 \leq n < r, \tag{2.4}$$

so that its squared coefficient of variation (SCV) is $c^2 = 1/r(r-2)$. This Pareto family is not too attractive because it does not allow small values, but modifications do. However, the Laplace transform of this distribution (its density) and its standard modifications are not expressible in terms of elementary functions.

We thus propose a new modification of (2.2), namely, a *Pareto mixture of exponentials (PME)*. For $r > 1$, let the PME density be

$$g_r(x) = \int_{(r-1)/r}^{\infty} f_r(y) y^{-1} e^{-x/y}\, dy \tag{2.5}$$

for f_r in (2.3). From (2.4) and (2.5), we see that the moments of g_r are

$$m_n \equiv m_n(g_r) = n! m_n' = n! \frac{r}{(r-n)}\left(\frac{r-1}{r}\right)^n. \tag{2.6}$$

e.g.,

$$m_1 = 1, \quad m_2 = \frac{2r}{(r-2)}\left(\frac{r-1}{r}\right)^2, \quad m_3 = \frac{6r}{(r-3)}\left(\frac{r-1}{r}\right)^3 \tag{2.7}$$

and the SCV is

$$c_r^2 = 1 + \frac{2}{r(r-2)}. \tag{2.8}$$

By performing a change of variables (x to x^{-1}), we can also represent the density as

$$g_r(x) = \int_0^{r/(r-1)} \phi_r(y) y \, e^{-yx} \, dy, \tag{2.9}$$

where

$$\phi_r(y) = r\left(\frac{r-1}{r}\right)^r y^{r-1}, \quad 0 < y < r/(r-1). \tag{2.10}$$

Let G_r and G_{re} be the cdf and stationary-excess cdf associated with the density g_r. We now express the complementary cdf G_r^c and the complementary stationary-excess cdf G_{re}^c directly in terms of the density g_r. This enables us to apply theorem 1 without doing any integration. The following can be deduced from (2.9); we omit the proof.

THEOREM 2

For $r > 1$,

$$G_r^c(x) = a_r g_{r-1}(a_r x), \quad x \geq 0, \tag{2.11}$$

and

$$G_{re}^c(x) = \int_x^\infty G_r^c(y) \, dy = b_r g_{r-2}(c_r x), \quad x \geq 0, \tag{2.12}$$

where

$$a_r = \frac{r(r-2)}{(r-1)^2}, \quad b_r = \frac{(r-1)(r-3)}{(r-2)^2} \quad \text{and} \quad c_r = \frac{r(r-3)}{(r-1)(r-2)}. \tag{2.13}$$

Since G_r is a Pareto mixture of exponential cdf's, it is completely monotone, so that the density g_r is monotone. For r integer, the density g_r is easily expressed in closed form; see (4.2.55) on p. 71 of Abramowitz and Stegun [5]. For $r = 2, 3$ and 4, the densities and complementary cdf's are

$$\begin{aligned}
g_2(x) &= \frac{1}{x^3}(1-(1+2x+2x^2)\,e^{-2x}), && x>0,\\
g_3(x) &= \frac{16}{3x^4}\left(1-\left(1+\frac{3x}{2}+\frac{9x^2}{8}+\frac{9}{16}x^3\right)e^{-3x/2}\right), && x>0,\\
g_4(x) &= \frac{243}{8x^5}\left(1-\left(1+\frac{4x}{3}+\frac{8x^2}{9}+\frac{128x^3}{9}+\frac{32x^4}{243}\right)e^{-4x/3}\right), && x>0,\\
G_2^c(x) &= \frac{1}{2x^2}(1-(1+2x)\,e^{-2x}), && x>0,\\
G_3^c(x) &= \frac{16}{9x^3}\left(1-\left(1+\frac{3x}{2}+\frac{9x^2}{8}\right)e^{-3x/2}\right), && x>0,\\
G_4^c(x) &= \frac{243}{32x^4}\left(1-\left(1+\frac{4x}{3}+\frac{8x^2}{9}+\frac{32x^3}{81}\right)e^{-4x/3}\right), && x>0.
\end{aligned} \tag{2.14}$$

We now describe the asymptotics of PME distributions. For r integer, the asymptotics follow easily from the explicit representation above. For non-integer r, we exploit properties of the incomplete gamma function, §6.5 of Abramowitz and Stegun [5]. For this purpose, let $\Gamma(z)$ be the gamma function and $\gamma(a,x)$ the incomplete gamma function in (6.1) and (6.5.2) of [5].

THEOREM 3

For $r > 1$,

$$\begin{aligned}
G_r^c(x) &\sim rF_r^c(x)\gamma(r, xr/(r-1)) && \text{as } x\to\infty\\
&\sim \Gamma(r+1)F_r^c(x) = \Gamma(r+1)\left(\frac{r-1}{r}\right)^r x^{-r} && \text{as } x\to\infty
\end{aligned} \tag{2.15}$$

and

$$G_{re}^c(x) \sim \Gamma(r)\left(\frac{r-1}{r}\right)^{r-1} x^{-(r-1)} \qquad \text{as } x\to\infty. \tag{2.16}$$

Proof

Use (6.5.2), (6.5.3) and (6.5.32) in [5] together with (2.9). Use lemma 1 to get the second asymptotic relation. □

It is important to note that the next term in the asymptotics $r\gamma(r, xr/(r-1)) \sim \Gamma(r+1)$ as $x \to \infty$ is exponentially small, so that further refinements to (2.15) and (2.16) do not help much.

A crucial point for our purposes is that we can express the Laplace transform of the density g_r conveniently. In particular,

$$\hat{g}_r(s) \equiv \int_0^\infty e^{-sx}\, dG_r(x) = \int_0^\infty e^{-sx} g_r(x)\, dx$$

$$= r\left(\frac{r-1}{r}\right)^r \int_0^{r/(r-1)} \frac{x^r}{s+x}\, dx. \tag{2.17}$$

For integer r, the expansion for the integral is given on p.58 of Gradshteyn and Ryzhik [21]. It implies that

$$\hat{g}_r(s) = \sum_{i=1}^{r} (-1)^{r-i} \frac{r}{i} \left(\frac{r-1}{r}\right)^{r-i} s^{r-i} + (-1)^r r \left(\frac{r-1}{r}\right)^r s^r \ln\left(1 + \frac{r}{(r-1)s}\right). \tag{2.18}$$

For example,

$$\hat{g}_2(s) = 1 - s + \frac{s^2}{2} \ln\left(1 + \frac{2}{s}\right),$$

$$\hat{g}_3(s) = 1 - s + \frac{4}{3}s^2 - \frac{8}{9}s^3 \ln\left(1 + \frac{3}{2s}\right),$$

$$\hat{g}_4(s) = 1 - s + \frac{4}{3}s^2 - \frac{27}{16}s^3 + \frac{81}{64}s^4 \ln\left(1 + \frac{4}{3s}\right). \tag{2.19}$$

Formula (2.18) is fine for computation when $|s|$ is small, but if s is large, then it can cause difficulties. Since the first r terms are exactly cancelled by the first r terms in the Taylor series expansion of the logarithm in (2.18) in powers of s^{-1}, we obtain the alternate form

$$\hat{g}_r(s) = \sum_{i=0}^{\infty} (-1)^i \frac{r}{r+1+i} \left(\frac{r}{r-1}\right)^{i+1} s^{-(i+1)} \tag{2.20}$$

to use for large $|s|$. When $|s|$ is large, it is appropriate to use a truncated version (e.g., the first several terms) of (2.20). We use an automatic algorithm that truncates the series whenever the most recent term is below 10^{-12}.

We can also give an explicit representation for the integral (2.17) when r is of the form $n + 1/2$ for n integer. In this case,

$$\hat{g}_r(s) = 1 - s + \frac{m_2}{2}s^2 - \ldots + \frac{(-1)^n m_n}{n!}s^n + (-1)^{n+1}2r\left(\frac{r-1}{r}\right)^r s^r \operatorname{Arctan}(\sqrt{r/(r-1)}s), \tag{2.21}$$

where m_k is the kth moment in (2.6). We obtain (2.21) from (2.17) by applying the following relation between the transforms for different values of r and 2.213.1 on p. 71 of Gradshteyn and Ryzhik [21]. For numerics involving (2.21), see (4.4.3.9) of [5]. We obtain the following from theorem 2.

THEOREM 4

For $r > 1$,

$$\hat{g}_{r+1}(s) = 1 - s\hat{g}_r(d_r s),$$

where $d_r = r^2/(r+1)(r-1)$.

For example,

$$\hat{g}_{2.5}(s) = 1 - s + \frac{9s^2}{5} - 5(0.6s)^{2.5} \operatorname{Arctan}(\sqrt{5/3}s). \tag{2.22}$$

From theorem 3, for $r = m + 1/2$,

$$G_r^c(x) \sim \frac{(1 \cdot 3 \cdot 5 \cdots (2m+1))}{2^{m+1}} \sqrt{\pi} F_{2.5}^c(x), \tag{2.23}$$

so that

$$G_{2.5}^c(x) \sim \frac{27}{40}\sqrt{0.6\pi}x^{-2.5} \quad \text{and} \quad G_{2.5e}^c(x) \sim \frac{9\sqrt{0.6\pi}}{20}x^{-1.5} \quad \text{as } x \to \infty. \tag{2.24}$$

We close this section by mentioning a source for other long-tail distributions. Since any mixture of exponential distributions is infinitely divisible, see p. 452 of Feller [19], the PME distributions are all infinitely divisible. Since infinitely divisible distributions on $[0, \infty)$ have convenient Laplace transform representations, see p. 450 of Feller [19], it is natural to consider them as a source for additional long-tail distributions.

3. $M/G/1$ asymptotic expansions

Three-term asymptotic expansions for the waiting-time tail probabilities $P(W > x)$ in $M/G/1$ queues with long-tail service-time distributions are described in Willekens and Teugels [31] and references cited there. The asymptotics are developed in the time domain, exploiting the theory of regularly varying functions, as in Feller [19], and the theory of subexponential distributions as in appendix 4 of Bingham et al. [8]. Here we point out that the $M/G/1$ asymptotic expansion can be derived from the Laplace transforms, although so far it remains to fully justify our procedure. Our procedure is appealing because it is quick and insightful; it clearly shows where the terms come from.

Our starting point is the idea that a polynomial in the Laplace transform does not affect the asymptotics when the dominant singularity is at the origin. This idea is discussed on p. 254 of Doetsch [15], p. 333 of Bingham et al. [8] and p. 596 of Feller [19] in a special case. Notice that such polynomial terms appear in the PME transforms in (2.18), (2.19) and (2.21). Indeed, any distribution on the nonnegative real line with finite kth moment and infinite $(k+1)$st moment has a Laplace–Stieltjes transform that can be expressed as $\sum_{i=0}^{k} a_i s^i = \hat{h}(s)$ where $\hat{h}(s)$ has a singularity at 0. We present our key tool as an operation principle. It can be verified in special cases.

FIRST OPERATIONAL PRINCIPLE

Suppose that the Laplace transform of a function g is of the form

$$\hat{g}(s) = \sum_{i=0}^{\infty} a_i s^i + \hat{h}(s) \tag{3.1}$$

for scalars a_i, where $\hat{h}$ is the Laplace transform of h, which has a singularity at 0, and the series is absolutely convergent for s suitably small. Then $g(x) \sim h(x)$ as $x \to \infty$.

We now give an instance in which a variant of the first operational principle has been demonstrated. This result is due to Bingham and Doney [7] and appears on p. 333 of Bingham et al. [8]. Let $L(x)$ be a slowly varying function, e.g., a constant; see [8, 9].

THEOREM 5 (Bingham and Doney)

Consider a probability distribution with cdf G and Laplace–Stieltjes transform $\hat{g}(s)$ having finite kth moment but infinite $(k+1)$st moment, $k \geq 0$, so that

$\hat{g}(s) = \sum_{i=0}^{k} a_i s + \hat{h}(s)$. Then $\hat{h}(s) \sim (-1)^{k+1} s^r L(1/s)$ as $s \to 0$ if and only if

$$G^c(x) \sim \frac{(-1)^{k+1} L(x)}{r\Gamma(-r)x^r} \quad \text{as } x \to \infty, \tag{3.2}$$

provided that $k < r < k+1$.

The integer case related to theorem 5 is not so clean. The time-domain limit (3.2) implies an analogous transform limit, but *not* conversely. We deduce the equivalence below from Bingham and Doney [7].

THEOREM 6

In the setting of theorem 5, $h(s) \sim (-1)^{k+1} s^{k+1} \log s$ as $s \to \infty$ if and only if

$$\int_0^x y^{k+1}\, dG(y) \sim (k+1)! \log x \quad \text{as } x \to \infty. \tag{3.3}$$

The asymptotic relation (3.3) is in turn implied by (but does not imply)

$$1 - G(x) \sim \frac{k!}{x^{k+1}} \quad \text{as } x \to \infty. \tag{3.4}$$

Proof

The first equivalence follows from theorem A of Bingham and Doney [7] by letting their slowly varying function be $\log x$. Using integration by parts [19, p. 150],

$$\int_0^x y^{k+1}\, dG(y) = -x^{k+1} G^c(x) + (k+1) \int_0^x y^k G^c(y)\, dy,$$

we see that (3.4) implies (3.3). Example 3.1 below shows that the converse is not true. □

Note that the cases covered by theorems 5 and 6 are especially of interest here because the PME distributions have precisely this asymptotic form, as shown in theorem 3. For these PME distributions we can derive asymptotics for the Laplace transforms directly from (2.18) and (2.21). For the integer case in (2.18), we get

$$\hat{h}(s) \sim (-1)^{n+1} \left(\frac{n-1}{n}\right)^n n s^n \log s \quad \text{as } s \to 0;$$

for the half-integer case $r = n + 1/2$ in (2.21), we get

$$\hat{h}(s) \sim (-1)^{n+1}\left(\frac{r-1}{r}\right)^r r\pi s^r \quad \text{as } s \to 0.$$

For the special case in which G has a Pareto density, $g(x) = Ax^{-(k+2)}$ for $x \geq x_0$, we can deduce the transform asymptotics in theorem 6 from exponential integrals, in particular, from 5.1.12 on p. 229 of Abramowitz and Stegun [5]. We now give an example showing that (3.3) does not imply (3.4). This is similar to example 1 in Abate et al. [2] showing that an extra condition is needed in the Tauberian theorem applied there. Here too the transform asymptotics cannot detect periodic behavior in the time domain.

EXAMPLE 3.1

To see that the transform asymptotics in theorem 6 does not directly imply (3.4), let

$$g(t) = \frac{2(1-\cos t)}{\pi t^2} = \frac{4}{\pi}\left(\frac{\sin(t/2)}{t}\right)^2, \tag{3.5}$$

from which we deduce that

$$\hat{g}(s) = \frac{2}{\pi}\arctan(1/s) - \frac{s}{\pi}\log\left(1+\frac{1}{s^2}\right). \tag{3.6}$$

By 4.4.42 on p. 81 of [5], as $s \to 0$,

$$\arctan(1/s) = \frac{\pi}{2} - s + \frac{s^3}{3} - \frac{s^5}{5} + \ldots.$$

Therefore, the "singularity part" of $\hat{g}(s)$ is the second term on the right in (3.6). Since log has a branch cut singularity on the negative real axis, the critical singularity of

$$\frac{s}{\pi}\log\left(1+\frac{1}{s^2}\right) = \frac{s}{\pi}\log(1+s^2) - \frac{2s}{\pi}\log s$$

is at $s = 0$. Moreover, this term is asymptotic to $(2/\pi)s\log s$ as $s \to 0$. Hence, we see that the transform asymptotics does not reflect the periodic cosine function in (3.5). □

We now turn to the $M/G/1$ queue. Recall from (1.2) that the Laplace transform of the stationary-excess service-time distribution is $\hat{g}_e(s) = (1 - \hat{g}(s))/s$.

LEMMA 2

For the $M/G/1$ queue with service-time distribution having mean 1,

$$\hat{W}^c(s) \equiv \int_0^\infty e^{-sx} P(W > x)\, dx = \sum_{k=1}^{\infty} (-1)^{k+1} \left(\frac{\rho}{1-\rho}\right)^k (1 - \hat{g}_e(s))^k s^{-1}, \tag{3.7}$$

where the series is absolutely convergent for positive s sufficiently small.

Proof

Recall that the Laplace–Stieltjes transform of W is

$$\begin{aligned} E e^{-sW} &= \frac{1-\rho}{1-\rho\left(\dfrac{1-\hat{g}(s)}{s}\right)} = \frac{1}{1+\left(\dfrac{\rho}{1-\rho}\right)\left[\dfrac{\hat{g}(s)-(1-s)}{s}\right]} \\ &= \frac{1}{1+\dfrac{\rho}{1-\rho}(1-\hat{g}_e(s))}. \end{aligned} \tag{3.8}$$

Since $\hat{W}^c(s) = [1 - E e^{-sW}]/s$, (3.7) follows from (3.8). Since $\hat{g}_e(s)$ is the Laplace transform of a probability density, $\hat{g}_e(0) = 1$ and, for any $\epsilon > 0$, $|1 - g_e(s)| < \epsilon$ for positive s suitably small. Hence, the series is indeed absolutely convergent for positive s suitably small. □

The idea now is to invert the series in (3.7) term by term. Borovkov justifies such a term-by-term approach for long-tail service time distributions in lemma 2 on p. 133 of [9]. When we do this, we get

$$\begin{aligned} W^c(t) = \frac{\rho}{1-\rho} G_e^c(t) &+ \left(\frac{\rho}{1-\rho}\right)^2 [1 - G_e^{*2}(t) - 2G_e^c(t)] \\ &+ \left(\frac{\rho}{1-\rho}\right)^3 \left(3[G_e^c(t) - (1 - G_e^{*2}(t))] + 1 - G_e^{*3}(t)\right) + \ldots . \end{aligned} \tag{3.9}$$

Note that the first term is just (1.4), as it should be. To make a connection to Willekens and Teugels [31], their (10) and (11) shows how to do asymptotics

in the time domain to obtain theorem 5 from the terms in (3.9). Instead, we reason directly from (3.7), exploiting the first operational principle above.

Let $\mathscr{L}^{-1}(\hat{g})$ denote the inverse Laplace transform of $\hat{g}$. Recall that $\mathscr{L}^{-1}(s^k\hat{g})$ is the kth derivative of g. The assumptions in the next operational principle express key properties of long-tail distributions (our case 3), and are satisfied by the PME distributions in section 2. Note, that the result agrees with the one-term asymptotics in theorem 1 and the three-term asymptotics in Willekens and Teugels [31]. With our approach, we can also obtain additional terms.

SECOND OPERATIONAL PRINCIPLE

Suppose that the service-time density g in an $M/G/1$ queue has Laplace transform of the form (3.1), where $\mathscr{L}^{-1}(s^{k+1}\hat{g}(-s))$ is asymptotically negligible compared to $\mathscr{L}^{-1}(s^k\hat{g})$ as $x \to \infty$ for $k \geq -2$ and $\mathscr{L}^{-1}(s^{-3}\hat{h}(s)^2)$ is asymptotically negligible compared to $\mathscr{L}^{-1}(\hat{h}(s))$ as $x \to \infty$. Assuming that the asymptotics of the sum in (3.7) is the sum of the asymptotics for each term, we obtain

$$P(W > x) \sim \left(\frac{\rho}{1-\rho}\right)\psi_1(x) + m_2\left(\frac{\rho}{1-\rho}\right)^2\psi_2(x) + \left[\frac{m_3}{3} + \frac{3m_2^2}{4}\left(\frac{\rho}{1-\rho}\right)\right]\left(\frac{\rho}{1-\rho}\right)^2\psi_3(x) \quad \text{as } x \to \infty, \tag{3.10}$$

with the two-term (three-term) expansion being valid when $m_2 < \infty$ ($m_3 < \infty$), where

$$\begin{aligned} &\psi_3(x) \sim g(x) \sim h(x) && \text{as } x \to \infty, \\ &\psi_2(x) \sim G^c(x) \sim H^c(x) \equiv \int_x^\infty h(y)\,dy && \text{as } x \to \infty, \\ &\psi_1(x) \sim G_e^c(x) = \int_x^\infty G^c(y)\,dy \sim \int_x^\infty H^c(y)\,dy && \text{as } x \to \infty. \end{aligned} \tag{3.11}$$

Partial proof

From (3.1), we see that $\mathscr{L}^{-1}(s^{k+1}\hat{h})$ is asymptotically negligible compared to $\mathscr{L}^{-1}(s^k\hat{h})$ as $x \to \infty$, given the corresponding property assumed for $\hat{g}$. Since $\mathscr{L}^{-1}(s^{-3}\hat{h}(s)^2)$ is asymptotically negligible compared to $\mathscr{L}^{-1}(\hat{h}(s))$ as $x \to \infty$, we

We now turn to the $M/G/1$ queue. Recall from (1.2) that the Laplace transform of the stationary-excess service-time distribution is $\hat{g}_e(s) = (1 - \hat{g}(s))/s$.

LEMMA 2

For the $M/G/1$ queue with service-time distribution having mean 1,

$$\hat{W}^c(s) \equiv \int_0^\infty e^{-sx} P(W > x)\, dx = \sum_{k=1}^{\infty} (-1)^{k+1} \left(\frac{\rho}{1-\rho}\right)^k (1 - \hat{g}_e(s))^k s^{-1}, \tag{3.7}$$

where the series is absolutely convergent for positive s sufficiently small.

Proof

Recall that the Laplace–Stieltjes transform of W is

$$\begin{aligned} E e^{-sW} &= \frac{1-\rho}{1-\rho\left(\dfrac{1-\hat{g}(s)}{s}\right)} = \frac{1}{1+\left(\dfrac{\rho}{1-\rho}\right)\left[\dfrac{\hat{g}(s)-(1-s)}{s}\right]} \\ &= \frac{1}{1+\dfrac{\rho}{1-\rho}(1-\hat{g}_e(s))}. \end{aligned} \tag{3.8}$$

Since $\hat{W}^c(s) = [1 - E e^{-sW}]/s$, (3.7) follows from (3.8). Since $\hat{g}_e(s)$ is the Laplace transform of a probability density, $\hat{g}_e(0) = 1$ and, for any $\epsilon > 0$, $|1 - g_e(s)| < \epsilon$ for positive s suitably small. Hence, the series is indeed absolutely convergent for positive s suitably small. □

The idea now is to invert the series in (3.7) term by term. Borovkov justifies such a term-by-term approach for long-tail service time distributions in lemma 2 on p. 133 of [9]. When we do this, we get

$$\begin{aligned} W^c(t) = \frac{\rho}{1-\rho} G_e^c(t) &+ \left(\frac{\rho}{1-\rho}\right)^2 [1 - G_e^{*2}(t) - 2G_e^c(t)] \\ &+ \left(\frac{\rho}{1-\rho}\right)^3 \left(3[G_e^c(t) - (1 - G_e^{*2}(t))] + 1 - G_e^{*3}(t)\right) + \ldots. \end{aligned} \tag{3.9}$$

Note that the first term is just (1.4), as it should be. To make a connection to Willekens and Teugels [31], their (10) and (11) shows how to do asymptotics

can disregard powers $\hat{h}(s)^k$ in expansions of $[\hat{g}(s) - (1-s)]$. In particular,

$$\begin{aligned}
\mathscr{L}^{-1}\left(s^{k-1}\left(\frac{\hat{g}(s)-(1-s)}{s^2}\right)^k\right) &= \mathscr{L}^{-1}\left(s^{k-1}\left(a_2 + a_3 s + \ldots + \frac{\hat{h}(s)}{s^2}\right)^k\right) \\
&\sim \mathscr{L}^{-1}\left(\frac{\hat{h}(s)}{s^2}\right) \quad \text{as } x \to \infty \qquad \text{for } k = 1, \\
&\sim \mathscr{L}^{-1}\left(2a_2\frac{\hat{h}(s)}{s} + a_3\hat{h}(s) + \frac{\hat{h}(s)^2}{s^3}\right) \qquad \text{for } k = 2, \\
&\sim \mathscr{L}^{-1}(3a_2^2\hat{h}(s)) \qquad \text{for } k = 3, \\
&\sim \mathscr{L}^{-1}(ka_2^k s^{k-2}\hat{h}(s)) \qquad \text{for } k > 3. \qquad (3.12)
\end{aligned}$$

We thus obtain (3.10) with (3.11) by applying (3.12) and (3.7), assuming the asymptotics can be done term by term (see p. 133 of [9]). □

To apply the second operational principle, we do not need to work directly with the transforms, because the result gives the asymptotics in terms of the service-time cdf G. However, we could also work with the transforms; e.g., (2.18), (2.19), (2.21) and (2.22). With (2.18) and (2.19), we need to work with the functions $s^n \log(1 + as^{-1})$. For relevant material on Laplace transforms involving logarithms, see pp. 67–69 of Abramowitz and Stegun [5], no. 43 on p. 320 of Doetsch [15], pp. 340–341 of Carslaw and Jaeger [10] and p. 104 of Smith [28].

We can also do asymptotics when $m_2 = \infty$ or $m_3 = \infty$. To illustrate, we consider the PME distribution with $r = 2$; then $m_2 = \infty$. Since we do not have complete theoretical justification, we call our result a conjecture. (Since the PME distribution is a mixture of exponentials, the PME distribution is completely monotone. This implies that the waiting-time distribution is completely monotone. Hence, an anomaly such as in example 3.1 seems very unlikely.) In [3] we saw that the two-term asymptotics when $r = 2$ performs pretty well in this case.

CONJECTURE

Consider the $M/G/1$ queue with PME service-time distribution having $r = 2$. Again, assuming that the asymptotics can be done term by term in (3.7), we obtain

$$P(W > x) \sim \frac{\rho}{2(1-\rho)x}\left(1 + \frac{\rho\log(2x) - 1)}{(1-\rho)x}\right) \quad \text{as } x \to \infty.$$

Partial proof

Reasoning as with the second operational principle, we get

$$\hat{W}^c(s) = \left(\frac{\rho}{1-\rho}\right)\left(\frac{\hat{g}_2(s)-(1-s)}{s^2}\right) - \left(\frac{\rho}{1-\rho}\right)^2\left(\frac{\hat{g}_2(s)-(1-s)}{s^2}\right)^2 s + \dots$$

$$= \frac{\rho}{2(1-\rho)}\log(1+2s^{-1}) - \left(\frac{\rho}{2(1-\rho)}\right)^2 (\log(1+2s^{-1}))^2 s + \dots .$$

Since $\log(1+2s^{-1}) \sim -\log s$, $(\log(1+2s^{-1}))^2 \sim -2(\log 2)\log s + (\log s)^2$, $\mathscr{L}^{-1}(-\log s) \sim x^{-1}$, $\mathscr{L}^{-1}(s\log s) \sim x^{-2}$ and $\mathscr{L}^{-1}(s(\log s)^2) \sim 2(\log x - 1)/x^2$ as $x \to \infty$, the result follows. □

4. Numerical examples

In this section we examine some numerical examples.

One-term asymptotics for GI/G/1

We start by investigating the quality of the simple one-term asymptotic approximation provided by the corollary to theorem 1. Table 1 covers the PME service-time distribution in section 2 with mean 1 and $r = 3.0$. Now we consider the PME distribution with mean 1 and $r = 4$. The first four moments of G_4 are $m_1 = 1$, $m_2 = 2.25$, $m_3 = 10.125$ and $m_4 = \infty$.

To focus on the influence of the interarrival-time distribution beyond its mean, we consider four different interarrival-time distributions, all with mean 1.25, so that $\rho = 0.8$. The distributions we consider are Erlang of order 2 (E_2), exponential (M), gamma with shape parameter 1/2 ($\Gamma_{1/2}$), and the long-tail distribution in section 2 with $r = 4$ (G_4). (The random variable is multiplied by 1.25 to make it have mean 1.25.) The SCVs of these distributions are 0.5, 1.0, 2.0 and 1.25, respectively. By the criterion of SCV, the long-tail G_4 distribution is actually less variable than the $\Gamma_{1/2}$ distribution.

Note that the $\Gamma_{1/2}$ and G_4 distributions do not have rational Laplace transforms. In these cases, neither the interarrival-time transform nor the service-time transform is rational, but that presents no difficulty for the numerical transform inversion algorithm based on Pollaczek's contour integrals in [3]. Since the service-time distribution is long-tailed, we used the procedure in §4 of [3] with damping parameter $\alpha = 10^{-8}$.

The steady-state waiting-time tail probabilities $P(W > x)$ are plotted for values of x ranging from 5 to 50 in table 2. Recall that the mean service time is 1

Table 2
A comparison of exact values with the asymptotic approximation (1.5) for the tail probabilities $P(W > x)$ in $GI/G/1$ queues with a long-tail service-time distribution having $r = 4$. Four different interarrival-time distributions are considered, all with $\rho = 0.8$.

	Interarrival-time distribution				
x	long-tail G_4 $c_a^2 = 1.25$	$\Gamma_{1/2}$ $c_a^2 = 2.0$	M $c_a^2 = 1.0$	E_2 $c_a^2 = 0.5$	Common asymptotics (1.5)
5	0.3462	0.4498	0.3200	0.2364	0.0810
10	0.1550	0.2417	0.1359	0.0833	0.0101
20	0.0328	0.0714	0.0262	0.0120	0.0013
30	0.0074	0.0215	0.0055	0.0021	0.00038
40	0.00180	0.00661	0.00128	0.00051	0.00016
50	0.00050	0.00209	0.00036	0.00018	0.00008

and the mean waiting time in the $M/G/1$ case is

$$EW = \frac{\rho(1 + c_s^2)}{2(1 - \rho)} = 4.5. \tag{4.1}$$

Thus the range is reasonable.

From table 2, we see that the exact tail probabilities $P(W > x)$ range from 1.8×10^{-4} to 2.1×10^{-3} at $x = 50$, while the asymptotic approximation is 8×10^{-5}. Thus, for these values of x, the influence of the interarrival-time distribution beyond its mean is still significant and the simple asymptotic approximation provided by (1.5) is not yet very good. Indeed, in the case of $\Gamma_{1/2}$ interarrival times, the exact value at $x = 50$ is about 2.1×10^{-3}, while the asymptotic approximation is too small by a factor of 25. For more variable gamma interarrival-time distributions (with shape parameter less than 1.2), the error is even greater.

Moreover, the difficulty with (1.5) is not alleviated by using (1.4) or (1.3). In particular, the asymptotic approximation (1.5) for G_{4e}^c in (1.4) is pretty good. Since $G_r^c(x)$ is decreasing in x, we see that the approximation (2.3) is always less than or equal to (1.4), but (1.4) and (1.5) are too low in table 2.

Also note, consistent with the interarrival-time SCVs, that the tail probabilities for the $GI/G/1$ queue when the interarrival-time distribution also has a long tail falling between the M and $\Gamma_{1/2}$ cases. Generally speaking, the long-tail property in the interarrival-times has much less impact upon waiting-time tail probabilities than the long-tail property in service times. As noted in [3], long-tail interarrival-time distributions present no difficulty for the algorithm there.

Two-term and three-term asymptotics for M/G/1

We also investigate the accuracy of the asymptotic approximations with more terms in the $M/G/1$ case discussed in section 3. We consider the same G_4 service-time distribution and we once again let $\rho = 0.8$. In this case we are able to apply the numerical inversion algorithms in [4] directly, which requires less computation.

The exact results and the asymptotic results are displayed in table 3. The correctness of the numerics and the asymptotics is evident from the larger values of x in table 3. The one-term, two-term and three-term approximations have 23%, 1.5% and 1.0% relative errors at $x = 250$. In particular, the two-term and three-term approximations perform well at $x = 250$, but then the exact probability is only about 7×10^{-7}.

It is also significant that the asymptotic approximations take a long time (large x) before they are good. For example, even the three-term approximation is off by a factor of 4 at $x = 40$.

Since the $M/G/1$ waiting-time tail probability has the exact representation

$$P(W > x) = 1 - (1 - \rho) \sum_{n=0}^{\infty} \rho^n G_{re}^{*nc}(x), \tag{4.2}$$

where G_{re}^{*nc} is the complementary cdf associated with the n-fold convolution of the stationary excess cdf G_{re} associated with G_r, we anticipate that the first term might provide a good approximation in light traffic, i.e.,

$$P(W > x) \approx 1 - (1 - \rho)(1 + \rho G_{re}(x)) \approx \rho G_{re}^{c}(x) \tag{4.3}$$

Table 3
A comparison of asymptotic approximations with exact values of the tail probabilities in the $M/G/1$ queue with long-tail service times having $r = 4$ and $\rho = 0.8$.

		Asymptotics		
x	Exact	one-term $\phi(x) \equiv \frac{10.125}{x^3}$	two-term $\phi(x)\left(1 + \frac{27}{x}\right)$	three-term $\phi(x)\left(1 + \frac{27}{x} + \frac{352}{x^2}\right)$
10	0.1359	0.0101	0.0374	0.0730
20	0.0262	0.0013	0.0031	0.0042
30	0.0055	0.00038	0.00072	0.00087
40	0.00128	0.00016	0.00027	0.00031
50	3.60×10^{-4}	8.10×10^{-5}	1.25×10^{-4}	1.39×10^{-4}
62.5	1.04×10^{-4}	4.15×10^{-5}	5.94×10^{-5}	6.31×10^{-5}
125	6.75×10^{-6}	5.18×10^{-6}	6.30×10^{-6}	6.42×10^{-6}
250	7.29×10^{-7}	6.48×10^{-7}	7.17×10^{-7}	7.22×10^{-7}

Table 4
A comparison of asymptotic approximations with exact values of the tail probabilities in the $M/G/1$ queue with long-tail service times having $r = 4$ and $\rho = 0.2$ (light traffic).

		Asymptotics		
x	Exact	one-term $\phi(x) \equiv \frac{0.63281}{x^3}$	two-term $\phi(x)\left(1 + \frac{1.35}{x}\right)$	three-term $\phi(x)\left(1 + \frac{1.35}{x} + \frac{11.39}{x^2}\right)$
8	1.7778×10^{-3}	1.236×10^{-3}	1.445×10^{-3}	1.665×10^{-3}
16	1.8336×10^{-4}	1.545×10^{-4}	1.675×10^{-4}	1.744×10^{-4}
24	5.0499×10^{-5}	4.578×10^{-5}	4.836×10^{-5}	4.927×10^{-5}
32	2.0663×10^{-5}	1.931×10^{-5}	2.012×10^{-5}	2.034×10^{-5}
40	1.0409×10^{-5}	9.888×10^{-6}	1.022×10^{-5}	1.029×10^{-5}

which is approximately equal to (1.4) in light traffic. Hence, we show numerical values for the $M/G/1$ queue with $\rho = 0.2$ in table 4. From table 4, we see that the asymptotic approximations do indeed perform significantly better in this case.

We conclude by making two observations about the differences between the asymptotics in (1.1) and (1.5). With (1.1), once the asymptotics provides a good approximation, it tends to stay good, whereas with (1.5) in table 4 it is better at $x = 8$ than at $x = 40$. Another is that (1.1) is better in heavy traffic, while (1.5) is better in light traffic.

Finite approximations

We now consider approximations based on finite approximations of the service-time distribution. We assume that the original service-time distribution is the PME distribution with $r = 3$, so that we have exact results to compare with, namely, the results in table 1.

We consider two approximations: first, (1.6) in which the approximating service-time distribution is stochastically less than the true service-time distribution and, second, (1.7) in which the point mass is put in the midpoint of the interval.

The upper limit of the approximating distribution is fixed at 500. The difference between successive boundary points is made to grow geometrically so that the ratio of the last difference to the first equals 200. The results are displayed in table 5.

Consistent with theory, the approximation based on (1.6) yields a lower bound, while the approximation based on (1.7) is neither a lower bound nor an upper bound. As one might anticipate, the approximation based on (1.7) is considerably more accurate than the one based on (1.6).

For the lower x values, the upper limit of 500 on the approximating distribution looks reasonable. However, for the higher x values, this seems to be not entirely true, because as we increase the number N of points, the approximations approach a value slightly lower than the true value.

Table 5
A comparison of approximate tail probabilities $P(W > x)$ based on finite service-time distribution approximations (1.6) and (1.7) using N points with exact values for the $M/G/1$ queue with PME distribution having $r = 3.0$ and $\rho = 0.8$ for several values of N. ($e - k$ stands for $\times 10^{-k}$).

		Approximations					
		(1.6)			(1.7)		
x	Exact	$N = 100$	$N = 300$	$N = 900$	$N = 100$	$N = 300$	$N = 900$
10	1.675e − 1	9.226e − 2	1.359e − 1	1.560e − 1	1.711e − 1	1.679e − 1	1.675e − 1
20	4.673e − 2	1.991e − 2	3.417e − 2	4.195e − 2	4.828e − 2	4.689e − 2	4.673e − 2
30	1.573e − 2	6.284e − 3	1.103e − 2	1.388e − 2	1.633e − 2	1.578e − 2	1.572e − 2
40	6.407e − 3	2.726e − 3	4.534e − 3	5.655e − 3	6.645e − 3	6.421e − 3	6.396e − 3
50	3.143e − 3	1.481e − 3	2.303e − 3	2.800e − 3	3.241e − 3	3.141e − 3	3.130e − 3
60	1.804e − 3	9.292e − 4	1.370e − 3	1.624e − 3	1.845e − 3	1.796e − 3	1.790e − 3
70	1.165e − 3	6.381e − 4	9.080e − 4	1.056e − 3	1.182e − 3	1.154e − 3	1.151e − 3
80	8.171e − 4	4.655e − 4	6.473e − 4	7.433e − 4	8.277e − 4	8.051e − 4	8.031e − 4
90	6.075e − 4	3.543e − 4	4.852e − 4	5.523e − 4	6.068e − 4	5.948e − 4	5.935e − 4
100	4.708e − 4	2.783e − 4	3.771e − 4	4.267e − 4	4.664e − 4	4.577e − 4	4.567e − 4

Overall, table 5 suggests that this finite approximation scheme is reasonable. Certainly the approximations provided by (1.6) and (1.7) are far superior to the approximation provided by the gamma distribution matching the first two moments, as can be seen by comparing tables 1 and 5.

5. The second case

In this section we point out that the second case of waiting-time asymptotics is not as pathological as it might seem. Indeed, it is closely related to the third case that we have considered.

For the discussion here, it is useful to *classify the service-time distributions* (all assumed to have mean 1). We assume that all interarrival-time distributions have finite mean and that $V - U$ has a nonlattice distribution. We say that a service-time distribution belongs to class I if (1.1) holds for all interarrival time distributions and all ρ, $0 < \rho < 1$. We say that a service-time distribution belongs to class II if, for each interarrival-time distribution, there exists ρ^* with $0 < \rho^* < 1$ such that (1.1) holds for all ρ with $\rho^* < \rho < 1$ but does not hold for any ρ with $0 < \rho < \rho^*$. We say that a service-time distribution belongs to class III if (1.1) never holds.

Since (1.1) is characterized by the equation $E\,e^{s(V-U)} = 1$ having a solution, e.g., see Abate et al. [1], Asmussen [6] or Borovkov [9], it is easy to classify the service-time distributions directly in terms of their Laplace–Stieltjes transforms $\hat{g}(s) \equiv E\,e^{-sV}$. If $\hat{g}(s)$ has no singularities, then it is class I. Henceforth, assume

that it does have singularities (necessary in the left half plane). Let $-s^*$ for $s^* > 0$ be the rightmost singularity of $\hat{g}(s)$ (s^* is the radius of convergence of $E\,e^{sV}$). Then g is classified as follows:

$$\begin{aligned} &\text{class I:} && s^* > 0 \text{ and } g(-s^*) = \infty; \\ &\text{class II:} && s^* > 0 \text{ and } 1 < g(-s^*) < \infty; \\ &\text{class III:} && s^* = 0. \end{aligned} \tag{5.1}$$

From (5.1) it is easy to see how to convert a class-III service-time distribution into a class-II service-time distribution and vice versa. For any $u > 0$ and any class-III service-time density $g(x)$, construct a new density by *exponential damping*, i.e., by setting

$$h_u(x) = \frac{e^{-ux} g(x)}{\hat{g}(u)}, \quad x \geq 0. \tag{5.2}$$

Clearly, $h_u(x)$ is a bonafide probability density function. Since $\hat{h}_u(s) = \hat{g}(s+u)/\hat{g}(u)$, $\hat{g}(0) = 1$ and $\hat{g}(-s) = \infty$ for all $s > 0$, $-u$ is the rightmost singularity of $\hat{h}_u(s)$ and $\hat{h}_u(-u) = 1/\hat{g}(u)$. Hence, h_u is indeed a class-II density. If u is very small, then h_u is very close to g.

Moreover, given any class-II density h, where $h(s)$ has rightmost singularity at $-s^*$, we can construct a class-III density by setting

$$\hat{g}(s) = \frac{\hat{h}(s - s^*)}{h(-s^*)}, \tag{5.3}$$

i.e., by setting

$$g(x) = \frac{e^{s^* x} h(x)}{h(-s^*)}, \quad x \geq 0. \tag{5.4}$$

In other words, we obtain classes II and III from each other simply by translation (to move the critical singularity) and renormalization with the transforms.

The asymptotics for a $GI/G/1$ queue with a class-II service-time distribution is given in Pakes [25, theorem 2] and Borovkov [9]. The version developed by Pakes seems very nice, because the asymptotic constant is directly expressed in terms of available transforms. The following is a restatement.

THEOREM 7 (Pakes)

Consider the $GI/G/1$ queue with $EV = 1$, $EU = \rho^{-1}$ and Laplace transforms $\hat{a}(s) = E\,e^{-sU}$ and $\hat{g}(s) = E\,e^{-sV}$, where the rightmost singularity of $\hat{g}(s)$ is $-s^*$,

$0 < s^* < \infty$, and $\hat{g}(-s^*) = d$, $1 < d < \infty$ (class II). For $\rho < \rho^*$, where $\hat{a}(\rho^*) = 1/d$,

$$P(W > x) \sim \frac{s^* \hat{w}(-s^*)\hat{a}(s^*)}{1 - d\hat{a}(s^*)} G_e^c(x) \quad \text{as } x \to \infty. \tag{5.5}$$

We can calculate the asymptotic constant in (5.5) by calculating the waiting-time transform value $\hat{w}(-s^*)$. To do this, we can use the Pollaczek [26, 27] integral representation, using exponential damping just as for the case-III service-time distributions in [3]. Note that, in contrast to (1.4), the asymptotic constant in (5.5) depends on the interarrival-time distribution beyond its mean via $\hat{a}(s^*)$ and $\hat{w}(-s^*)$.

COROLLARY

If, in addition to the assumptions of theorem 7, the interarrival-time distribution is exponential, then

$$P(W > x) \sim \frac{\rho(1-\rho)}{(1 - [\rho/\rho^*])^2} G_e^c(x) \quad \text{as } x \to \infty. \tag{5.6}$$

Proof

Since $\hat{a}(s) = \rho/(\rho + s)$, $\rho^* = s^*/(d-1)$ and

$$\frac{s^* \hat{w}(-s^*)\hat{a}(s^*)}{1 - d\hat{a}(s^*)} = \frac{\rho s^*}{\rho + s^* - d\rho} \frac{1-\rho}{1 - \rho\hat{g}_e(-s^*)} = \frac{\rho(1-\rho)}{(1 - [\rho/\rho^*])^2}. \qquad \square$$

From the close connection between classes II and III shown in (5.2)–(5.4), we anticipate that the asymptotic formula for class II also does not yield very good approximations, and this is our experience. The poor performance of (5.6) is illustrated by example 5 of [1].

Note added in proof: Additional discussion of class-II service-time distributions and their role in the $M/G/1$ queue (in a risk theory context) appears in: P. Embrechts, A property of the generalized inverse Gaussian distribution with some applications, J. Appl. Prob. 20 (1983) 537–544.

6. Conclusions

In this paper we have focused on the $GI/G/1$ queue with long-tail service-time distributions. We have seen that long-tail service-time distributions not only make the steady-state waiting-time tail probabilities bigger (see table 1), but they significantly affect the ways we can analyze the model. In particular, we have focused on two issues: asymptotic approximations and numerical algorithms.

In section 3 we showed that $M/G/1$ asymptotics is remarkably easy to derive with Laplace transforms, although the procedure is not fully justified. We have also seen that the asymptotics provides remarkably simple formulas. Unfortunately, however, the approximations for the tail probabilities $P(W > x)$ provided by the asymptotics in theorem 1 and section 3 for the case of long-tail service times can be quite poor for typical x of interest. Indeed, the quality of the approximations in the case of long-tail service-time distributions is as remarkably bad as the quality of the approximations in the case of (1.1) is remarkably good. (This can be explained, at least in part, by the rates of convergence.) Moreover, the asymptotic approximation with long-tail service-time distribution often underestimates the true tail probability. Thus, we suggest caution when applying asymptotics for the waiting-time tail probabilities in the case of long-tail service-time distributions.

Fortunately, the algorithms in Abate et al. [3], Abate and Whitt [4] and Choudhury et al. [12] are effective for computation when the Laplace transform of the service-time distribution is known. Since Laplace transforms do not seem to be known for the classical long-tail distributions, we introduced a new class of long-tail distributions with explicit Laplace transforms in section 2. This class seems promising for applications with data.

We have also suggested approximating other distributions by finite approximations in order to obtain an approximation of the required Laplace transform. With data, we suggest using empirical Laplace transforms. With data, the significant remaining difficulty is the validity of the independence and stationarity assumptions underlying the $GI/G/1$ model.

In this paper we have focused on long-tail service-time distributions rather than long-tail interarrival-time distributions. We have found that long-tail interarrival-time distributions have much less impact on waiting-time tail probabilities. Indeed, large waiting times are primarily due to long service times and short interarrival times. However, care should be taken in the analysis whenever long-tail distributions might be present.

We have analyzed the steady-state performance of the $GI/G/1$ queue with the FIFO discipline. The long-tail property makes steady-state be approached more slowly. Hence, it is even more important than usual to consider the transient behavior. The long-tail property also suggests that we should consider other service disciplines besides FIFO such as processor sharing.

In section 5 we considered the second case, intermediate between the long-tail

service-time distributions and the exponential asymptotics in (1.1). We showed that the second case is closely related to the third. We reviewed the asymptotics for the second case and noted that the quality of the associated approximations is also typically not good.

Acknowledgement

We thank Donald P. Gaver for helpful comments.

References

[1] J. Abate, G.L. Choudhury and W. Whitt, Exponential approximations for tail probabilities in queues, I: waiting times, Oper. Res., to appear.
[2] J. Abate, G.L. Choudhury and W. Whitt, Asymptotics for steady-state tail probabilities in structured Markov queueing models, Stoch. Models 10 (1994) 99–143.
[3] J. Abate, G.L. Choudhury and W. Whitt, Calculation of the $GI/G/1$ waiting time distribution and its cumulants from Pollaczek's formulas, Archiv für Elektronik und Übertragungstechnik (1993) 311–321.
[4] J. Abate and W. Whitt, The Fourier-series method for inverting transforms of probability distributions, Queueing Syst. 10 (1992) 5–88.
[5] M. Abramowitz and I.A. Stegun, *Handbook of Mathematical Functions* (National Bureau of Standards, Washington, DC, 1972).
[6] S. Asmussen, *Applied Probability and Queues* (Wiley, 1987).
[7] N.H. Bingham and R.A. Doney, Asymptotic properties of supercritical branching processes I: the Galton–Watson process, Adv. Appl. Prob. 6 (1974) 711–731.
[8] N.H. Bingham, C.M. Goldie and J.L. Teugels, *Regular Variation* (Cambridge University Press, Cambridge, England, 1989).
[9] A.A. Borovkov, *Stochastic Processes in Queueing Theory* (Springer, 1976, translation of 1972 Russian edition).
[10] H.S. Carslaw and J.C. Jaeger, *Conduction of Heat in Solids*, 2nd ed. (Clarendon Press, 1959).
[11] G.L. Choudhury, D.M. Lucantoni and W. Whitt, Squeezing the most out of ATM (1993), submitted.
[12] G.L. Choudhury, D.M. Lucantoni and W. Whitt, An algorithm for a large class of $G/G/1$ queues, in preparation.
[13] G.L. Choudhury and W. Whitt, Heavy-traffic asymptotic expansions for the asymptotic decay rates in the $BMAP/GI/1$ queue, Stoch. Models 10, no. 2 (1994), to appear.
[14] J.W. Cohen, Some results on regular variation for distributions in queueing and fluctuation theory, J. Appl. Prob. 10 (1973) 343–353.
[15] G. Doetsch, *Introduction to the Theory and Application of Laplace Transformation* (Springer, 1974).
[16] D.E. Duffy, A.A. McIntosh, M. Rosenstein and W. Willinger, Analyzing telecommunications traffic data from working common channel signaling subnetworks, *Proc. INTERFACE '93*, to appear.
[17] D.E. Duffy, A.A. McIntosh, M. Rosenstein and W. Willinger, Analyzing CCSN/SST traffic data from working CCS subnetworks: implications for engineering and modeling, Bellcore, Morristown, NJ (1993).
[18] A. Erdélyi, *Asymptotic Expansions* (Dover, 1956).

[19] W. Feller, *An Introduction to Probability Theory and its Applications*, Vol. 2, 2nd ed. (Wiley, 1971).
[20] D.P. Gaver and P.A. Jacobs, Nonparametric estimation of the probability of a long delay in the $M/G/1$ queue, J. R. Statist. Soc. B 50 (1988) 392–402.
[21] I.S. Gradshteyn and I.M. Ryzhik, *Table of Integrals, Series and Products* (Academic Press, 1980).
[22] C.M. Harris, The Pareto distribution as a queue service distribution, Oper. Res. 16 (1968) 307–313.
[23] N.L. Johnson and S. Kotz, *Distributions in Statistics: Continuous Univariate Distributions-1* (Wiley, 1970).
[24] K.S. Meier-Hellstern, P.E. Wirth, Y.L. Yan and D.A. Hoeflin, Traffic models for ISDN data users: office automation application, *Teletraffic and Data Traffic in a Period of Change, ITC 13*, eds. A. Jensen and B. Iversen (Elsevier, Amsterdam, 1991) pp. 167–172.
[25] A.G. Pakes, On the tails of waiting-time distributions, J. Appl. Prob. 12 (1975) 555–564.
[26] F. Pollaczek, Fonctions caractéristiques de certaines répartitions définies au moyen de la notion d'ordre. Application à la théorie des attentes, C. R. Acad. Sci. Paris 234 (1952) 2334–2336.
[27] F. Pollaczek, *Problèmes Stochastiques Poseés par le Phénomène de Formation d'une Queue d'Attente à un Guichet et par des Phénomènes Apparentes*, Mémorial des Sciences Mathématiques, fac. 136 (Gauthier–Villars, Paris, 1957).
[28] M.G. Smith, *Laplace Transform Theory* (Van Nostrand, 1966).
[29] W. Whitt, On approximations for queues I: extremal distributions, AT&T Bell Lab. Tech. J 63 (1984) 115–138.
[30] W. Whitt, Planning queueing simulations, Manag. Sci. 35 (1989) 1341–1366.
[31] E. Willekens and J.L. Teugels, Asymptotic expansions for waiting time probabilities in an $M/G/1$ queue with long-tailed service time, Queueing Syst. 10 (1992) 295–312.