CALCULATION OF THE GI/G/1 WAITING-TIME DISTRIBUTION AND ITS CUMULANTS FROM POLLACZEK'S FORMULAS

by

Joseph Abate,[1] *Gagan L. Choudhury*[2] *and Ward Whitt*[3]

AT&T Bell Laboratories

A contribution to the 1993 special issue of *AEÜ, Archiv für Elektronik und Übertragungstechnik*
(International Journal of Electronics and Communication)

on Teletraffic Theory and Engineering

in memory of Félix Pollaczek (1892-1981)

December 1, 1992

Revision: February 9, 1993

[1] AT&T Bell Laboratories Retired, 900 Hammond Road, Ridgewood, NJ 07450-2908

[2] AT&T Bell Laboratories, Room 3K-603, Holmdel, NJ 07733-3030

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

Abstract

The steady-state waiting time in a stable GI/G/1 queue is equivalent to the maximum of a general random walk with negative drift. Thus, the distribution of the steady-state waiting time in the GI/G/1 queue is characterized by Spitzer's (1956) formula. However, earlier, Pollaczek (1952) derived an equivalent contour-integral expression for the Laplace transform of the GI/G/1 steady-state waiting time. Since Spitzer's formula is easier to understand probabilistically, it is better known today, but it is not so easy to apply directly except in special cases. In contrast, we show that it is easy to compute the GI/G/1 waiting-time distribution and its cumulants (and thus its moments) from Pollaczek's formulas. For the waiting-time tail probabilities, we use numerical transform inversion, numerically integrating the Pollaczek contour integral to obtain the transform values. For the cumulants and the probability of having to wait, we directly integrate the Pollazcek contour integrals numerically. The resulting algorithm is evidently the first for a GI/G/1 queue in which neither the transform of the interarrival-time distribution nor the transform of the service-time transform distribution need be rational. The algorithm can even be applied to long-tail distributions, i.e., distributions with some infinite moments. To treat these distributions, we approximate them by suitable exponentially-damped versions of these distributions. Overall, the algorithm is remarkably simple compared to alternative algorithms requiring more structure.

key words: Félix Pollaczek, queueing theory, computational probability, GI/G/1 queue, waiting time distribution, tail probabilities, random walk, Spitzer's formula, numerical transform inversion, numerical integration, contour integrals.

1. Introduction and Summary

In 1987 the Operations Research Society of America celebrated its 35th anniversary by making special commemorative T-shirts. One of these, now frequently worn by six-year old Daniel Whitt as a night shirt, displays the famous Pollaczek-Khintchine formula for the mean steady-state waiting time in the M/G/1 queue:

$$EW = \frac{\rho\tau}{2(1-\rho)} \left[1 + \frac{\sigma^2}{\tau^2} \right] . \tag{1}$$

As a consequence, ‘‘Pollaczek’’ is a household word in some households. Of course, the number of these households may not be so large, but surely every student of queueing theory learns (1) and comes to know the name of Félix Pollaczek. Among queueing theorists, Félix Pollaczek is highly honored as a pioneer in the serious mathematical study of queueing models. Félix Pollaczek is also honored for his persistence in face of adversity; see Cohen (1981). Thus, we are pleased to be able to contribute to this special issue in memory of Félix Pollaczek on the 100th anniversary of his birth.

It is natural to wonder what Pollaczek would think of queueing theory today. What questions would interest him if he were still with us? During his long life, Pollaczek had the opportunity to see the beginning of the rapid development of computers. Nevertheless, he probably would be impressed with the much greater emphasis today on computational methods for queues and related stochastic models. No doubt he would enjoy the vivid color graphics displaying queueing results. He would see how the focus on numerical algorithms has contributed to the development of new theory, including new ways of looking at old models. Now we are much more inclined to look for expressions that will lead to effective numerical algorithms; e.g., see Neuts (1981, 1989).

However, new algorithmic goals do not necessarily mean that we must abandon the established results. For example, many distributions of interest in queueing theory have been characterized in the form of transforms. Instead of lamenting about the ‘‘Laplace curtain’’ obscuring the desired queueing descriptions, and looking for new characterizations, we can try to calculate the distributions and their moments by numerically inverting these transforms. We have been exploring this approach, and we have found it to be very effective. We have found that it is possible to calculate probability distributions directly from their transforms; see Abate and Whitt (1992a,b,c, 1993). We have found that it is possible to calculate moments and asymptotic parameters from moment generating functions, see Choudhury and Lucantoni (1992). We also have recently developed algorithms for BMAP/GI/1 queues that combine matrix-analytic methods in Lucantoni (1991) with transform inversion; see Choudhury (1992) and Abate, Choudhury and Whitt (1993a,b).

With this in mind, we think that Pollaczek might well want to see what numerical algorithms he could develop from his own results. He might hope that the new computational emphasis would not eliminate the need for his analytical results, but instead would help make his early analytical results even more useful. Since he no longer can do this, we intend to do it for him. If he is watching, we hope that he is pleased.

In fact, a significant start in this direction was already made by De Smit (1983a,b). Pollaczek had the opportunity to appreciate De Smit’s (1971, 1973) generalization (in a thesis directed by J. W. Cohen) of his analysis of the difficult GI/G/s model in Pollaczek (1961, 1965), but he did not have the opportunity to see the later numerical algorithms for the special case of hyperexponential service-time distributions based on Wiener-Hopf methods in the 1983 papers.

Pollaczek’s general GI/G/s results actually do not require that the service-time distribution be H_k (hyperexponential) or even have a rational Laplace transform. In particular, the double transform of the waiting times of successive customers $\sum_{n=1}^{\infty} r^n E e^{-\theta W_n}$ is obtained from the solution of s simultaneous linear integral equations, involving s-dimensional contour integrals;

see pp. 157-159 of De Smit (1973). We believe that this representation can be the basis for effective algorithms with general distributions, at least when s is not too large, but this remains an important topic for future research.

In this paper we have a more modest goal. Here we investigate whether it is possible to do numerical calculations from Pollaczek's more elementary GI/G/1 formulas. Pollaczek (1952) derived contour-integral expressions for the Laplace transform and the cumulants of the steady-state waiting-time distribution in the general GI/G/1 queue. This GI/G/1 model has one server, the first-in first-out service discipline, and i.i.d. (independent and identically distributed) service times that are independent of i.i.d. interarrival times, where both the interarrival times and service times can have general distributions.

To state Pollaczek's results, let V be a generic service time and let U be a generic interarrival time. We assume throughout that $0 < EV < EU < \infty$, so that $\rho \equiv EV/EU < 1$ and the steady-state waiting time has a proper distribution; see Asmussen (1987). Let $G(t)$ be the cumulative distribution (cdf) of $V - U$ and let $\phi(z)$ be its transform, defined by

$$\phi(z) = Ee^{z(V-U)} \equiv \int_{-\infty}^{\infty} e^{zt}\, dG(t) = Ee^{zV} Ee^{-zU} , \tag{2}$$

which we assume is analytic for complex z in the strip $|\mathrm{Re}\, z| < \delta$ for some $\delta > 0$. A natural sufficient condition for this analyticity condition is for the service-time and interarrival-time distributions to have finite moment generating functions in a neighborhood of the origin, and thus moments of all orders, but *neither the transform of the interarrival-time distribution nor the transform of the service-time distributions need be rational.* (As a consequence, neither distribution need be phase type.)

Moreover, as noted on p. 40 of Pollaczek (1965) and in §II.5.9 on p. 31 of Cohen (1982), it is possible to treat the case of more general service-time distributions by considering limits of service-time distributions that satisfy this analyticity condition. We discuss this extension here in

§4. Now we assume that $\phi(z)$ in (2) is indeed analytic for complex z in the strip $|\mathrm{Re}\, z| < \delta$ for some $\delta > 0$.

Let W be the steady-state waiting time. Our first Pollaczek formula is for the Laplace transform of W, namely,

$$Ee^{-sW} = \exp\left\{- \frac{1}{2\pi i} \int_C \frac{s}{z(s-z)} \log[1-\phi(-z)]\, dz\right\}, \tag{3}$$

where s is a complex number with $Re(s) \geq 0$, C is a contour to the left of, and parallel to, the imaginary axis, and to the right of any singularities of $\log[1-\phi(-z)]$ in the left half plane; see (xii) on p. 33 of Syski (1965), Theorem 3 on p. 41 of Syski (1967) or Chapter 5 of Cohen (1982) as well as Pollaczek (1952, 1957) and Le Gall (1962) (in French). Formula (3) was also derived by Kingman (1962b); see p. 348.

Let $c_n(W)$ be the n^{th} cumulant of W, i.e., the n^{th} derivative of $\log Ee^{sW}$ evaluated at $s = 0$. Recall that the first cumulant is the mean, the second cumulant is the variance, the third cumulant is the central third moment $E(W - EW)^3$, and the fourth cumulant is

$$c_4(W) = E(W - EW)^4 - 3(E(W - EW)^2)^2 ; \tag{4}$$

e.g., see p. 37 of Riordan (1958). From (3), we obtain two more Pollaczek formulas

$$c_n(W) = \frac{(-1)^n n!}{2\pi i} \int_C \log[1-\phi(-z)] \frac{dz}{z^{n+1}} \tag{5}$$

and

$$P(W > 0) = 1 - \exp\left\{- \frac{1}{2\pi i} \int_C \log[1-\phi(-z)] \frac{dz}{z}\right\} \tag{6}$$

where C is the contour in (3); see (5.7)–(5.9) on p. 42 of Syski (1967).

Our goal is to compute using (3), (5) and (6). However, first we point out the connection to the well known Spitzer (1956) formula

$$Ee^{-sW} = \exp\left\{ \sum_{n=1}^{\infty} n^{-1} E[e^{-sS_n^+} - 1] \right\} , \tag{7}$$

where s is again a complex number with Re $s \geq 0$, S_n is the n^{th} partial sum of i.i.d. copies of $V - U$ (with distribution the n-fold convolution of G), and $x^+ = \max\{x, 0\}$; see (5.50) on p. 280 of Cohen (1982) and (4.3) on p. 174 of Asmussen (1987). A nice derivation of (7) appears in §8.5 of Chung (1974). For additional discussion, see Spitzer (1956, 1957, 1960, 1964), Kingman (1961, 1962a,b, 1966). Kemperman (1961), Le Gall (1962) and Chapter VIII of Siegmund (1985).

Paralleling (5) and (6), we also have the formulas

$$EW = \sum_{n=1}^{\infty} n^{-1} E(S_n^+) , \quad \mathrm{Var}(W) = \sum_{n=1}^{\infty} n^{-1} E((S_n^{+2})) \tag{8}$$

and

$$P(W > 0) = 1 - \exp\{ - \sum_{n=1}^{\infty} n^{-1} P(S_n > 0) \} ; \tag{9}$$

for (9) see p. 350 of Kingman (1962b).

Even though (3) and (7) look quite different, they are easily shown to be equivalent, as was done by Syski (1965, 1967), and as evidently was known to Pollaczek (e.g., see p. 40 of Pollaczek (1965)). Theorem 2 of Kingman (1962b) also established (3) from Spitzer's formula, exploiting an alternative integral representation established by Spitzer (1957), but without reference to Pollaczek. Kingman (1962b) also shows that (3) is equivalent to yet another integral representation for the Laplace transform of W in the GI/G/1 queue established by Smith (1953). Later, Kingman (1966) discusses the connection to Pollaczek (1952, 1957).

It is significant that the integral representations equivalent to (3) have played an important role in asymptotics. In particular, Kingman (1961, 1962a) used these integral representations to establish his classic heavy-traffic limit theorems for the GI/G/1 queue; also see p. 596 of Cohen (1982).

In the remainder of this paper we point out that it is relatively straightforward to numerically calculate the tail probabilities $P(W > x)$ from (3), the cumulants $c_n(W)$ from (5) and the probability of delay $P(W > 0)$ from (6). In §2 we describe the algorithms; in §3 we discuss a few examples; in §4 we discuss the case in which the transform $\phi(z)$ in (2) is not analytic in z for z in a strip $|\mathrm{Re}\, z| < \delta$ for some $\delta > 0$; in §5 we discuss how to obtain the waiting-time asymptotic decay rate η in (12) below from the cumulants in (5); and in §6 we state our conclusions.

Pollaczek (1952, 1957) also derived expressions for the transient distributions, which can also be the basis for effective algorithms. We intend to discuss algorithms for the transient behavior in a subsequent paper.

2. The Algorithms

We first describe an algorithm to compute the tail probabilities $P(W > x)$ based on (3). Then we describe an algorithm to compute the cumulants $c_n(W)$ and the probability of delay $P(W > 0)$ based on (5) and (6). All computations are done using double precision.

2.1 Tail Probabilities

Abate and Whitt (1992a, 1993) describe algorithms for computing tail probabilities $F^c(x) \equiv P(W > x)$ by numerically inverting the Laplace transform

$$\hat{F}^c(s) \equiv \int_0^\infty e^{-sx} F^c(x)\, dx = \frac{1 - Ee^{-sW}}{s} \ . \tag{10}$$

For example, the algorithm EULER there reduces to a finite weighted sum of terms $\mathrm{Re}\, \hat{F}^c(u + kvi)$ over integers k for appropriate real numbers u and v (the number of different k

might be as low as 30.) To apply this algorithm, it suffices to compute Re $\hat{F}^c(s)$ in (10) for s of the required form $s = u + kvi$. For this purpose, it suffices to compute Ee^{-sW} in (3) for s of this same form.

The standard expression for (3), e.g., as in Syski (1965, 1967), has the contour just to the left of the imaginary axis, but this poses numerical difficulties because of the singularity in the first portion of the integrand, $s/z(s-z)$, and in the second portion, $\log[1-\phi(-z)]$, at $z = 0$. However, this difficult is easily avoided by moving the vertical contour of integration to the left, but still keeping it to the right of the singularity of $\log[1-\phi(-z)]$ in the left halfplane closest to the origin, which we denote by $-\eta$. It turns out that this critical singularity of $\log[1-\phi(-z)]$ also corresponds to the singularity of Ee^{-sW} in the left halfplane closest to the origin; i.e.,

$$\eta = \sup\{s > 0 : Ee^{sW} < \infty\} . \tag{11}$$

Given a reasonable estimate of η, we perform the integration (3) by putting the contour at $-\eta/2$. On this contour, $z = -\eta/2 + iy$ and y ranges from $-\infty$ to $+\infty$. Equation (3) becomes $Ee^{-sW} = \exp(-I)$, where

$$\begin{aligned} I &= \frac{1}{2\pi}\left[\int_{-\infty}^{0} \frac{s}{z(s-z)}\log[1-\phi(z)]\,dy + \int_{0}^{\infty} \frac{s}{z(s-z)}\log[1-\phi(-z)]\,dy\right] \\ &= \frac{1}{2\pi}\int_{0}^{\infty}\left[\frac{s}{\bar{z}(s-\bar{z})}\log[1-\phi(-\bar{z})] + \frac{s}{z(s-z)}\log[1-\phi(-z)]\right]dy \end{aligned} \tag{12}$$

with $\bar{z} = -\eta/2 - iy$. In general, I in (12) is complex; we compute its real and imaginary parts by integrating the real and imaginary parts of the integrand, respectively. However, if s is real, then so is I. In that case, the real parts of the two components of the integrand are the same, thereby simplifying the computation somewhat.

For the GI/G/1 queue, the desired parameter η in (11) can usually be easily found by solving the transform equation

$$\phi(\eta) \ = \ 1 \tag{13}$$

for ϕ given in (2); see p. 269 of Asmussen (1987) and Abate, Choudhury and Whitt (1992a). (We describe yet another way to get η in (11) from the cumulants using (5) in §5.) Since ϕ is convex, η in (13) is easily found by search. It suffices to restrict attention to the interval $(0,\eta_s)$, where

$$\eta_s \ = \ \sup\{ s \ \geq \ 0 : Ee^{sV} \ < \infty \} \tag{14}$$

with V being a service time. (Of course η_s can be infinite, but that presents no major difficulty; in that case we start the search in the interval $(0,1)$. If the interval does not contain a root of (13), then we geometrically increase the upper limit until it contains the root.) The value of η_s in (14) is easily found from the service-time Laplace transform using the algorithm in Choudhury and Lucantoni (1992).

However, it can happen that transform equation (13) does not have a root even though the transform ϕ in (2) satisfies the analyticity condition; e.g., see the M/G/1 example in Example 5 of Abate, Choudhury and Whitt (1992a). This means that $\eta \ = \ \eta_s \ > \ 0$ for η in (11) and η_s in (14), so that we can still put the vertical contour at $-\eta/2$. We remark that our algorithm has no difficulty with Example 5 of Abate, Choudhury and Whitt (1992a).

The specific numerical integration procedure we used is fifth-order Romberg integration, as described in §4.3 of Press, Flannery, Teukolsky and Vetterling (1988). We first divide the integration interval $(0,\infty)$ in (12) into a number of subintervals. If η is not too close to 0, then no special care is needed and it suffices to use the two subintervals $(0,1)$, and $(1,\infty)$ and then transform the infinite interval into $(0,1)$ using the transformation in (4.4.2) of Press et al. (1988).

However, more care is required for less well behaved distributions (e.g., highly variable, nearly deterministic, or when η is close to 0). Then we examine the integrand more carefully and choose subintervals so that the ratio of the maximum to the minimum value within any subinterval is at most 10 or 100. This helps ensure that computational effort is expended where it is needed. Indeed, a version of the algorithm was developed to do this automatically. In this automatic procedure, the integration interval $(0,\infty)$ in (12) is divided into $m+1$ subintervals: $(0,b_1)$, $(b_1,b_2),\ldots,(b_{m-1},b_m)$, (b_m,∞). The last infinite subinterval (b_m,∞) is transformed into the finite interval $(0,b_m^{-1})$ using the transformation in (4.4.2) of Press et al. (1988). Within each subinterval, a fifth-order Romberg integration procedure is performed. An error tolerance of 10^{-12} is specified and the program generates successive partitions (going from n to $2n$ points) until the estimated improvement is no more than either the tolerance value itself or the product of the tolerance and the accumulated value of the integral so far (in the current subinterval as well as in earlier subintervals).

The specific procedure used for choosing the subintervals is as follows. If the integrand doesn't differ by more than a factor of 10 in the interval $(0,1)$ then b_1 is chosen as 1. Otherwise, b_1 is chosen such that the integrand roughly changes by a factor of 10 in the interval $(0,b_1)$. The endpoint b_1 is roughly determined by evaluating the integrand at 0 and at the points 10^{-n} with $n = 10,9,\ldots,0$. For $2 \le i \le m$, the ratio b_i/b_{i-1} is assumed to be a constant K, where K is an input parameter. The number m is determined by looking at the ratio of the contribution from the subinterval (b_{i-1},b_i) to the total contribution so far. If this ratio is less than a constant ε, where ε is a second input parameter, then m is set to i, i.e., the next interval is made the last interval. A good choice of K and ε depends on the service-time and interarrival-time distributions. Typically less well behaved distributions require smaller K and/or ε. Our numerical experience indicates that $K = 3$ and $\varepsilon = 10^{-4}$ works pretty well for most cases of interest. Specifically, we observed that, with these choices of K and ε, the computation time for each point of the waiting-time distribution is less than 1 minute (often much less) on a SUN workstation for all the numerical examples reported in this paper.

The Laplace transform inversion algorithm EULER also gives an estimate of the final error. If it is close to or below the 10^{-8} precision specified, we can be fairly confident of a good computation.

2.2 The Cumulants and the Probability of Delay

The algorithms for the cumulants $c_n(W)$ and the probability of delay $P(W > 0)$ using (5) and (6) are easier than the algorithm for the tail probabilities $P(W > x)$ based on (3), because they do not require the numerical transform inversion step. We simply calculate the integrals in (5) and (6), using the same contour, the same subintervals and the same Romberg integration.

Since $c_n(w)$ and $P(W > 0)$ are real quantities (unlike Ee^{-sW}), there is further simplification in the integrals. From (5) and (6), we get

$$c_n(W) = \frac{(-1)^n n!}{\pi} \int_0^\infty \mathrm{Re}\left[\frac{\log[1-\phi(-z)]}{z^{n+1}}\right] dy \tag{15}$$

and

$$P(W > 0) = 1 - \exp\left\{-\frac{1}{\pi} \int_0^\infty \mathrm{Re}\left[\frac{\log[1-\phi(-z)]}{z}\right] dy\right\} , \tag{16}$$

where $z = -\eta/2 + iy$. We observed that the computation time for each cumulant of the waiting-time distribution is typically between a fraction of a second to a few seconds on a SUN workstation.

3. Gamma Queue Examples

We tried a variety of examples, including examples with deterministic and two-point distributions. As noted in Abate and Whitt (1992a), such deterministic distributions tend to be more difficult for the numerical transform inversion, but satisfactory accuracy can usually be obtained at the expense of somewhat greater computational effort.

The computation tends to be much easier with smooth densities. There is no requirement that the transforms be rational. To illustrate, in this section we consider $\Gamma_\alpha / \Gamma_\beta / 1$ queues, where Γ denotes the gamma distribution, and α and β are the shape parameters of the interarrival-time and service-time distributions, respectively. The gamma distribution with scale parameter λ and shape parameter α has density

$$f(x) = \frac{1}{\Gamma(\alpha)} \lambda^\alpha x^{\alpha - 1} e^{-\lambda x} , \ x > 0 , \tag{17}$$

mean α / λ, variance α / λ^2 and Laplace transform

$$E e^{-sV} \equiv \int_0^\infty e^{-sx} f(x)\, dx = \left[\frac{\lambda}{\lambda + s} \right]^\alpha . \tag{18}$$

The transform in (18) is rational if, and only if, the shape parameter α is a positive integer. When $\alpha = k$ for an integer k, the gamma distribution is also called Erlang of order k (E_k). From §12 of Abate and Whitt (1992a), which considers convolutions of exponential distributions, we expect that this distribution will not be very difficult, at least when α is not too small.

We stipulate that the mean service time is 1 and that the arrival rate is ρ. The remaining two parameters α and β of the $\Gamma_\alpha / \Gamma_\beta / 1$ queue are the shape parameters of the interarrival-time and service-time distribution. Since the squared coefficient of variation (SCV, variance divided by the square of the mean) is the reciprocal of the shape parameter, it suffices to specify the SCVs c_a^2 and c_s^2 of the interarrival-time and service-time distributions.

We checked our algorithm against known results by considering the $E_k/\Gamma/1$ and $\Gamma/E_k/1$ special cases. These are special cases of the PH/G/1 and GI/PH/1 queues, for which algorithms were developed by Choudhury (1992), exploiting results for the M/G/1 and GI/M/1 paradigms in Neuts (1981, 1989) and Lucantoni (1991). For example, the new algorithm here agrees for the

$E_2/\Gamma_{1/2}/1$ queue, which is considered in Example 7.3 of Choudhury and Whitt (1992). We also compared our numerical results with numerical results in Chaudhry, Agarwal and Templeton (1992). Also, we found good agreement with results for the $E_{10}/E_{100}/1$ queue in Table 10 on p. 141 of Chaudhry, Agarwal and Templeton (1992).

Some other algorithms for $E_k/E_m/1$ queues get more difficult as k and m increase. Hence, we performed calculations for $E_k/E_k/1$ models with large k.

Example 3.1 $E_k/E_k/1$ Queues.

We did calculations for the $E_k/E_k/1$ queue for $k = 10$, $k = 100$, $k = 1000$ and $k = 10,000$. In this case the transform equation in (13) for the asymptotic decay rate η becomes

$$\left[\frac{k}{k-\eta} \right]^k \left[\frac{k}{k+\eta/\rho} \right]^k = 1 \; , \tag{19}$$

from which we easily obtain

$$\eta = k(1-\rho) \; . \tag{20}$$

Since E_k is approaching a deterministic distribution as k increases, to avoid having negligible probabilities we let $\rho \equiv \rho_k$ increase with k. In particular, we let $\rho_k = 1 - k^{-1}$. With this choice, $\eta \equiv \eta_k = 1$ for all k. Also W_k, the steady-state waiting time in model k, converges to an exponential random variable with mean 1 as $k \to \infty$, as can be seen by applying the heavy-traffic argument of Kingman (1961) using (3).

Numerical values of some tail probabilities and cumulants are given for $E_k/E_k/1$ queues for these cases in Table 1. The exponential limit is displayed as well under the heading $k = \infty$. None of these presented any numerical difficulties.

Interestingly, from Table 1, we see that for these cases W is quite well approximated by a mixture of an atom at 0 with probability $1/\sqrt{k} = \sqrt{1-\rho}$ and an exponential with mean 1 with probability $1 - 1/\sqrt{k}$.

Example 3.2. Constant Sum of SCVs, Low Variability

A simple approximation for the mean steady-state waiting time in a GI/G/1 queue, for which $EV = 1$, is

$$EW = \frac{\rho(c_a^2 + c_s^2)}{2(1 - \rho)} \ . \tag{21}$$

In order to see the impact of unusual variability (e.g., departure from the *M* assumption), it is interesting to see how the queue behaves as c_a^2 and c_s^2 change with $c_a^2 + c_s^2$ held fixed. Hence, we consider two cases: very low variability, where $c_a^2 + c_s^2 = 0.20$, and very high variability, where $c_a^2 + c_s^2 = 100$. Tables 2 and 3 display numerical results for these two cases.

For the low variability case, we think of the $E_{10}/E_{10}/1$ (or $\Gamma_{10}/\Gamma_{10}/1$) system as our reference system. We consider four other systems: $\Gamma_{(1/.05)}/\Gamma_{(1/.15)}/1$, $\Gamma_{(1/.15)}/\Gamma_{(1/.05)}/1$, $\Gamma_{(1/.01)}/\Gamma_{(1/.19)}/1$ and $\Gamma_{(1/.19)}/\Gamma_{(1/.01)}/1$. The last two cases are relatively extreme within this class. They clearly correspond approximately to $D/E_5/1$ and $E_5/D/1$.

We use traffic intensity $\rho = 0.8$ to have non-negligible values, but not to be too much in heavy traffic. As $\rho \to 1$, we know that the distribution is asymptotically exponential with the mean in (19). Consistent with the heavy traffic approximation, we see from Table 2 that the congestion measures in this low variability case do not differ much as we change c_a^2 with $c_a^2 + c_s^2$ fixed at 0.2. From Table 2, we also see that $P(W > 0)$ and *EW* decrease with α, consistent with *lower congestion,* but the higher tail probabilities $P(W > k)$, the higher cumulants $c_k(W)$ and the reciprocal of the asymptotic decay rate, η^{-1}, increase with α, indicating *higher congestion.* This may be considered consistent with intuition: higher α corresponds to a more variable service time, which we expect to make big delays more likely, while lower α corresponds to more variable interarrival times, which we expect to make the probability of having to wait higher.

Example 3.3. Constant Sum of SCVs, High Variability.

Now we consider the high variability case. For the high variability case, we think of $\Gamma_{(1/50)}/\Gamma_{(1/50)}/1$ as the reference system. We consider four other systems: $\Gamma_{(1/20)}/\Gamma_{(1/80)}/1$, $\Gamma_{(1/80)}/\Gamma_{(1/20)}/1$, $\Gamma_{(1/.1)}/\Gamma_{(1/99.9)}/1$ and $\Gamma_{(1/99.9)}/\Gamma_{(1/.1)}/1$. The last two cases are relatively extreme cases within this class. They clearly correspond approximately to $E_{10}/\Gamma_{(1/100)}/1$ and $\Gamma_{(1/100)}/E_{10}/1$. They are interesting because one distribution has very high variability while the other distribution has very low variability.

Since the variability is so high, we use traffic intensity $\rho = 0.1$ to obtain moderate values. Since ρ is very small and the variability is very high, we expect that the congestion measures will vary much more than in the low-variability case, and this is demonstrated by Table 3. For example, the probability of delay $P(W > 0)$ ranges from 0.085 to 0.965. These cases are rather pathological and would be stressful for the algorithm without an adaptive choice of the subintervals for the numerical integration. However, with the adaptive choice of integration subintervals, these cases present no serious difficulty. The error estimates on the final distribution calculations were good, all being of order 10^{-10}.

4. Long-Tail Service-Time Distributions

Pollaczek’s integral representations in (3), (5) and (6) depend on an analyticity condition that is satisfied when the interarrival-time and service-time distributions have finite moment generating functions in a neighborhood of the origin; i.e., when $Ee^{sU} < \infty$ and $Ee^{sV} < \infty$ for some $s > 0$. However, as noted on p. 40 of Pollaczek (1965) and on p. 310 of Cohen (1982), it is possible to treat the general case by representing a general distribution as a limit of a sequence of distributions each of which satisfies this analyticity condition. It is known that the associated sequence of steady-state waiting-time distributions will converge to a proper limit provided that the distributions and their means converge to proper limits; see p. 194 of Asmussen (1987). (The moment condition is actually on $(V-U)^+ = \max\{V-U, 0\}$.)

In fact, the long-tail interarrival-time distributions actually present no difficulty. It suffices to have $\phi(z)$ in (2) analytic in the strip $0 < \mathrm{Re}(z) < \delta$ for some $\delta > 0$. However, the service-time distribution poses a real problem.

Hence, if $H^c(x)$ is the given service-time complementary cdf with Laplace transform $\hat{H}^c(s)$ and mean $m = \hat{H}^c(0)$, then it suffices to find an approximating sequence of service-time complementary cdf's $\{H_n^c(x) : n \geq 1\}$ with associated Laplace transforms $\{\hat{H}_n^c(s) : n \geq 1\}$ and means $\{m_n = \hat{H}_n^c(0) : n \geq 1\}$ such that $\hat{H}_n^c(s) \to \hat{H}^c(s)$ as $n \to \infty$ for all s. Then $H_n^c(x) \to H^c(x)$ as $n \to \infty$ for all x that are continuity points of the limiting complementary cdf H^c and $m_n \to m$ as $n \to \infty$.

A natural way to obtain a sequence of approximating service-time distributions with finite moment generating functions in some neighborhood of the origin when this condition is not satisfied originally is to simply truncate the given service-time distribution, as on p. 310 of Cohen (1982). For example, we can choose a sequence of truncation points $\{t_n : n \geq 1\}$ and let

$$H_n^c(x) = \begin{cases} H^c(x) - H^c(t_n) \ ; & 0 < x < t_n \\ 0 \ , & x \geq t_n \end{cases} \tag{22}$$

In this scheme we move the mass in the interval $[t_n, \infty)$ to the origin. Alternatively, we could move the mass in the interval $[t_n, \infty)$ to the point t_n.

For our purposes, a difficulty with truncation is that it does not seem so easy to calculate the modified Laplace transform $\hat{H}_n^c(s)$ given the original transform $\hat{H}^c(s)$. An easier procedure is to introduce *exponential damping* in the Laplace-Stieltjes transform with respect to H, as on pp. 15, 29 of Abate and Whitt (1992a). In particular, for any $\alpha > 0$ let the α-damped complementary cdf be

$$H_\alpha^c(x) \;=\; \int_x^\infty e^{-\alpha t} dH(t) \;, \quad x \geq 0 \;. \tag{23}$$

Since we want a proper probability distribution, we put mass $1 - H_\alpha^c(0)$ at 0. If the original service-time distribution has mean 1 and we want the new service-time distribution also to have mean 1, then we also divide the random variable V_α with cdf H_α by the new mean m_α, i.e., we let the complimentary cdf be $H_\alpha^c(m_\alpha x)$.

The direct approximation in (23) makes the service-time distribution stochastically smaller than the original service-time distribution, which in turn makes the new steady-state waiting-time distribution stochastically smaller than the original one, see Stoyan (1983), which may be helpful for interpretation. However, keeping the same mean seems to give somewhat better numbers. Here we keep the mean fixed at 1.

From (23), it is easy to see that, if $\hat{h}(s)$ is the original Laplace-Stieltjes transform of *H*, then the Laplace-Stieltjes transform of $H_\alpha(m_\alpha x)$ with mean 1 is

$$\hat{h}_\alpha(s) \;=\; \hat{h}(\alpha \;+\; (s/m_\alpha)) \;+\; 1 - \hat{h}(\alpha) \;, \tag{24}$$

where

$$m_\alpha \;=\; -\hat{h}_\alpha'(0) \;=\; -\hat{h}\,'(\alpha) \;. \tag{25}$$

As in (10), the Laplace transform $\hat{H}_\alpha^c(s)$ of $H_\alpha^c(m_\alpha x)$ for $H_\alpha^c(x)$ in (21) is

$$\hat{H}_\alpha^c(s) \;=\; \frac{1 \;-\; \hat{h}_\alpha(s)}{s} \tag{26}$$

for $\hat{h}_\alpha(s)$ in (22). Hence, given $\hat{h}(s)$, we can readily calculate $\hat{H}_\alpha^c(s)$ for any $\alpha \;>\; 0$.

However, this approach is not trivial to implement, because it often requires a very small α before $H_\alpha(x)$ is a satisfactory approximation for $H(x)$, and a small α means a small η in (13). Indeed, $0 \;<\; \eta \;\leq\; \eta_s \;=\; \alpha$. In turn, a small η means a relatively difficult computation, because the contour at $-\eta/2$ is near the singularity at 0. However, this can be handled by being careful with the numerical integration. Indeed, our algorithm employing an adaptive choice of integration subintervals was developed to handle this case. (We found that the algorithm works well in other cases too.)

Example 4.1. An M/G/1 Queue.

To illustrate how this damping approach works, we consider an M/G/1 queue with service-time density

$$h(x) = x^{-3}(1-(1+2x+2x^2)e^{-2x}) , \quad x \geq 0 . \tag{27}$$

This distribution has first two moments $m_1 = 1$ and $m_2 = \infty$, so that there is a proper steady-state waiting-time distribution which has infinite mean; see pp. 181-184 of Asmussen (1987). It is easy to see that our service-time distribution has complementary cdf

$$H^c(x) = (2x^2)^{-1}(1-(1+2x)e^{-2x}) , \; x \geq 0 , \tag{28}$$

and Laplace transform

$$\hat{h}(s) = 1-s+\frac{s^2}{2}\ln(1+(2/s)) . \tag{29}$$

We use the M/G/1 queue to make it easier to compare our numerical results with other known results. First, since the Laplace transform Ee^{-sW} is available from the Pollaczek-Khintchine (transform) formula for the M/G/1 queue, we can also directly apply the numerical inversion algorithms in Abate and Whitt (1992a,b) to this example. Moreover, we can determine the asymptotic behavior of the steady-state waiting-time complementary cdf $P(W > x)$, as we show in Abate, Choudhury and Whitt (1993c). In particular, the first two terms are given by

$$P(W > x) \sim \frac{\rho}{2(1-\rho)x}\left[1 + \frac{\rho}{(1-\rho)x}[\ln(2x) - 1]\right] \text{ as } x \to \infty , \tag{30}$$

where $f(x) \sim g(x)$ means that $f(x)/g(x) \to 1$; e.g., see §22 of Borovkov (1976) and Willekens and Teugels (1992) for related results.

In Table 4 we display numerical results for this M/G/1 example in the case $\rho = 0.8$. We display numerical results for the tail probabilities $P(W > x)$ for five values of x: $x = 4$, $x = 20$, $x = 100$, $x = 500$ and $x = 2500$. We display the exact results (no damping, $\alpha = 0$) obtained by the algorithm EULER in Abate and Whitt (1992a,b) and both the one-term and two-term asymptotic approximations based on (28). We also display the approximations obtained from five values of the damping parameter α: $\alpha = 10^{-2}$, $\alpha = 10^{-3}$, $\alpha = 10^{-4}$, $\alpha = 10^{-6}$ and $\alpha = 10^{-8}$. The numerical results based on the new algorithm here and EULER agreed to the stated precision in all cases except $\alpha = 10^{-8}$. For $\alpha = 10^{-8}$, both numerical results are given, from which we see that the agreement is excellent.

From Table 4, we see that the damping parameter α needs to be smaller and smaller as x increases in order for the calculations based on the approximating cdf H_α to be accurate. However, the calculation gets more difficult as x increases and α decreases. For the smaller α values reported, it was important to carefully choose the subintervals for the Romberg integration so that the integrand does not fluctuate too greatly within the subinterval. This was done by the automatic procedure described in §2.1.

In this example we are able to obtain a good calculation for all x because the asymptotics apply before the computation gets difficult. The relative percent error for the one-term (two-term) approximations at $x = 100$, $x = 500$ and $x = 2,500$ are, respectively, 19%, 5.2% and 1.4% (5.6%, 0.8% and 0.1%).

5. Using the Cumulants to Get the Decay Rate

We have suggested solving the transform equation (13) in order to get the decay rate η and thus the location of the contour, $-\eta/2$. However, η can also be estimated from higher cumulants using the cumulant contour integral (5). In particular, $c_j/(j-1)! \sim \eta^{-j}$ as $j \to \infty$, so that $jc_j/c_{j+1} \to \eta$ as $j \to \infty$.

Proposition. *Let X be a nonnegative random variable. If $P(X > x) \sim \alpha e^{-\eta x}$ as $x \to \infty$ for positive constants α and η, then $c_j \equiv c_j(X) \sim (j-1)!/\eta^j$ as $j \to \infty$.*

Proof. Let $\mu_j = EX^j$. Under the exponential asymptotics condition, $\mu_j \sim \alpha j! \eta^{-j}$ as $j \to \infty$ by Theorem 1 of Choudhury and Lucantoni (1992). Therefore, the radius of convergence of $M(z) = \sum_{j=0}^{\infty} \mu_j z^j / j!$ is η and by the final value theorem for generating functions,

$$M(z) = (1 - z\eta^{-1})^{-1} M_1(z) , \tag{31}$$

where $M_1(z)$ is analytic at η and $M_1(z) \to \alpha$ as $z \to \eta$. It follows that $\log M(z)$ also has radius of convergence η, but $\log M(z) = \sum_{j=0}^{\infty} c_j z^j / j!$ Hence,

$$\begin{aligned} \log M(z) &= \sum_{j=0}^{\infty} \frac{c_j z^j}{j!} = -\log(1 - z\eta^{-1}) + \log M_1(z) \\ &= \sum_{j=1}^{\infty} \frac{z^j}{\eta^j j} + \sum_{j=0}^{\infty} m_{1j} z^j , \end{aligned}$$

where the second term is the Taylor series expansion of $\log M_1(z)$ about $z = 0$, so that

$$\frac{c_j}{j!} = \frac{1}{j\eta^j} + m_{1j} , \; j \geq 1 . \tag{32}$$

However, $M_1(z)$ has a radius of convergence $\eta_1 > \eta$, which implies that $j\eta^j m_{1j} \to 0$ as $j \to \infty$, so that $c_j / j! \sim 1/j\eta^j$ as $j \to \infty$. ∎

As a consequence of the Proposition, we can estimate η by jc_j / c_{j+1} for j suitably large. By looking at successive j we can see if convergence is taking place. If we do not estimate η by finding the root to (13), then we can start by guessing an appropriate place for the contour. If the computed η is consistent with this choice, then we can be confident that we have estimated η correctly.

From the Proposition, we see that if $P(W > x) \sim \alpha e^{-\eta x}$ as $x \to \infty$, we can estimate η from the cumulants $c_j(W)$. However, we also see that the asymptotic behavior of the cumulants is independent of the asymptotic constant α. If we are interested in α, then we can apply (3) together with the algorithm in Choudhury and Lucantoni (1992) to directly compute the moments and estimate α as well as η. We have implemented this algorithm and found that it works well. We are interested in the asymptotic parameters α and η because we have found that the one-term asymptotic $\alpha e^{-\eta x}$ is often a remarkably good approximation for $P(W > x)$; see Abate, Choudhury and Whitt (1993a,b).

6. Conclusions

In this paper we have shown that it is not difficult to compute the steady-state waiting-time distribution and its cumulants from Pollaczek's (1952, 1957) formulas (3), (5) and (6). When the contour is near singularities, some care is needed in the numerical integration, but overall it is not difficult. For the tail probabilities, we use numerical transform inversion together with numerical integration of Pollaczek's contour integral (3). For the probability of having to wait before beginning service and the cumulants, we only need to integrate the contour integrals (5) and (6). It is significant that these procedures do not require that the interarrival-time or service-time transform be rational. Moreover, as shown in §4, with appropriate modifications, these procedures apply to long-tail distributions that violate Pollaczek's analyticity condition. Overall, our experience indicates that queueing transforms are very useful for extracting numerical results.

References

J. Abate, G. L. Choudhury and W. Whitt (1993a) Exponential approximations for tail probabilities in queues, I: waiting times, *Oper. Res.*, to appear.

J. Abate, G. L. Choudhury and W. Whitt (1993b) Exponential approximations for tail probabilities in queues, II: sojourn time and workload, submitted.

J. Abate, G. L. Choudhury and W. Whitt (1993c) Waiting-time tail probabilities in queues with long-tail service-time distributions, in preparation.

J. Abate and W. Whitt (1992a) The Fourier-series method for inverting transforms of probability distributions, *Queueing Systems* 10, 5-88.

J. Abate and W. Whitt (1992b) Numerical inversion of probability generating functions, *Oper. Res. Letters* 12, 245-251.

J. Abate and W. Whitt (1992c) Solving probability transform functional equations for numerical inversion, *Oper. Res. Letters* 12, 275-281.

J. Abate and W. Whitt (1993) Numerical inversion of Laplace transforms of probability distributions, *ORSA J. Comput.*, to appear.

S. Asmussen (1987) *Applied Probability and Queues,* Wiley, New York.

A. A. Borovkov (1976) *Stochastic Processes in Queueing Theory,* Springer-Verlag, New York.

M. L. Chaudhry, M. Agarwal and J. G. C. Templeton (1992) Exact and approximate numerical solutions of steady-state distributions arising in the GI/G/1 queue, *Queueing Systems* 10, 105-152.

G. L. Choudhury (1992) An algorithm for a large class of G/G/1 queues, in preparation.

G. L. Choudhury and D. M. Lucantoni (1992) Numerical computation of the moments of a probability distribution from its transforms, submitted.

G. L. Choudhury and W. Whitt (1992) Heavy traffic asymptotic expansions for the asymptotic decay rates in the BMAP/GI/1 queue, submitted.

K. L. Chung (1974) *A Course in Probability Theory,* second ed., Academic Press, New York.

J. W. Cohen (1982) *The Single Server Queue,* second ed., North-Holland, Amsterdam.

J. W. Cohen (1981) Obituary: Félix Pollaczek, *J. Appl. Prob.* 18, 958-963.

J. H. A. De Smit (1971) *Many Server Queueing Systems,* Ph.D. dissertation, University of Technology, Delft.

J. H. A. De Smit (1973) Some general results for many server queues, *Adv. Appl. Prob.* 5, 153-169.

J. H. A. De Smit (1983a) The queue GI/M/s with customers of different types or the queue GI/H_m/s, *Adv. Appl. Prob.* 15, 392-419.

J. H. A. De Smit (1983b) A numerical solution for the multi-server queue with hyperexponential service times, *Oper. Res. Letters* 2, 217-224.

J. H. B. Kemperman (1961) *The Passage Problem for a Stationary Markov Chain,* University of Chicago Press, Chicago.

J. F. C. Kingman (1961) The single server queue in heavy traffic, *Proc. Camb. Phil. Soc.* 57, 902-904.

J. F. C. Kingman (1962a) On queues in heavy traffic, *J. Roy Stat. Soc.,* Ser B, 24, 383-392.

J. F. C. Kingman (1962b) The use of Spitzer's identity in the investigation of the busy period and other quantities in the queue GI/G/1. *J. Aust. Math. Soc.* 2, 345-356.

J. F. C. Kingman (1965) The heavy traffic approximation in the theory of queues, pp. 137-159 in *Proceedings of the Symposium on Congestion Theory,* W. L. Smith and W. E. Wilkinson, eds., The University of North Carolina Press, Chapel Hill.

J. F. C. Kingman (1966) *On the Algebra of Queues,* Methuen, London.

P. Le Gall (1962) *Les Systèmes avec ou sans Attente et les Processus Stochastiques*, Tome I, Dunod.

D. M. Lucantoni (1991) New results on the single server queue with a batch Markovian arrival process, *Stochastic Models* 7, 1-46.

M. F. Neuts (1981) *Matrix-Geometric Solutions in Stochastic Models,* The Johns Hopkins University Press, Baltimore.

M. F. Neuts (1989) *Structured Stochastic Matrices of M/G/1 Type and Their Applications,* Marcel Dekker, New York.

F. Pollaczek (1952) 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, 2334-2336.

F. Pollaczek (1957) *Problèmes Stochastiques Posés par le Phénomène de Formation d'une Queue d'Attente à un Guichet et par des Phénomènes Apparentés* (Mémorial des Sciences Mathématiques, fac. 136), Gauthier-Villars, Paris.

F. Pollaczek (1961) *Théorie Analytique des Problèmes Stochastiques Relatifs à un Groupe de Lignes Teléphoniques avec Dispositif d'Attente,* Mémorial des Sciences Mathématiques, Gauthier-Villars, Paris.

F. Pollaczek (1965) Concerning an analytic method for the treatment of queueing problems, pp. 1-25 and 34-42 in *Proceedings of the Symposium on Congestion Theory,* W. L. Smith and W. E. Wilkinson (eds.), The University of North Carolina Press, Chapel Hill.

W. H. Press, B. P. Flannery, S. A. Teukolsky and W. T. Vetterling (1988) *Numerical Recipes,* FORTAN version, Cambridge University Press, Cambridge, England.

J. Riordan (1958) *An Introduction to Combinatorial Analysis,* Wiley, New York.

D. Siegmund (1985) *Sequential Analysis,* Springer-Verlag, New York.

F. L. Spitzer (1956) A combinatorial lemma and its application to probability theory, *Trans. Amer. Math. Soc.* 82, 323-339.

F. L. Spitzer (1957) The Wiener-Hopf equation whose kernel is a probability density, *Duke Math J.* 24, 327-343.

F. L. Spitzer (1960) A Tauberian theorem and its probability interpretation, *Trans. Amer. Math. Soc.* 94, 150-169.

F. L. Spitzer (1964) *Principles of Random Walk,* Van Nostrand, Princeton, NJ.

D. Stoyan (1983) *Comparison Methods for Queues and Other Stochastic Models,* Wiley, New York.

R. Syski (1965) Discussion on Dr. Pollaczek's paper, pp. 30-34 in *Proceedings of the Symposium on Congestion Theory,* W. L. Smith and W. E. Wilkinson (eds.), The University of North Carolina Press, Chapel Hill.

R. Syski (1967) Pollaczek's method in queueing theory. pp. 33-60 in *Queueing Theory, Recent Developments and Applications,* R. Cruon, (ed.), Engl. Univ. Press, London.

E. Willekens and J. T. Teugels (1992) Asymptotic expansions for waiting time probabilities in an M/G/1 queue with long-tailed service time, *Queueing Systems* 10, 295-312.