AN INTERPOLATION APPROXIMATION FOR THE MEAN WORKLOAD IN A GI/G/1 QUEUE

WARD WHITT

AT&T Bell Laboratories, Murray Hill, New Jersey

(Received April 1988; revision received July 1988; accepted September 1988)

This paper develops a closed form approximation for the mean steady-state workload or virtual waiting time in a GI/G/1 queue, using the first two moments of the service-time distribution and the first three moments plus the density at the origin of the interarrival-time distribution, with default values provided in case information is unavailable. The approximation is based on light and heavy traffic limiting behavior. The essential ideas are exposed by using S.L. Brumelle's formula to relate the mean workload to the mean waiting time and K.T. Marshall's formula to relate the mean waiting time to the first two moments of the idle period. Both formulas extend to general single server models without independence conditions, so this approach provides a basis for extensions, but a convenient exact expression for the second order heavy traffic term is evidently not possible even for GI/G/1. For the GI/G/1 second order heavy traffic term, an approximation is proposed, based on the relatively nice expression established for the GI/M/1 case by S. Halfin. The interpolation between light and heavy traffic limits, which can be applied to other performance measures and models whenever the limits can be determined or approximated, is chosen to satisfy differentiability and monotonicity regularity conditions.

In this paper, we develop a closed form approximation for the mean steady-state workload (or virtual waiting time) in a GI/G/1 queue, based on the first two moments of the service-time distribution and the first three moments plus the density at the origin of the interarrival-time distribution. The approximation formula is obtained by interpolating between light traffic and heavy traffic limits (the limiting behavior as $\rho \to 0$ and $\rho \to 1$ where ρ is the traffic intensity), in the spirit of Burman and Smith (1983, 1986) and Reiman and Simon (1988, 1989). We first determine or approximate the limits; then we determine an appropriate interpolation given the limits. This interpolation applies to more general models, provided that the light and heavy traffic behavior can be determined. Indeed, we apply this interpolation in Fendick and Whitt (1989) to approximate the mean workload in general single server queues without any independence conditions, where exact or approximate light and heavy traffic limits are either calculated from model parameters or estimated from traffic measurements. The approach with measurements applies without specifying a detailed probability model. The GI/G/1 model is a revealing reference case because we can determine the limits more precisely and evaluate the approximation more easily for the GI/G/1 model.

With respect to the literature on interpolation approximations, our major contribution is to incorporate the second order heavy traffic behavior. We also discuss how to do the interpolation, and how to evaluate the reliability of the fit. With respect to the literature on refined heavy traffic approximations (notably, Siegmund 1979, Köllerström 1981, Halfin 1985, Asmussen 1984, 1987, and Knessl 1990), our major contribution is to apply the results via the interpolation, using further approximations. We also show how the light traffic and heavy traffic limits can be identified easily by applying basic formulas due to Brumelle (1971) and Marshall (1968); see (3) and (21) below.

The Normalized Mean Workload in the GI/G/1 Queue

Our queue has one server, unlimited waiting space and a work-conserving discipline; see p. 418 of Heyman and Sobel (1982). Since we focus on the workload, the particular work-conserving discipline does not matter. As usual, GI/G/1 means that the interarrival and service times come from independent sequences of independent and identically distributed (i.i.d.) random variables. We assume that the interarrival and service times have finite third moments. Without loss of generality, let the mean service time be 1, so that the arrival rate (the reciprocal of the mean interarrival time) coincides with the traffic intensity ρ. We consider a family of GI/G/1 models

Subject classifications: Queues, approximations: GI/G/1 mean workload. Queues, limit theorems: light-traffic and heavy-traffic interpolation.

0030-364X/89/3706-0936 $01.25

indexed by ρ obtained by scaling the arrival process; i.e., let T be a nonnegative random variable with mean 1 and let $\rho^{-1}T$ be an interarrival time in the GI/G/1 model with traffic intensity ρ. Let Z_ρ be the steady-state workload as a function of ρ, which we assume exists. (It suffices to have $\rho < 1$ and the interarrival-time distribution be nonlattice; see p. 188 of Asmussen 1987.) We focus on the *normalized mean* (steady-state) *workload*, defined by

$$c_z^2(\rho) = \frac{2(1-\rho)}{\rho} E(Z_\rho). \tag{1}$$

As discussed in Fendick and Whitt, the normalized mean workload is convenient to *see* the effect of the variability in the interarrival-time and service-time distributions upon the mean workload. The normalized mean workload is the ratio of $E(Z_\rho)$ for the GI/G/1 model to what it is in the M/D/1 model. (A similar normalization is used in Burman and Smith 1983.) As a consequence, assuming that the interarrival-time distribution has no atom at the origin (for light traffic), it is relatively easy to show that $c_z^2(\rho)$ has nondegenerate limiting values as $\rho \to 0$ and as $\rho \to 1$. In particular (see Sections 1 and 2)

$$\lim_{\rho \to 0} c_z^2(\rho) = c_z^2(0) = 1 + c_s^2$$

and (2)

$$\lim_{\rho \to 1} c_z^2(\rho) = c_z^2(1) = c_a^2 + c_s^2$$

where c_a^2 and c_s^2 are the squared coefficients of variation (variance divided by the square of the mean) of the interarrival-time and service-time distributions. Of course, for the M/G/1 model $c_z^2(\rho)$ is constant, with $c_a^2 = 1$ in (2).

Derivatives and Approximations

The primary goal of this paper is to determine, at least approximately, and apply the derivatives $\dot{c}_z^2(0)$ and $\dot{c}_z^2(1)$ at the endpoints as well. In Section 1, we discuss the light traffic values $c_z^2(0)$ and $\dot{c}_z^2(0)$; in Section 2, we discuss the heavy traffic values $c_z^2(1)$ and $\dot{c}_z^2(1)$. These are obtained from results for the steady-state waiting W_ρ using the first-in first-out discipline plus Brumelle's formula

$$E(Z_\rho) = \rho E(W_\rho S) + \rho \frac{E(S^2)}{2}$$

$$= \rho E(W_\rho) + \frac{\rho(c_s^2 + 1)}{2} \tag{3}$$

where S is the service time of the customer with waiting time W_ρ; see pp. 408–412 of Heyman and Sobel. (Brumelle's formula stems from the generalization of $L = \lambda W$ known as $H = \lambda G$; it is not limited to GI/G/1.)

From (3) it is clear that an expression (exact or approximate) for $E(Z_\rho)$ yields an expression for $E(W_\rho)$ and vice versa. Exact expressions for both $E(W_\rho)$ and $E(Z_\rho)$ are known (e.g., Chapter VIII of Asmussen), and algorithms for numerical computation are available for a large class of interarrival-time and service-time distributions (e.g., pp. 216, 239 of Asmussen 1987, Ramaswami and Latouche 1987 and Tijms 1986), but relatively simple formulas are nevertheless helpful (for example, for understanding, optimization, and approximation of larger models such as queueing networks, as in Whitt 1983 and Segal and Whitt 1988.)

A familiar elementary approximation for $E(W_\rho)$ is

$$E(W_\rho) \approx \frac{\rho(c_a^2 + c_s^2)}{2(1-\rho)} \tag{4}$$

e.g., see (8) of Whitt (1982). Combining (1), (3) and (4), we obtain the simple linear interpolation of (2) as an approximation for the normalized mean workload

$$c_z^2(\rho) \approx 1 + c_s^2 + \rho(c_a^2 - 1)$$

$$= \rho c_z^2(1) + (1-\rho)c_z^2(0). \tag{5}$$

Indeed, a nice way to *derive* (4) is to apply (1)–(3) and (5). A better two-moment approximation for $E(W_\rho)$ than (4), and thus for $E(Z_\rho)$ via (3), was developed by Kraemer and Langenbach-Belz (1976), but we aim to do even better by using the third moment and the density at the origin of the interarrival-time distribution, which are the additional quantities that appear in the light traffic derivative and our approximation for the heavy traffic derivative.

A byproduct of our analysis is additional support for the conclusion that, given the first two moments of the interarrival-time and service-time distribution, $E(W_\rho)$ and $E(Z_\rho)$ depend more on additional information about the interarrival-time distribution than the service-time distribution; see Section VII of Whitt (1984b). Moreover, for fixed first two arrival and service moments, $E(W_\rho)$ and $E(Z_\rho)$ tend to be *increasing* in the interarrival-time density at the origin and *decreasing* in the interarrival time third moment. Indeed, this monotonicity in the third moment was proved for a large class of GI/G/1 queues by Halfin.

In this paper, we only consider the expected steady-state workload, but similar asymptotics can be determined for the expectation of other functions of the

steady-state workload such as higher moments, for example, see (10) in Siegmund (1979), (3.1) and (3.3) of Halfin, and pp. 183–185, 272–276, 281–287 of Asmussen and Knessl.

The Interpolation

The interpolation problem is to pick a reasonable real-valued function f defined on the unit interval [0, 1] that matches specified values $f(0)$ and $f(1)$, and derivatives $f'(0)$ and $f'(1)$. Our solution is a function of the form

$$f(\rho) = a_0 + a_1\rho + a_2\rho^2 + a_3\rho^{b_3} + a_4(1 - \rho)^{b_4} \qquad 0 \leq \rho \leq 1 \quad (6)$$

where the coefficients a_i and the exponents b_i depend on $f(0)$, $f(1)$, $f'(0)$ and $f'(1)$, as described in Section 3. Obviously, the fit to (6) is not unique.

The interpolation problem may seem elementary, but we believe that it deserves some care. For example, one might consider the unique cubic function ((6) with $b_3 = 3$ and $a_4 = 0$) determined by these four values. This cubic function often is reasonable, but unfortunately, in general, it has undesirable properties. First, in some cases, the cubic function takes negative values which is clearly unreasonable for $c_z^2(\rho)$ in (1). Second, the cubic function or its derivative may fail to be monotone when the four values $f(0)$, $f(1)$, $f'(0)$ and $f'(1)$ permit monotonicity. For example, this occurs when we apply the cubic fit to the normalized mean workload in the $E_2/M/1$ queue (Erlang interarrival times and exponential service times). For this model, the exact normalized mean workload is monotone with $c_z^2(0) = 2.0 > 1.5 = c_z^2(1)$, $\dot{c}_z^2(0) = -2.0$ and $\dot{c}_z^2(1) = -0.167$ (all exact), so that a monotone fit is desirable, but the cubic fit $(1.500 + 0.167\rho - 0.833\rho^2 + 1.367\rho^3)$ is not monotone. Moreover, the cubic fit does not constitute a reasonable fit compared to what we might do by eye.

Thus, we develop *fitting criteria* and find parameters so that (6) meets these criteria in every case. As a minimal condition, we require that f be nonnegative. Of course, this always can be achieved by considering $\max\{0, f(\rho)\}$, but that may make f nondifferentiable, which suggests a poor fit. Our approach is to look for a *nice* fit. Of course, *all* normalized mean workloads are *not* nice; see Section 6. However, we *assume* that for the model of interest, the normalized mean workload *is* relatively nice, subject to the constraints of the four specified values. By *nice* we mean smooth and monotone. We require that f be at least twice differentiable and that the second derivative f'' be continuous. The most important criterion we call the *low order derivative monotonicity criterion: First, the fitted function f should be monotone whenever possible and, second, the lowest order derivative that can be monotone should be.* Whether or not f can be monotone, we try to make f' monotone. If f' cannot be monotone, then we try to make f'' monotone, and so forth. In fact, we can make f'' monotone whenever f' is not monotone, so we need go no further. Indeed, f'' is monotone in all cases of Section 3 except in Case 1.2.3c and other cases that reduce to this case. (Since f' is monotone in this case, the low order derivative monotonicity criterion is satisfied.) In fact, for Case 1.2.3c, it can be shown that f'' cannot always be made monotone.

Of course, the smoothness and monotonicity criteria do not completely specify the function. We select (6), in part, based on the subjective criterion of simplicity; we want a relatively tractable expression, so that the fitting is easy to do and the function is relatively easy to apply, for example, in optimization applications where the best arrival or service rate is to be determined. For such optimization applications, it is significant that the parameters a_i and b_i in (6) are independent of ρ.

As a consequence of the low order derivative monotonicity criterion mentioned above, the fitted function must lie in a certain bounding polygon, which we describe in Section 4. This bounding polygon is useful to describe the reliability of the fit, given the fitting criteria.

Numerical Comparisons

In Section 5, we briefly describe the performance of our approximation by making numerical comparisons. Of particular interest are cases with highly variable interarrival-time distributions such as the hyperexponential distribution with $c_a^2 = 12$, as in Table II of Whitt (1984b). Then additional information about the interarrival-time distribution beyond the first two moments is essential to obtain a reasonable approximation of the mean workload.

The approximation developed here typically performs very well, but there are exceptional *pathological* cases. To put the approximation in perspective, we describe some of these in Sections 5 and 6.

Interpolation Approximations for General Single Server Queues

As mentioned at the outset, we are primarily motivated by the desire to develop an approximation for the mean workload in general single server queues, which may have dependence among successive

interarrival times, among successive service times, and between interarrival and service times. In Fendick and Whitt, we develop expressions (exact or approximate) for $c_z^2(0)$, $\dot{c}_z^2(0)$, $c_z^2(1)$ and $\dot{c}_z^2(1)$ in terms of a function that we call the *index of dispersion for work* (IDW). If $X(t)$ represents the total work in service time to enter the queue in the interval $[0, t]$, then the IDW is the function of time

$$I_w(t) = \frac{\text{Var}[X(t)]}{E(S)E[X(t)]}, \quad t \geq 0 \tag{7}$$

where $E(S)$ is the mean service time. The proposed approximations are

$$c_z^2(0) \approx I_w(0) = J_w(0)$$

$$\dot{c}_z^2(0) \approx \dot{I}_w(0)I_w(0) = \dot{J}_w(0)J_w(0)$$

$$c_z^2(1) \approx I_w(\infty) = J_w(1) \tag{8}$$

and

$$\dot{c}_z^2(1) \approx 2\dot{J}_w(1)/\max\{1, J_w(1)\}$$

where $J_w(\rho) = I_w(\rho/(1-\rho))$, $0 \leq \rho \leq 1$, and $I_w(t)$ is determined for the case $\rho = 1$. (See Fendick and Whitt for a discussion of scaling.) From a sample path of $X(t)$, we obtain the four values $J_w(0)$, $J'_w(0)$, $J_w(1)$ and $J'_w(1)$ from estimates of $I_w(t)$. For a large class of models, such as multiclass queues in which each class provides GI/G/1 input, we can also calculate $J_w(0)$, $J'_w(0)$, $J_w(1)$ and $J'_w(1)$ *analytically* in terms of basic model building blocks (see Fendick and Whitt, and Fendick, Saksena and Whitt 1989b), so that we can generate approximations for $c_z^2(\rho)$ *without measurements*. Obviously, the interpolation developed here applies equally well with (8). Moreover, it is significant that for the special case of the GI/G/1 model the general approximations in (8) are closely related to the GI/G/1 limits developed here. Indeed, all except $\dot{c}_z^2(1)$ coincide; the difference in $\dot{c}_z^2(1)$ is between the correction factors in (41) and (44). Thus, the GI/G/1 analysis here provides theoretical support for the approximate description of more complicated models.

1. LIGHT TRAFFIC

Brumelle's formula (3) is convenient for reducing the light traffic behavior of $E(Z\rho)$ and $c_z^2(\rho)$ to the light traffic behavior of $E(W_\rho)$; that is, from (1) and (3) it follows that

$$c_z^2(\rho) = (1-\rho)(1+c_s^2) + 2(1-\rho)E(W_\rho). \tag{9}$$

From (9) it is easy to see what $c_z^2(0)$ and $\dot{c}_z^2(0)$ should be. First, assuming that customers arrive one at a time, obviously $W_\rho \to 0$ w.p.1 and $E(W_\rho) \to 0$, so that we obtain the first relation in (2). (A formal proof for GI/G/1 follows from (2.3) on p. 185 of Asmussen (1987); there, $S_n^+ \to 0$ as $\rho \to 0$ for each n, and the moment conditions here imply the uniform integrability to get $E(S_n^+) \to 0$ as $\rho \to 0$ for each n.)

From (9), it is apparent that the derivative can be expressed as

$$\dot{c}_z^2(0) \equiv \lim_{\rho \to 0} \left[\frac{c_z^2(\rho) - c_z^2(0)}{\rho}\right]$$
$$= -(1 + c_s^2) + 2\lim_{\rho \to 0} \rho^{-1}E(W_\rho) \tag{10}$$

so that $\dot{c}_z^2(0)$ depends on the derivative of $E(W_\rho)$ at $\rho = 0$. As in Section 6.8 of Newell (1982), it is easy to see, heuristically, what this derivative should be. To compute this derivative, we only have to consider the interaction of two customers, that is, as $\rho \to 0$ we have the natural approximation

$$E(W_\rho) \approx E[(S - \rho^{-1}T)^+] \tag{11}$$

where $(x)^+ = \max\{x, 0\}$. (When the arrival process is Poisson or can be appropriately constructed in terms of a Poisson process, this step can be justified by the techniques of Section 1 of Whitt (1986), Reiman and Simon (1988, 1989) or Reiman and Weiss (1989).) To extract the light traffic limit from (11), we use the service time stationary-excess variable S_e, where

$$E(S)P(S_e > t)$$
$$\equiv \int_t^\infty P(S > y)dy = \int_0^\infty P(S > y + t)dy$$
$$= \int_0^\infty P(S - t > y)dy = E[(S-t)^+]. \tag{12}$$

Assume that T has a density $g(t)$ that is continuous at the origin. Then, under regularity conditions that allow us to move limits inside the integral

$$\lim_{\rho \to 0} 2\rho^{-1}E(W_\rho)$$
$$\approx \lim_{\rho \to 0} 2\rho^{-1}E[(S - \rho^{-1}T)^+]$$
$$= \lim_{\rho \to 0} 2\rho^{-1} \int_0^\infty E[(S-t)^+]\rho g(\rho t)dt \tag{13}$$
$$= 2g(0) \int_0^\infty E[(S-t)^+]dt$$
$$= 2g(0)(ES) \int_0^\infty P(S_e > t)dt$$
$$= 2g(0)E(S)E(S_e) = g(0)(c_s^2 + 1).$$

Combining (10) and (13), we see that

$$\dot{c}_z^2(0) = (1 + c_s^2)(g(0) - 1). \tag{14}$$

As indicated in Newell, and Fendick and Whitt, this heuristic argument extends to more general models.

For GI/G/1, (14) can be rigorously justified by applying the results by Daley and Rolski (1984, 1990). Let $G(t) = P(T \leq t)$. For general GI/G/1 models with our interarrival-time scaling, they conclude that

$$\lim_{\rho \to 0} \rho^{-\alpha} E(W_\rho) = E(S^{1+\alpha})\gamma/(1 + \alpha) \tag{15}$$

for $0 \leq \alpha < \infty$ and $0 \leq \gamma < \infty$ if $t^{-\alpha}G(t) \to \gamma$ as $t \to 0$ and $E(S^{2+\alpha}) < \infty$.

From (9) and (15), using $\alpha = 0$, it follows that if $G(0) = 0$ (T has no atom at 0), then indeed $E(W_\rho) \to 0$ as $\rho \to 0$, so that $c_z^2(0) = 1 + c_s^2$ as in (2). By considering the case $\alpha = 1$, we see that if $E(S^3) < \infty$ (as we have assumed), $G(0) = 0$ and $G(t)$ has a derivative $g(0)$ at 0, then

$$\lim_{\rho \to 0} \rho^{-1} E(W_\rho) = \frac{E(S^2)g(0)}{2} = \frac{(1 + c_s^2)g(0)}{2} \tag{16}$$

which is consistent with (13), so that we have rigorously justified (14).

Note that $g(0)$ is the density of the unscaled interarrival-time T at 0. If T has a density $g(t)$ for all t, then $\rho^{-1}T$ has density $g_\rho(t) = \rho g(\rho t)$, $t \geq 0$, so that $g(0) = \rho^{-1}g_\rho(0)$ in (13), (14) and (16). In applications, we consider $g_\rho(0)$ for some ρ; that is, (14) is equivalent to

$$\dot{c}_z^2(0) = (1 + c_s^2)\left(\frac{g_\rho(0)}{\rho} - 1\right). \tag{17}$$

Finally, from (14) or (17) note that

$$\dot{c}_z^2(0) \geq -c_z^2(0) \tag{18}$$

because $g(0) \geq 0$.

From this analysis, we see that the mean workload is substantially more robust in light traffic than the mean waiting time, so that it should be easier to obtain fairly good light traffic approximations for the mean workload without considering the fine structure of the model, that is, without considering more than c_s^2 and $g(0)$. On the other hand, we can expect to obtain better approximations, especially for the mean waiting time, by exploiting (15) with the precise α yielding a nondegenerate limit.

In applications, the value $\rho^{-1}g_\rho(0)$ in (17) may be unavailable. As a default value, we suggest

$$g(0) \equiv \rho^{-1}g_\rho(0) = \begin{cases} 2c_a^2/(c_a^2 + 1) & c_a^2 > 1 \\ (c_a^2)^4 & c_a^2 \leq 1. \end{cases} \tag{19}$$

The default value when $c_a^2 > 1$ is the exact value for an H_2 distribution (hyperexponential, a mixture of two exponentials) with balanced means, having a density

$$g(t) = p\mu_1 e^{-\mu_1 t} + (1 - p)\mu_2 e^{-\mu_2 t} \quad t \geq 0 \tag{20}$$

where $p/\mu_1 = (1 - p)/\mu_2 = 1/2$. For $c_a^2 < 1$, the default value is based on the observation that $g(0)$ is often small when $c_a^2 < 1$; for example, for all E_k (Erlang) distributions, for which $c_a^2 = k^{-1}$, $g(0) = 0$, whereas $g(0) = 1$ for an exponential distribution. (The exponent 4 in (19) is rather arbitrary.)

2. HEAVY TRAFFIC

The heavy traffic limit in (2) is well known, following from Kingman's (1962) seminal paper; see p. 196 of Asmussen (1987). (The third moment condition here implies the uniform integrability assumed there in order to get convergence of the moments. Apply (3) to go from $E(W_\rho)$ to $E(Z_\rho)$.) The extensive heavy traffic literature provides a basis for determining $c_z^2(1)$ in much more general models; a general characterization is given in (8). See Fendick, Saksena and Whitt (1989a, b) and Fendick and Whitt for further discussion.

In contrast, the heavy traffic derivative $c_z^2(1)$ is not so well understood, although there has been a surprising amount of work related to the GI/G/1 special case, as can be seen from Siegmund (1979), Köllerström (1981), Chapter X of Siegmund (1985), Halfin (1985), Chapter XII of Asmussen (1987) and Knessl (1990).

It is actually easy to see what the derivative $\dot{c}_z^2(1)$ should be by applying Marshall's formula

$$E(W_\rho) = \frac{E(\rho^{-1}T - S)^2}{2E(\rho^{-1}T - S)} - \frac{E(I_\rho^2)}{2E(I_\rho)} \quad 0 < \rho < 1 \tag{21}$$

where I_ρ is the idle period (see p. 76 of Stoyan 1983, Halfin or (2.4) on p. 185 of Asmussen 1987) so that

$$E(W_\rho) = \frac{c_a^2 + c_s^2}{2(1 - \rho)} + \frac{\rho^{-1}c_a^2 - c_s^2}{2} + \frac{(1 - \rho)}{2\rho} - \frac{E(I_\rho^2)}{2E(I_\rho)}. \tag{22}$$

Moreover, an extension of (21) exists for general single server queues without any independence conditions, see (5.0.21) of Stoyan, but we will have our hands full with the standard GI/G/1 model.

Unfortunately, even for GI/G/1 the idle period I_ρ is rather complicated, so that (22) is not immediately informative. However, upon reflection, it is intuitively clear that unlike W_ρ, I_ρ usually converges to a proper limit I_1 as $\rho \to 1$ with $E(I_\rho^k) \to E(I_1^k)$ for $k = 1$ and 2. For example, for any M/G/1 queue, I_ρ has the same distribution as an interarrival-time $\rho^{-1}T$, which converges to T as $\rho \to 1$. From basic stochastic comparison theory, we can conclude that the ratio $E(I_\rho^2)/2E(I_\rho)$ is bounded below by a term of the order $1 - \rho$; that is, since

$$E(I_\rho^2)/2E(I_\rho) \geq (1 - \rho)/2\rho \tag{23}$$

(see (5.5.4) of Stoyan), we have the classical Kingman bound, expressed as a one-sided expansion in powers of $(1 - \rho)$

$$E(W_\rho) \leq \frac{c_a^2 + c_s^2}{2(1 - \rho)} + \frac{c_a^2 - c_s^2}{2} + \frac{(1 - \rho)c_a^2}{2\rho} \tag{24}$$

(see (5.6.2) of Stoyan). However, we have no tight bound on the other side, so (24) does not nearly resolve the issue.

Here we simply *assume* that $E(I_\rho^2)/2E(I_\rho) \to E(I_1^2)/2E(I_1)$ as $\rho \to 1$. Under this assumption, we can write

$$E(W_\rho) = \frac{c_a^2 + c_s^2}{2(1 - \rho)} + \frac{c_a^2 - c_s^2}{2} - \frac{E(I_1^2)}{2E(I_1)} + o(1) \quad \text{as } \rho \to 1 \tag{25}$$

where $o(h(\rho))$ is a quantity that converges to 0 as $\rho \to 1$ after being divided by $h(\rho)$. Combining (1), (3) and (25), we obtain

$$c_z^2(\rho) = (c_a^2 + c_s^2) - (1 - \rho)\left(\frac{E(I_1^2)}{E(I_1)} - (c_a^2 + 1)\right) + o(1 - \rho) \quad \text{as } \rho \to 1 \tag{26}$$

so that $c_z^2(1) = c_a^2 + c_s^2$ as in (2) and

$$\dot{c}_z^2(1) = \left(\frac{E(I_1^2)}{E(I_1)} - (c_a^2 + 1)\right). \tag{27}$$

Two difficulties remain: First, we need to know when (25) is justified (presumably a technical matter of little practical concern) and, second, we need to express $E(I_1^2)/2E(I_1)$ in terms of the distributions of T and S (obviously a serious issue of great practical concern). In fact, both issues have been investigated fairly extensively, but neither is completely resolved. In support of (25), Köllerström proved that under the condition of finite fourth moments, $E(I_\rho^2)/E(I_\rho)$ is *bounded* as $\rho \to 1$; see Lemma 2, (60) and (63) there. (It appears that only a finite third moment is needed for this part of his analysis. The idle period I_ρ coincides with the descending ladder height H_1^- there.) Moreover, Halfin, who also starts from Marshall's formula (2.1), shows that for any GI/M/1 queue I_ρ converges in distribution to I_1, where, in that case, I_1 has the stationary-excess distribution associated with the interarrival-time T, that is,

$$P(I_1 > t) = (ET)^{-1} \int_t^\infty P(T > s)ds \quad t \geq 0 \tag{28}$$

so that the mystery term in (25) and (27) has the very simple form

$$\lim_{\rho \to 1} \frac{E(I_\rho^2)}{2E(I_\rho)} = \frac{E(I_1^2)}{2E(I_1)} = \frac{E(T^3)}{3E(T^2)}. \tag{29}$$

Further support for (25) and extensions are provided by Siegmund (1979), which is further discussed in Chapter X of Siegmund (1985) and Chapter XII of Asmussen (1987). Siegmund uses a change of measure argument (not unlike Reiman and Weiss) in the context of conjugate distributions (additional assumptions on the distribution of $(S - \rho^{-1}T)$) to establish the expansion

$$E(W_\rho) = \Delta_\rho^{-1} - \frac{E(I_1^2)}{2E(I_1)} + \frac{\Delta_\rho}{2}\left(\frac{E(I_1^3)}{3E(I_1)} - \left[\frac{E(I_1^2)}{2E(I_1)}\right]^2\right) + o(\Delta_\rho) \quad \text{as } \rho \to 1 \tag{30}$$

where $\Delta_\rho \to 0$ as $\rho \to 1$. Since the convex function ψ there is asymptotically quadratic as $\rho \to 1$ (Taylor series), it is possible to show that (30) is in fact consistent with (25). Of course, (30) also provides an additional third order term.

As indicated above, the pressing applied problem is to characterize $E(I_1^2)/2E(I_1)$. In fact, much is known about the idle period I_ρ from the general theory of random walks. For any ρ, $-I_\rho$ coincides with the weak descending ladder height of the random walk with steps $(S - \rho^{-1}T)$; see p. 182 of Asmussen (1987) or p. 53 of Prabhu (1980). The Laplace transform of I_ρ for any ρ can be obtained from the Wiener-Hopf theory, and is given in general in (5.61) on p. 284 and (5.136) on p. 304 of Cohen (1982), and for a large class of special cases in (5.194) on p. 325 and (5.206) on p. 331. Moreover, conditions for the moments $E(I_\rho^k)$ to be finite have been determined (with some ingenuity for the case $\rho = 1$). Spitzer (1960) showed that if

$0 < E(T - S)^2 < \infty$, then

$$0 < E(I_1) = \frac{E(S-T)^2}{\sqrt{2}} \exp\left\{\sum_{n=1}^{\infty} n^{-1}\left[\frac{1}{2} - P(U_n \leq 0)\right]\right\} < \infty \tag{31}$$

where

$$U_n = \sum_{i=1}^{n} (S_i - T_i), \quad n \geq 1 \tag{32}$$

(see p. 282 of Chung 1974 or p. 199 of Spitzer 1964). By a similar argument (Problem 6 on p. 232 of Spitzer 1964), $E(I_1^2) < \infty$ provides that $E(|S - T|^3) < \infty$. Hence, under our assumptions

$$0 < \frac{E(I_1^2)}{E(I_1)} < \infty \tag{33}$$

so that the right side of (27) is well defined.

Higher moments $E(I_1^k)$ were computed by Lai (1976), who also showed that under finite third moments

$$0 < \frac{E(I_1^2)}{2E(I_1)} = \frac{E[(T-S)^3]}{6} + \frac{1}{2}\sum_{n=0}^{\infty}\left[\sqrt{2}\binom{-1/2}{n}(-1)^n - n^{-1}E(|U_n|)\right] < \infty \tag{34}$$

where U_n is given by (32) and $n^{-1}E(|U_n|) = 0$ for $n = 0$. For numerical calculations, an integral representation is also available (p. 225 of Siegmund 1985 or p. 276 of Asmussen 1987) in terms of the characteristic function $\phi(t) = E[e^{it(T-S)}]$, namely

$$\frac{E(I_1^2)}{2E(I_1)} = \frac{E[(T-S)^3]}{6} - \frac{1}{\pi}\int_0^{\infty} t^{-2}\,\mathrm{Re}\log\{2[1-\phi(t)]/t^2\}dt. \tag{35}$$

In summary, even though there remains a technical gap ((30) does not cover all GI/G/1 queues), it seems reasonable to treat (27) plus (34) or (35) as the correct formula. For practical purposes, the real difficulty is that, in general, the ratio $E(I_1^2)/E(I_1)$ that appears in (27) has no simple expression that depends on only a few parameters of the interarrival-time and service-time distributions. (It is easy to show that (29) does not hold for all GI/G/1 queues.) Indeed, it is remarkable that the simple formula (29) emerges in the GI/M/1 case.

An extremely important insight from this analysis, applying with even more force than Daley and Rolski's observation about the light traffic limit (15), is the fact that a simple exact expression for $\dot{c}_z^2(1)$ for any GI/G/1 queue is evidently *not possible*, so that additional approximation is necessary in order to obtain the kind of elementary formula we want. This insight is confirmed by Knessl who identifies terms in an expansion for the distribution of Z_ρ in powers of $(1 - \rho)$, assuming that such an expansion is valid, using the theory of matched asymptotic expansions applied to the forward Kolmogorov (Takács) equation. Knessl's analysis is very different from the other approaches, all of which are based on random walk theory. It seems to have the potential for making the higher order heavy traffic terms more numerically accessible, but for general GI/G/1 systems the second order term seems to require the solution of a Wiener-Hopf problem.

Köllerström suggests *simulating* I_1 to estimate $E(I_1^2)/2E(I_1)$. This closely parallel's Minh and Sorli's (1983) method for estimating $E(W_\rho)$ by simulating $E(I_\rho^2)/2E(I_\rho)$ and applying (21). Simulation of I_1 is not straightforward, however, because the mean length of the busy cycle containing the idle period is infinite when $\rho = 1$. As an approximation, one might use I_ρ for $\rho = 0.9$, but long busy cycles remain a problem. For practical purposes, however, it appears that little will be lost by simulating successive busy cycles to estimate $E(I_1^k)$ and terminating all runs that fail to complete a busy cycle by a prescribed time. We, thus, fail to sample in a region with small probability, but this does not seem too harmful because the conditional expectations of I_1^k given a long busy period are not too large.

We believe that useful approximations for $c_z^2(\rho)$ can be obtained by interpolating using *approximations* for the derivative $\dot{c}_z^2(1)$. This idea is supported by Halfin, who suggests converting (21) into an approximation by approximating the idle period distribution by the stationary-excess interarrival-time distribution, so that we obtain

$$\frac{E(I_\rho^2)}{2E(I_\rho)} \approx \frac{E(\rho^{-1}T)^2}{3E(\rho^{-1}T^2)} = \frac{E(T^3)}{3\rho E(T^2)} \tag{36}$$

which is, of course, supported by (29). Based on (21) and (36), Halfin proposed the general approximation

$$E(W_\rho) \approx \frac{E(\rho^{-1}T - S)^2}{2E(\rho^{-1}T - S)} - \frac{E(T^3)}{3\rho E(T^2)} \tag{37}$$

which he notes is not good (e.g., negative) when $E(T^3)$ becomes large. (He also develops refinements.) Of course, this same difficulty applies to (30), (34) and (35), even with three terms. To avoid this difficulty, we suggest applying (36) only to estimate the *derivative* $\dot{c}_z^2(1)$. In particular, we suggest as an initial approximation to (27)

$$\dot{c}_z^2(1) \approx \left(\frac{2E(T^3)}{3E(T^2)} - (c_a^2 + 1)\right) = \left(\frac{2E(T^3)}{3(c_a^2 + 1)} - (c_a^2 + 1)\right). \tag{38}$$

We also suggest a way to refine (38), which should have many other applications. We suggest using numerical values for high traffic intensities to estimate $\dot{c}_z^2(1)$. In particular, we refined (38) by using numerical estimates of the derivative $\dot{c}_z^2(1)$ based on algorithmically generated values in the tables of Seelen, Tijms and van Hoorn (1985). (In fact $E(Z_\rho)$ does not appear in the tables; it is obtained from the mean queue length there via $L = \lambda W$ and (3).) The tabled values for the case $\rho = 0.98$ were used to estimate $\dot{c}_z^2(1)$ via

$$\dot{c}_z^2(1) \approx \frac{c_z^2(1) - c_z^2(0.98)}{0.02} \tag{39}$$

which should be pretty reliable because $(1 - \rho)^2 = 0.0004 \ll 0.02 = (1 - \rho)$; we can safely neglect higher order terms. Based on the tabled values for $\rho = 0.98$, we propose the approximation

$$\dot{c}_z^2(1) \approx X(c_a^2, c_s^2)\left(\frac{2E(T^3)}{3(c_a^2 + 1)} - (c_a^2 + 1)\right) \tag{40}$$

where

$$X(c_a^2, c_s^2) = \begin{cases} 1 + \dfrac{(1 - c_s^2) + 2(1 - c_s^2)}{4} & c_a^2 < 1,\ c_s^2 \leqslant 1 \\ 1 + \dfrac{(1 - c_s^2)}{4} & c_a^2 > 1,\ c_s^2 \leqslant 1 \\ 1 & c_a^2 < 1,\ c_s^2 > 1 \\ 1 - \dfrac{1}{2}\left(\dfrac{c_s^2 - 1}{c_s^2 + 0.5}\right)^2 & c_a^2 > 1,\ c_s^2 > 1. \end{cases} \tag{41}$$

If the third moment $E(T^3)$ is unavailable, then we suggest the default value

$$E(T^3) \equiv \frac{E(T^3)}{(ET)^3} = \begin{cases} 3c_a^2(c_a^2 + 1) & c_a^2 > 1 \\ (c_a^2 + 1)(2c_a^2 + 1) & c_a^2 \leqslant 1 \end{cases} \tag{42}$$

as on p. 2804 of Whitt (1983), which is based on the exact results for the H_2 distribution with balanced means in (20) when $c_a^2 > 1$ and the E_k distribution when $c_a^2 \leqslant 1$.

Note that (41) yields $X(c_a^2, 1) = 1$, so that (40) agrees with (38) for GI/M/1. The adjustment factor $X(c_a^2, c_s^2)$ in (40) has the disadvantage of not being continuous in c_a^2 at 1, but in the case of an exponential interarrival-time distribution with $c_a^2 = 1$, $E(T^3) = 6 = 3(c_a^2 + 1)^2/2$, so that the other term in (40) becomes 0; i.e., for an M/G/1 queue, $\dot{c}_z^2(1) = 0$.

Since the tables in Seelen et al. only include distributions characterized by their first two moments, only c_a^2 and c_s^2 could reasonably enter into (41). Our idea is that the third interarrival-time moment should be pretty well captured by (38) and that the third service-time moment should not matter too much for $E(W_\rho)$ and $E(Z_\rho)$. There is certainly room for further refinement, but even (38) seems to be pretty good.

The first general approximation for $\dot{c}_z^2(1)$ developed in Fendick and Whitt was $\dot{c}_z^2(1) \approx 2J_w'(1)/J_w(1)$, as in (8) without the maximum adjustment in the denominator. For GI/G/1, this initial approximation reduces to (40) with

$$X(c_a^2, c_s^2) = \frac{c_a^2 + 1}{c_a^2 + c_s^2} \tag{43}$$

instead of (41). (The general approximation for $\dot{c}_z^2(0)$ agrees with (17).) From numerical comparisons (e.g., Section 5), we find that (40) performs well for typical distributions with all three adjustment factors (41), (42) and $X = 1$ as in (38)). Indeed, even though X can vary significantly, the values of $c_z^2(\rho)$ do not change dramatically. Evidently $c_z^2(\rho)$ is not very sensitive to changes in $\dot{c}_z^2(1)$; for example, a 50% change in X produces a 50% change in $\dot{c}_z^2(1)$, but typically less than a 5% change in $c_z^2(\rho)$ for any ρ. This robustness obviously provides strong support for the general procedure (6) and (8). The robustness does not mean that $\dot{c}_z^2(1)$ does not matter at all; it obviously matters whether $\dot{c}_z^2(1)$ is much less than, nearly the same as, or much greater than $c_z^2(1) - c_z^2(0)$.

Even though the three adjustments do not differ greatly, (41) seems best overall. For example, for the E_2/D/1 model, (38) and (41) produce maximum (over all ρ) relative errors of about 5.0% and 1.7%, respectively. Adjustment (43) is similar to (41) except when $c_a^2 + c_s^2$ is very small, as in E_{10}/D/1; then (41) is much better. This suggests that (43) be replaced by

$$X(c_a^2, c_s^2) = \frac{c_a^2 + 1}{\max\{1, c_a^2 + c_s^2\}} \tag{44}$$

which led us to introduce the maximum adjustment in the denominator of $\dot{c}_z^2(1)$ in (8).

3. THE INTERPOLATION FUNCTION

In this section, we determine the coefficients a_i and exponents b_i of the function $f(\rho)$ in (6) in order to match the four given values $f(0), f'(0), f(1)$ and $f'(1)$, and satisfy the positivity and low order derivative monotonicity criteria. From (6), we see that

$$f(0) = a_0 + a_4$$

$$f(1) = a_0 + a_1 + a_2 + a_3$$

$$f'(0) = a_1 - a_4 b_4 \tag{45}$$

and

$$f'(1) = a_1 + 2a_2 + a_3 b_3$$

provided that $b_3 \neq 1$ and $b_4 \neq 1$. The final parameters are chosen to satisfy the regularity criteria. By (2), we can assume that $f(0)$ and $f(1)$ are strictly positive and finite, but we make no such assumption for the derivatives. If $f'(0)$ or $f'(1)$ is $+\infty(-\infty)$, then we replace it with a large positive (negative) value.

We consider three cases: 1) $f(0) < f(1)$, 2) $f(0) > f(1)$ and 3) $f(0) = f(1)$, each with several subcases.

Case 1. $f(0) < f(1)$

We try to make f strictly increasing; we get two subcases depending on whether or not this is possible.

Case 1.1. Increasing fit: $f'(0) \geq 0$ and $f'(1) \geq 0$

In this case, we can make f strictly increasing, and do. Next, we try to make the derivative f' constant and, if that is not possible, monotone. In all five subcases f'' is monotone.

Case 1.1.1. Linear fit: $f'(0) = f'(1) = [f(1) - f(0)]$

Let $a_0 = f(0)$, $a_1 = f(1) - f(0)$, and $a_2 = a_3 = a_4 = 0$.

Case 1.1.2. Convex fit: $0 \leq f'(0) < [f(1) - f(0)] < f'(1)$

In this case we make f and f' increasing, and all higher derivatives monotone. Let

$$a_0 = f(0), \quad a_1 = f'(0), \quad a_2 = a_4 = 0$$

$$a_3 = f(1) - f(0) - f'(0) > 0$$

and (46)

$$b_3 = \frac{f'(1) - f'(0)}{a_3} > 1.$$

Case 1.1.3. Concave fit: $0 \leq f'(1) < [f(1) - f(0)] < f'(0)$

We use the same fit as in Case 1.1.2; here $a_3 < 0$ and $b_3 > 1$, so that f' is decreasing instead of increasing.

Case 1.1.4. Increasing f'' fit: $f'(0) \geq [f(1) - f(0)]$ and $f'(1) \geq [f(1) - f(0)]$

The case of two equalities is covered by Case 1.1.1, so we assume that one inequality is strict. In this case, we cannot make f' monotone, but we make both f and f'' increasing. Let $a_1 = a_2 = 0$

$$a_0 = \frac{f(0) + f(1)}{2}$$

$$a_3 = \frac{f(1) - f(0)}{2} = -a_4$$

$$b_3 = \frac{2f'(1)}{f(1) - f(0)} \geq 2 \tag{47}$$

and

$$b_4 = \frac{2f'(0)}{f(1) - f(0)} \geq 2.$$

Case 1.1.5. Decreasing f'' fit: $f'(0) \leq [f(1) - f(0)]$ and $f'(1) \leq [f(1) - f(0)]$

Again, the case of two equalities is covered by Case 1.1.1, so we assume that one inequality is strict. As in Case 1.1.4, we cannot make f' monotone, but we make both f and f'' monotone. In this case, we do a simple cubic fit, i.e., let $b_3 = 3$ and $a_4 = 0$. In addition, let $a_0 = f(0)$, $a_1 = f'(0)$

$$a_2 = -f'(1) - 2f'(0) + 3[f(1) - f(0)] > 0 \tag{48}$$
$$a_3 = f'(1) + f'(0) - 2[f(1) - f(0)] < 0.$$

Since $a_2 \geq -a_3$, f is indeed increasing.

Case 1.2. Nonmonotone fit: $f'(0) < 0$ or $f'(1) < 0$

We cannot fit a monotone f, because $f'(0) < 0$ or $f'(1) < 0$. We try to make f' monotone and, if not f', then f''. We make f'' monotone in every subcase, except Case 1.2.3c below, where f' is monotone. When $f'(0) < 0$, we have $f(\rho) < f(0)$ for some ρ, so that we must also ensure that $f(\rho) > 0$ for all ρ.

Case 1.2.1 Concave fit: $f'(1) < 0 < [f(1) - f(0)] < f'(0)$

Let $a_4 = 0$, $a_0 = f(0)$ and $a_1 = f'(0)$.

a. If $[f(1) - f(0)] < f'(0) < 2[f(1) - f(0)]$, then let

$$a_2 = a_3 = \frac{-(f'(0) - [f(1) - f(0)])}{2}$$

and (49)

$$b_3 = \frac{f(1) - f(0) - f'(1)}{-a_3} \geqslant 1.$$

b. If $f'(0) > 2[f(1) - f(0)]$, then let

$$a_2 = f(0) - f(1)$$

$$a_3 = -(f'(0) - 2[f(1) - f(0)]) \quad (50)$$

and

$$b_3 = \frac{a_3 + f'(1)}{a_3} > 1.$$

Case 1.2.2. Decreasing f'' fit: $f'(1) < 0 \leqslant f'(0) \leqslant [f(1) - f(0)]$

Use the cubic fit in Case 1.1.5.

Case 1.2.3. Positive convex fit: $f'(0) < 0 < [f(1) - f(0)] < f'(1)$

We consider three subcases.

a. If $-f'(0) \leqslant 2f(0)$ and $-f'(0) < 2(f'(1) - [f(1) - f(0)])$, then let $a_0 = f(0)$, $a_1 = f'(0)$, $a_2 = -f'(0)/2$, $a_3 = f(1) - f(0) - f'(0)/2$, $a_4 = 0$ and

$$b_3 = \frac{2f'(1)}{2[f(1) - f(0)] - f'(0)} > 1. \quad (51)$$

Note that $a_1 < 0$, but

$$(a_0 + a_1\rho + a_2\rho^2) = f(0) + f'(0)\rho - \frac{f'(0)}{2}\rho^2 \geqslant \frac{-f'(0)}{2}(1 - \rho)^2. \quad (52)$$

b. If $f(0) \geqslant -f'(0)/2 \geqslant f'(1) - [f(1) - f(0)]$, then let $a_0 = f(0) - a_4$, $a_1 = a_2 = 0$, $a_3 = f(1) - f(0) + a_4$

$$b_3 = \frac{f'(1)}{f(1) - f(0) + a_4}, \quad b_4 = \frac{-f'(0)}{a_4}$$
$$a_4 = f'(1) - [f(1) - f(0)] - z > 0 \quad (53)$$

where

$$z = \min\left\{\frac{f'(1) - [f(1) - f(0)]}{2}, 0.48f'(1)\right\} > 0. \quad (54)$$

Since $1 < b_3 < 2$ and $b_4 \geqslant 2$, both f' and f'' are monotone.

c. The remaining subcase occurs when $-f'(0) > 2f(0)$. This subcase cannot arise directly in our application, because $f'(0) \geqslant -f(0)$, as noted in (18), but it can arise indirectly via Cases 2 and 3.3. For this case, proceed just as in b, except let

$$a_4 = \min\{f'(1) - [f(1) - f(0)] - z, f(0)\} \quad (55)$$

which implies that $b_3 > 1$ and $b_4 > 2$. When $b_3 > 2$ and $b_4 > 2$, f'' need not be monotone. Indeed, for this subcase we cannot guarantee that f'' is monotone.

Case 1.2.4. Positive with decreasing f'' fit: $f'(0) < 0 \leqslant f'(1) \leqslant [f(1) - f(0)]$

Let $a_0 = a_3 = 0$, $a_1 = 2f(1) - f'(1)$, $a_2 = f'(1) - f(1)$, $a_4 = f(0)$ and

$$b_4 = [2f(1) - f'(1) - f'(0)]/f(0) > 2. \quad (56)$$

Case 1.2.5. Positive with two-zero f' fit: $f'(0) < 0$ and $f'(1) < 0$

Not only cannot f and f' be monotone, but f' must have two zeros in the interval (0, 1). Let f be defined as in Case 1.2.4.

Case 2. $f(0) > f(1)$

We reduce this to Case 1 by considering the time reversed function

$$h(\rho) = f(1 - \rho), \quad 0 \leqslant \rho \leqslant 1. \quad (57)$$

Obviously, $h(0) = f(1) < f(0) = h(1)$, so that h is covered by Case 1, with $h'(0) = -f'(1)$ and $h'(1) = -f'(0)$. Applying Case 1 to h, we obtain

$$h(\rho) = \hat{a}_0 + \hat{a}_1\rho + \hat{a}_2\rho^2 + \hat{a}_3\rho^{\hat{b}_3} + \hat{a}_4(1 - \rho)^{\hat{b}_4} \quad (58)$$

so that

$$f(\rho) = h(1 - \rho) = \hat{a}_0 + \hat{a}_1(1 - \rho) + \hat{a}_2(1 - \rho)^2 + \hat{a}_3(1 - \rho)^{\hat{b}_3} + \hat{a}_4\rho^{\hat{b}_4} \quad (59)$$

which is transformed into (6) by setting

$$a_0 = \hat{a}_0 + \hat{a}_1 + \hat{a}_2$$
$$a_1 = -(\hat{a}_1 + 2\hat{a}_2), \quad a_2 = \hat{a}_2$$
$$a_3 = \hat{a}_4, \quad a_4 = \hat{a}_3 \quad (60)$$
$$b_3 = \hat{b}_4$$

and

$$b_4 = \hat{b}_3.$$

Case 3. $f(0) = f(1)$.

This case is not really needed, because we could perturb $f(0)$ or $f(1)$ slightly to obtain Case 1 or 2, but if $f'(0)$ and $f'(1)$ are nearly equal, it may be preferable to treat them as equal and apply this case.

Case 3.1. Constant fit: $f'(0) = f'(1) = 0$.

Let $a_0 = f(0)$ and $a_1 = a_2 = a_3 = a_4 = 0$.

Case 3.2. Concave fit: $f'(1) < 0 < f'(0)$

Let

$$a_0 = f(0) - a_4, \quad a_1 = a_2 = 0$$

$$a_3 = a_4 = -\min\{f'(0), -f'(1)\}$$

$$b_3 = \frac{f'(1)}{a_3} \geq 2 \tag{61}$$

and

$$b_4 = \frac{-f'(0)}{a_3} \geq 2.$$

Case 3.3. Positive convex fit: $f'(0) < 0 < f'(1)$

We apply Case 1.2.3 to the function

$$h(\rho) = f(\rho) + c\rho - \frac{f(0)}{2}, \quad 0 < \rho < 1 \tag{62}$$

where $2c = \min\{f(0), -f'(0)\}$. Since

$$h(0) = \frac{f(0)}{2} < \frac{f(0)}{2} + c = h(1)$$

$$h'(0) = f'(0) + c < 0 < h(1) - h(0) \tag{63}$$
$$= c < f'(1) + c = h'(1)$$

h belongs to Case 1.2.3. Given $h(\rho)$ as in (58) based on Case 1.2.3, we obtain $f(\rho)$ in (6) by letting $a_0 = \hat{a}_0 + f(0)/2$, $a_1 = \hat{a}_1 - c$, and $a_i = \hat{a}_i$ plus $b_i = \hat{b}_i$ for $i \geq 2$.

Case 3.4. Positive with decreasing f'' fit: $f'(0) \leq 0$ and $f'(1) \leq 0$

Apply Case 1.2.4.

Case 3.5. Positive with increasing f'' fit: $f'(0) > 0$ and $f'(1) > 0$

Paralleling Case 2, we obtain this from Case 3.4 by considering the time reversed function h in (57). We apply Case 3.4, and thus Case 1.2.4, to determine h in (58). We obtain f itself from (59) and (60).

4. BOUNDING POLYGONS

Through the low order derivative monotonicity criterion, we have ensured that f or f' is concave or convex in every case of Section 3. As a consequence, we conclude that in every case the fitted function lies in a bounded polygon determined by (some of) the six lines

$$\begin{aligned} &f_1(\rho) = 0, \quad f_2(\rho) = f(0) \\ &f_3(\rho) = f(1), \quad f_4(\rho) = f(0) + \rho f'(0) \\ &f_5(\rho) = f(1) - (1 - \rho)f'(1), \\ &f_6(\rho) = f(0) + \rho[f(1) - f(0)], \quad 0 \leq \rho \leq 1 \end{aligned} \tag{64}$$

and the two vertical lines that correspond to $\rho = 0$ and $\rho = 1$. Table I shows the line indices that form edges of the polygon in each case. Figure 1 depicts the bounding polygons for each of the ten subcases of Case 1.

The bounding polygon gives some idea about the range of possible fits consistent with the fitting data. Of course, we have no guarantee that an unknown function lies in the bounded polygon, but it will if it satisfies the fitting criteria. Thus, in graphical displays, it is helpful to display the bounding polygon as well as the fitted function.

In Section 3 we have specified a fit for every case, but the cases differ greatly with regard to the *reliability of the fit*. For example, suppose that $f(1) = f(0) + 1$. If $f'(0) = 2$ and $f'(1) = \frac{1}{2}$ (Case 1.1.3), then the concave increasing fit is fairly reliable. However, if

Table I
Indices of the Lines in (64) That Form Edges of the Bounding Polygon in Each Case of the Interpolation (Depicted in Figure 1)

Case	Line Indices
1.1.1	6 (= 4 = 5)
1.1.2	4, 5, 6
1.1.3	4, 5, 6
1.1.4	2, 3, 4, 5
1.1.5	4, 5
1.2.1	4, 5, 6
1.2.2	4, 5
1.2.3	1, 4, 5, 6
1.2.4	1, 4, 5
1.2.5	1, 4, 5
3.1	6 (= 2 = 3 = 4 = 5)
3.2	4, 5, 6
3.3	1, 4, 5, 6
3.4	1, 4, 5
3.5	1, 4, 5

instead $f'(0) = f'(1) = 0.01$ (Case 1.1.5), then the monotone fit with nonmonotone derivative is less reliable. If $f'(0) < 0$ and $f'(1) < 0$ (Case 1.2.5), then the resulting nonmonotone fit will be even less reliable.

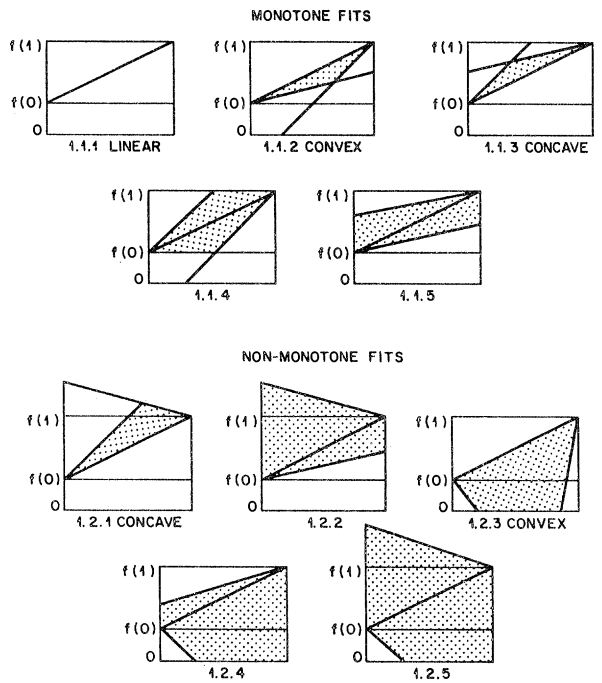

Figure 1. The bounding polygons for the ten subcases of Case 1.

To provide some guidance about the reliability of the fit we suggest two *fitting reliability measures.* The first is the area of the bounding polygon and the second is the length of the ρ-section of the bounding polygon as a function of ρ; i.e., if $\mathscr{P}$ is the polygon in R^2, then the ρ-section of $\mathscr{P}$ is

$$L(\rho) = \{f(\rho):(\rho, f(\rho)) \in \mathscr{P}\} \quad 0 \leq \rho \leq 1. \tag{65}$$

5. NUMERICAL COMPARISONS

In this section, we present a few numerical comparisons between the approximation for the normalized mean workload $c_z^2(\rho)$ in (1) and exact values. The approximation is (6) with $c_z^2(0)$ and $c_z^2(1)$ in (2), $\dot{c}_z^2(0)$ in (17), and $\dot{c}_z^2(1)$ in (40) with the adjustment factor $X(c_a^2, c_s^2)$ in (41). First, comparisons for $E_2/G/1$ and $E_{10}/G/1$ queues are displayed in Tables II and III. Four different service-time distributions are considered: D, E_k, M and H_2 with balanced means as in (20). The exact values come from Seelen et al. Their exact values for the mean queue length (excluding the customer in service) are transformed into $c_z^2(\rho)$ via $L = \lambda W$, (1) and (3). From Tables II and III, it is evident that the accuracy of the approximations in these cases is very good (beyond what is typically needed in practice), even though neither the light traffic nor the heavy traffic behavior is pinned down precisely. In light traffic, $g(0) = 0$, so that we do not fully exploit (15).

Table II

A Comparison of Approximations for the Normalized Mean Workload $c_z^2(\rho)$ in (1) With Exact Values for Four $E_2/G/1$ Queues from Seelen, Tijms and van Hoorn. (The H_2 Distribution has Balanced Means as in (20).)

Traffic Intensity	Service-Time Distribution							
	D, $c_s^2 = 0.0$		E_2, $c_s^2 = 0.5$		M, $c_s^2 = 1.0$		H_2, $c_s^2 = 4.0$	
ρ	Exact	Approx.	Exact	Approx.	Exact	Approx.	Exact	Approx.
1.00	0.500	0.500	1.000	1.000	1.500	1.500	4.500	4.500
0.98	0.506	0.506	1.004	1.004	1.503	1.503	4.504	4.503
0.95	0.514	0.515	1.011	1.010	1.508	1.508	4.507	4.508
0.90	0.529	0.530	1.022	1.021	1.518	1.517	4.513	4.517
0.80	0.561	0.559	1.046	1.042	1.538	1.533	4.531	4.533
0.70	0.596	0.591	1.073	1.064	1.560	1.550	4.549	4.550
0.60	0.634	0.626	1.105	1.088	1.587	1.569	4.570	4.567
0.50	0.678	0.666	1.140	1.118	1.618	1.591	4.594	4.583
0.40	0.724	0.712	1.183	1.155	1.657	1.620	4.623	4.600
0.30	0.777	0.766	1.233	1.206	1.703	1.664	4.660	4.619
0.20	0.838	0.831	1.297	1.275	1.766	1.731	4.710	4.646
0.10	0.911	0.908	1.381	1.374	1.854	1.837	4.788	4.722
0.00	1.000	1.000	1.500	1.500	2.000	2.000	5.000	5.000
Max (over ρ) % error	—	1.8	—	2.4	—	2.3	—	1.4

Table III

A Comparison of Approximations for the Normalized Mean Workload $c_z^2(\rho)$ in (1) With Exact Values in Four $E_{10}/G/1$ Queues from Seelen, Tijms and van Hoorn. (The H_2 Distribution has Balanced Means as in (20).)

Traffic Intensity ρ	Service-Time Distribution							
	D, $c_s^2 = 0.0$		E_3, $c_s^2 = 0.333$		M, $c_s^2 = 1.0$		H_2, $c_s^2 = 4.0$	
	Exact	Approx.	Exact	Approx.	Exact	Approx.	Exact	Approx.
1.00	0.100	0.100	0.433	0.433	1.100	1.100	4.100	4.100
0.98	0.113	0.113	0.443	0.442	1.106	1.106	4.105	4.106
0.95	0.134	0.135	0.456	0.456	1.116	1.115	4.113	4.115
0.90	0.169	0.172	0.482	0.481	1.132	1.131	4.127	4.130
0.80	0.244	0.254	0.535	0.540	1.169	1.166	4.157	4.160
0.70	0.325	0.339	0.600	0.608	1.212	1.210	4.191	4.190
0.60	0.412	0.427	0.661	0.685	1.260	1.265	4.232	4.220
0.50	0.505	0.518	0.745	0.771	1.325	1.334	4.279	4.253
0.40	0.601	0.611	0.838	0.867	1.401	1.421	4.336	4.291
0.30	0.700	0.706	0.946	0.971	1.500	1.528	4.407	4.347
0.20	0.800	0.803	1.068	1.083	1.630	1.658	4.501	4.444
0.10	0.900	0.901	1.200	1.204	1.802	1.815	4.644	4.633
0.00	1.000	1.000	1.333	1.333	2.000	2.000	5.000	5.000
Max (over ρ) % error	—	4.3	—	3.6	—	1.7	—	1.4

Next, comparisons between the approximation for $c_z^2(\rho)$ and exact values for $H_2/M/1$ queues with $c_a^2 = 2.0$ and $c_a^2 = 12.0$ are displayed in Tables IV and V. Here different cases are obtained by allowing the third moment of the interarrival time to vary. In contrast to Tables II and III, a dramatic improvement over two moment approximations is evident in Tables IV and V. The exact values here come from Tables I and II of Whitt (1984b). Again, the mean queue length (including the customer in service) is transformed into $c_z^2(\rho)$ via $L = \lambda W$, (1) and (3). The accuracy of the approximation in these cases is usually very good, but not uniformly so, even though $\dot{c}_z^2(1)$ is exact by (29) and $\dot{c}_z^2(0)$ uses the full force of (15) because $0 < f(0) < \infty$. The accuracy degrades significantly when ρ and $E(T^3)$ are both small or large. However, this is to be expected, because $\dot{c}_z^2(0) \downarrow 0$ and $\dot{c}_z^2(1) \uparrow \infty$ as $E(T^3) \uparrow \infty$, while $\dot{c}_z^2(0) \uparrow \infty$ and $\dot{c}_z^2(1) \downarrow 0$ as $ET^3) \downarrow 1.5(E(T^2))^2$, the lower bound for $E(T^3)$. The reliability measures in Section 4 clearly depict the problem. The very large errors in these extreme cases obviously reveal limitations in the approximation here or the general procedure in (8), but the performance in these extreme cases can be interpreted more positively. The approximate function is actually reasonably close to the exact function $\{c_z^2(\rho): 0 \leq \rho \leq 1\}$; that is, the shape

Table IV

A Comparison of Approximations of the Normalized Mean Workload $c_z^2(\rho)$ in (1) With Exact Values for the $H_2/M/1$ Queue With $c_a^2 = 2.0$ and Different Third Moments (From Table I, p. 170, of Whitt 1984b)

Arrival Third Moment $E(T^3)$	$\rho = 0.3$		$\rho = 0.7$		$\rho = 0.9$	
	Exact	Approx.	Exact	Approx.	Exact	Approx.
13.5	3.000	3.000	3.000	3.000	3.000	3.000
14.6	2.604	2.512	2.902	2.875	2.974	2.970
16.2	2.357	2.333	2.781	2.756	2.936	2.932
18.0	2.244	2.251	2.675	2.679	2.897	2.897
20.9	2.156	2.170	2.546	2.567	2.838	2.842
32.1	2.067	2.051	2.302	2.268	2.659	2.654
167.9	2.007	2.006	2.044	2.014	2.159	2.042
∞	2.000	2.000	2.000	2.000	2.000	2.000

Table V
A Comparison of Approximations for the Normalized Mean Workload $c_z^2(\rho)$ in (1) With Exact Values for the $H_2/M/1$ Queue With $c_a^2 = 12.0$ and Different Third Moments (From Table II, p. 170, of Whitt 1984b).

Arrival Third Moment $E(T^3)$	$\rho = 0.3$		$\rho = 0.7$		$\rho = 0.9$	
	Exact	Approx.	Exact	Approx.	Exact	Approx.
253.5	13.00	13.00	13.00	13.00	13.00	13.00
254.0		9.23		12.38		12.94
280.7	9.93	6.00	12.41	10.96	12.84	12.64
312.6	7.02	5.09	11.73	10.57	12.66	12.50
351.6	4.85	4.60	10.92	10.12	12.45	12.32
401.2	3.63	4.26	9.93	9.67	12.17	12.10
468.0	2.99	3.93	8.71	9.11	11.80	11.82
565.1	2.62	3.45	7.17	8.56	11.27	11.44
722.7	2.38	2.89	5.36	7.08	10.44	10.74
1031.7	2.22	2.25	3.70	5.40	8.93	9.52
1946.0	2.10	2.07	2.61	2.77	5.64	6.82
18287.0	2.01	2.01	2.01	2.01	2.20	2.02
∞	2.00	2.00	2.00	2.00	2.00	2.00

is essentially the same in these cases. The discrepancies are primarily due to the curves being very steep at one end.

Table VI compares four different approximations for $c_z^2(\rho)$ with exact values in the $H_2/D/1$ queue, where the H_2 distribution has balanced means as in (2) with $c_a^2 = 2.0$. The exact values come from Seelen et al. The first three approximations are (40) with the correction factor $X = 1$ as in (38), (41) and (44). The fourth approximation is the simple linear interpolation in (5). The three refined approximations do significantly better than the simple interpolation for $\rho \leq 0.6$ because they reflect the light traffic derivative $\dot{c}_z^2(0) = 0.333$. The refinement (41) comes closest to matching the exact values in heavy traffic, e.g., at $\rho = 0.95$ and 0.98, but approximation (41) does not perform exceptionally well overall, even though $\dot{c}_z^2(1)$ is close and $\dot{c}_z^2(0)$ is exact and uses the full force of (15). In part, this can be explained by the fact that the interpolation produces an increasing convex fit (Case 1.1.2 in Section 3), while the actual function $c_z^2(\rho)$ is *not* convex, as can be seen from the exact values at 0.98 and 0.8; $c_z^2(\rho)$ is necessarily decreasing for some ρ above 0.8. In this case, a simple cubic fit as in Case 1.1.5 of Section 3 actually performs better. Evidently, the deterministic service times interact with the H_2 interarrival times in a rather strange way. (See Whitt 1984c for further evidence.) The maximum relative error using (41) with (40) is 8.3% for $H_2/D/1$, but only 0.4% for $H_2/M/1$ (see Table 1 of Fendick and Whitt). Table VI is one of the worst cases for (41).

Finally, comparisons between the approximation for $c_z^2(\rho)$ and the exact values for $H_2/H_2/1$ queues are displayed in Table VII. Here the third moments of both the interarrival and service times are allowed to vary. As noted in Whitt (1984b), from which the exact values are obtained, $c_z^2(\rho)$ does not change much as the third moment of the service time changes, justifying having the approximation depend more on the interarrival-time distribution than the service-time distribution, beyond the first two moments. Note that $c_z^2(\rho)$ is *constant* when $E(T^3)$ is at its lower or upper bound; see Whitt (1982, 1984) for further discussion.

6. HOW BAD CAN THE APPROXIMATION BE?

In Tables IV and V we saw how poorly the approximation can perform in $H_2/M/1$ queues when the third interarrival-time moment $E(T^3)$ and ρ are either both small or large, but this can be predicted because the derivatives $\dot{c}_z^2(0)$ and $\dot{c}_z^2(1)$ do not permit a reliable fit. As observed in Section 5, the curves are actually quite close in that example. In general, the situation can be much worse, as we show in this section.

First, we show that the interarrival-time density at the origin, $g(0)$, actually provides *no* rigorous control on $\dot{c}_z^2(\rho)$. Given any service-time distribution with $0 < c_s^2 < \infty$, any finite interarrival-time density $g(t)$ with $0 < c_a^2 < \infty$ and any positive ϵ_1 and ϵ_2, it is possible to modify $g(t)$ in the interval $[0, \delta]$ for sufficiently small δ so that $g(0)$ assumes any value we wish, while $c_z^2(\rho)$ changes by at most ϵ_1 for any ρ in the interval $[\epsilon_2, 1]$. This is an elementary consequence of

Table VI

A Comparison of Four Approximations for the Normalized Mean Workload $c_z^2(\rho)$ in (1) With Exact Values for the $H_2/D/1$ Queue With $c_a^2 = 2.0$. (The H_2 Distribution Has Balanced Means as in (20).)

		Approximation (40) With Adjustments			
Traffic Intensity ρ	Exact	$X = 1$ as in (38)	GI/G/1 (41)	General (44)	Linear Interp. (5)
1.00	2.000	2.000	2.000	2.000	2.000
0.98	1.973	1.980	1.975	1.970	1.980
0.95	1.931	1.948	1.938	1.926	1.950
0.90	1.859	1.894	1.877	1.854	1.900
0.80	1.712	1.779	1.757	1.718	1.800
0.70	1.567	1.658	1.642	1.590	1.700
0.60	1.433	1.536	1.530	1.473	1.600
0.50	1.316	1.417	1.424	1.365	1.500
0.40	1.221	1.304	1.322	1.267	1.400
0.30	1.145	1.202	1.227	1.181	1.300
0.20	1.085	1.115	1.140	1.107	1.200
0.10	1.037	1.046	1.061	1.045	1.100
0.00	1.000	1.000	1.000	1.000	1.000
$c_z^2(0)$	1.000	1.000	1.000	1.000	
$\dot{c}_z^2(0)$	0.333	0.333	0.333	0.333	
$c_z^2(1)$	2.000	2.000	2.000	2.000	
$\dot{c}_z^2(1)$	$\approx$1.350	1.000	1.250	1.500	
Fitting case (Sec. 3)	—	1.1.5 Cubic	1.1.2 Convex	1.1.2 Convex	—
Max. (over ρ) % error	—	6.8	8.3	3.8	14.7

continuity or stability results for the GI/G/1 queue; e.g., p. 194 of Asmussen (1987). Hence, $g(0)$ is useful only to the extent that it genuinely represents the interarrival-time distribution for nonnegligible small values. When this is in doubt, better approximations may be obtained by using $t^{-1}G(t)$ for small t (e.g., $t =$ 0.1 or 0.2) instead of $g(0)$ in the formula for $\dot{c}_z^2(0)$, but the full cdf is less likely to be available. The main point is that the light traffic limit simply does not rigorously control the behavior of $\dot{c}_z^2(\rho)$.

It remains to determine how well the first two moments of the service-time distribution and the first three moments of the interarrival-time distribution rigorously control $c_z^2(\rho)$. For the special case of an exponential service-time distribution, which is the only case in which we have a nice exact expression for $\dot{c}_z^2(1)$, this issue was addressed by Whitt (1984a, b) and Klincewicz and Whitt (1984); see especially Section IV of Whitt (1984a). There it was found that three interarrival-time moments do substantially limit the

Table VII

A Comparison of Approximations for the Normalized Mean Workload With Exact Values for Several $H_2/H_2/1$ Queues With $\rho = 0.7$ and $c_a^2 = c_s^2 = 2.0$ (from Table III of Whitt 1984b). The Third Moments of Both the Interarrival Time and Service Time are Allowed to Vary.

	Exact Values					Approximations			
Arrival Third Moment	H_2 Service-Time Distribution Third Moments					(40) With			Linear Interp.
$E(T^3)$	13.6	14.6	18.0	32.1	167.9	$X = 1$	(41)	(44)	(5)
13.5	4.00	4.00	4.00	4.00	4.00	4.00	4.00	4.00	3.70
14.6	3.93	3.93	3.91	3.91	3.90	3.88	3.89	3.89	3.70
18.0	3.75	3.74	3.72	3.69	3.60	3.70	3.71	3.74	3.70
32.1	3.39	3.39	3.38	3.35	3.32	3.29	3.31	3.37	3.70
∞	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.70

possible values of $E(Z_\rho)$ and thus $c_z^2(\rho)$, with three moments doing much better than two. Moreover, for any ρ, the smallest and largest possible values of $c_z^2(\rho)$ are attained by special two-point extremal distributions, so that we can easily calculate the set of possible $c_z^2(\rho)$ values in any case. This analysis of $g(0)$ and $E(T^3)$ supports the notion that the heavy traffic derivative is typically more useful than the light traffic derivative for determining $c_z^2(\rho)$.

For the GI/G/1 model in general, it remains to determine how much $E(W_\rho)$ and $c_z^2(\rho)$ are constrained by the first few moments of S and T or even of $(S - \rho^{-1}T)$. To see the possibilities, we conclude with a somewhat pathological example. We use a two-point service-time distribution

$$P(S = 0) = 0.1 = 1 - P(S = 10/9) \tag{66}$$

so that $E(S) = 1$ and $c_s^2 = 1/9$. We use a three-point interarrival-time distribution

$$\begin{aligned} P(T = 100) &= 0.0001, \\ P(T = 1) &= 0.99, \\ P(T = 0) &= 0.0099 \end{aligned} \tag{67}$$

so that $E(T) = 1$, $c_a^2 = 0.99$, $E(T^3) = 100.99$, $c_z^2(0) = 10/9 = c_z^2(1) + 0.1$, and $\dot{c}_z^2(0) = +\infty$. Since $P(S - \rho^{-1}T \leq 0) = 0.9901$ for $\rho \leq 0.9$, $E(W_\rho) \approx 0$ and $c_z^2(\rho) \approx 10(1 - \rho)/9$ for $\rho \leq 0.9$, but $E(W_\rho)$ increases dramatically for $\rho > 0.9$, with $c_z^2(1) \approx 1.0$. The interpolation approximation certainly does not describe $c_z^2(\rho)$ well for $\rho \leq 0.9$. Since $c_z^2(0) \approx c_z^2(1) \approx 1.0$, we might think that this model behaves like an M/D/1 queue, which is not nearly the case. Of course, the derivatives $\dot{c}_z^2(0)$ and $\dot{c}_z^2(1)$ suggest otherwise, but in this case only the heavy traffic derivative $\dot{c}_z^2(1)$, which we do not know exactly, provides useful information. In summary, Sections 5 and 6 support the conclusion drawn in Whitt (1984a, c) and Klincewicz and Whitt that relatively simple approximations should perform well *provided* that the distributions are not too strange.

ACKNOWLEDGMENT

I am very grateful to Kerry Fendick for preparing the interpolation program and collaborating on the research that motivated this paper.

REFERENCES

ASMUSSEN, S. 1984. Approximations for the Probability of Ruin Within Finite Time. *Scand. Act. J.* 31–57.

ASMUSSEN, S. 1987. *Applied Probability and Queues.* John Wiley & Sons, New York.

BRUMELLE, S. L. 1971. On the Relation Between Customer and Time Averages in Queues. *J. Appl. Prob.* **8,** 508–520.

BURMAN, D. Y., AND D. R. SMITH. 1983. Asymptotic Analysis of a Queueing Model With Bursty Traffic. *Bell Syst. Tech. J.* **62,** 1433–1453.

BURMAN, D. Y., AND D. R. SMITH. 1986. An Asymptotic Analysis of a Queueing System With Markov-Modulated Arrivals. *Opns. Res.* **34,** 105–119.

CHUNG, K. L. 1974. *A Course in Probability Theory,* 2nd ed. Academic Press, New York.

COHEN, J. W. 1982. *The Single Server Queue,* 2nd ed. North-Holland, Amsterdam.

DALEY, D. J., AND T. ROLSKI. 1984. Light Traffic Approximation for a Single-Server Queue. *Math. Opns. Res.* **9,** 624–628.

DALEY, D. J., AND T. ROLSKI. 1990. Light Traffic Approximations in Queues. *Math. Opns. Res.* **15** (to appear).

FENDICK, K. W., V. R. SAKSENA AND W. WHITT. 1989a. Dependence in Packet Queues. *IEEE Trans. Commun.* **37** (to appear).

FENDICK, K. W., V. R. SAKSENA AND W. WHITT. 1989b. Investigating Dependence in Packet Queues With the Index of Dispersion for Work. (Submitted for publication).

FENDICK, K. W., AND W. WHITT. 1989. Measurements and Approximations to Describe the Offered Traffic and Predict the Average Workload in a Single-Server Queue. *Proc. IEEE* **77,** 171–194.

HALFIN, S. 1985. Delays In Queues, Properties and Approximations. In *Teletraffic Issues in an Advanced Information Society, ITC-11,* M. Akiyama (ed.). Elsevier, Amsterdam, 47–52.

HEYMAN, D. P., AND M. J. SOBEL, 1982. *Stochastic Models in Operations Research,* Vol. I. McGraw-Hill, New York.

KINGMAN, J. F. C. 1962. On Queues in Heavy Traffic. *J. Roy. Statist. Soc.* **B24,** 383–392.

KLINCEWICZ, J. G., AND W. WHITT. 1984. On Approximations for Queues, II: Shape Constraints. *AT&T Bell Lab. Tech. J.* **63,** 139–162.

KNESSL, C. 1990. Refinements to Heavy Traffic Limit Theorems in Queueing Theory. *Opns. Res.* **38** (to appear).

KÖLLERSTRÖM, J. 1981. A Second-Order Heavy Traffic Approximation for the Queue GI/G/1. *Adv. Appl. Prob.* **13,** 167–185.

KRAEMER, W., AND M. LANGENBACH-BELZ. 1976. Approximate Formulae for the Delay in the Queueing System GI/G/1. *Proc. 8th Int. Teletraffic Cong.,* Melbourne, pp. 235-1–8.

LAI, T. L. 1976. Asymptotic Moments of Random Walks With Application to Ladder Variables and Renewal Theory. *Ann. Prob.* **4,** 51–66.

MARSHALL, K. T. 1968. Some Inequalities in Queueing. *Opns. Res.* **16,** 651–665.

MINH, D. L., AND R. H. SORLI. 1983. Simulating the GI/G/1 Queue in Heavy Traffic. *Opns. Res.* **31,** 966–971.

NEWELL, G. F. 1982. *Applications of Queueing Theory,* 2nd ed. Chapman & Hall, London.

PRABHU, N. U. 1980. *Stochastic Storage Processes.* Springer-Verlag, New York.

RAMASWAMI, V., AND G. LATOUCHE. 1987. An Experimental Evaluation of the Matrix-Geometric Method for the GI/PH/1 Queue. Bell Communications Research, Morristown, N.J.

REIMAN, M. I., AND B. SIMON. 1988. An Interpolation Approximation for Queueing Systems With Poisson Input. *Opns. Res.,* **36,** 454–469.

REIMAN, M. I., AND B. SIMON. 1989. Open Queueing Systems in Light Traffic. *Math. Opns. Res.* **14,** 26–59.

REIMAN, M. I., AND A. WEISS. 1989. Light Traffic Derivatives via Likelihood Ratios. *IEEE Trans. Inf. Thy.* **35,** 648–654.

SEELEN, L. P., H. C. TIJMS AND M. H. VAN HOORN. 1985. *Tables for Multi-Server Queues.* North-Holland, Amsterdam.

SEGAL, M., AND W. WHITT. 1988. A Queueing Network Analyzer for Manufacturing. In *Teletraffic Science for New Cost-Effective Systems, Networks and Services, ITC-12,* M. Bonatti (ed.). Elsevier, Amsterdam, 1146–1152.

SIEGMUND, D. 1979. Corrected Diffusion Approximations in Certain Random Walk Problems. *Adv. Appl. Prob.* **11,** 701–719.

SIEGMUND, D. 1985. *Sequential Analysis.* Springer-Verlag, New York.

SPITZER, F. 1960. A Tauberian Theorem and Its Probability Interpretation. *Trans. Am. Math. Soc.* **94,** 150–169.

SPITZER, F. 1964. *Principles of Random Walk.* Van Nostrand, Princeton, N.J.

STOYAN, D. 1983. *Comparison Methods for Queues and Other Stochastic Models.* D. J. Daley, (ed.) John Wiley & Sons, New York.

TIJMS, H. C. 1986. *Stochastic Modelling and Analysis.* John Wiley & Sons, New York.

WHITT, W. 1982. The Marshall and Stoyan Bounds for IMRL/G/1 Queues are Tight. *Opns. Res. Letters* **1,** 209–213.

WHITT, W. 1983. The Queueing Network Analyzer. *Bell System Tech. J.* **62,** 2779–2815.

WHITT, W. 1984a. On Approximations for Queues, I; Extremal Distributions. *AT&T Bell Lab. Tech. J.* **63,** 115–138.

WHITT, W. 1984b. On Approximations for Queues, III: Mixtures of Exponential Distributions. *AT&T Bell Lab. Tech. J.* **63,** 163–175.

WHITT, W. 1984c. Minimizing Delays in the GI/G/1 Queue. *Opns. Res.* **32,** 41–51.

WHITT, W. 1986. Deciding Which Queue to Join: Some Counterexamples. *Opns. Res.* **34,** 55–62.