PRODUCTION AND OPERATIONS MANAGEMENT
Vol. 2, No. 2, Spring 1993
Printed in U.S.A.

APPROXIMATIONS FOR THE GI/G/m QUEUE*

WARD WHITT
AT&T Bell Laboratories, Murray Hill, New Jersey 07974-0636, *USA*

Queueing models can usefully represent production systems experiencing congestion due to irregular flows, but exact analyses of these queueing models can be difficult. Thus it is natural to seek relatively simple approximations that are suitably accurate for engineering purposes. Here approximations for a basic queueing model are developed and evaluated. The model is the GI/G/m queue, which has m identical servers in parallel, unlimited waiting room, and the first-come first-served queue discipline, with service and interarrival times coming from independent sequences of independent and identically distributed random variables with general distributions. The approximations depend on the general interarrival-time and service-time distributions only through their first two moments. The main focus is on the expected waiting time and the probability of having to wait before beginning service, but approximations are also developed for other congestion measures, including the entire distributions of waiting time, queue-length and number in system. These relatively simple approximations are useful supplements to algorithms for computing the exact values that have been developed in recent years. The simple approximations can serve as starting points for developing approximations for more complicated systems for which exact solutions are not yet available. These approximations are especially useful for incorporating GI/G/m models in larger models, such as queueing networks, wherein the approximations can be components of rapid modeling tools.

(PERFORMANCE EVALUATION; ROUGH CUT ANALYSES; RAPID MODELING TOOLS; QUEUEING THEORY; QUEUEING NETWORKS; PRODUCTION NETWORKS; CONGESTION MEASURES; MULTISERVER QUEUES; WAITING TIMES; HEAVY-TRAFFIC LIMIT THEOREMS; APPROXIMATIONS)

1. Introduction

I present relatively simple approximations for the principal steady-state congestion measures describing the standard GI/G/m queueing model. The GI/G/m model has a single service facility with m identical servers, unlimited waiting room and the first-come first-served queue discipline. The service times are independent and identically distributed (IID) with a general distribution, the interarrival times are also IID with a general distribution, and the interarrival times are independent of the service times. I assume that the general interarrival-time and service-time distributions are each partially specified by their first two moments. Equivalently, I assume that the arrival process is partially specified by the *arrival rate* λ (the mean interarrival time

* Received December 1992; accepted June 1993.

1059-1478/93/0202/0000$1.25

is λ^{-1}) and the *squared coefficient of variation* (SCV, variance divided by the square of the mean) of an interarrival time, denoted by c_a^2. (I use the SCV instead of the variance, because it is a dimensionless quantity, easily interpreted independently of the mean.) Similarly, I assume that the service-time distribution is partially specified by its *mean* τ and its SCV c_s^2. All descriptions of this model thus depend only on the basic parameter 5-tuple $(\lambda, c_a^2, \tau, c_s^2, m)$.

Because the GI/G/m queueing model is generic, it has many potential applications (e.g., to computer, communication and production systems). Here I emphasize applying the model to the performance analysis of production systems, especially production networks (for background, see Buzacott and Shanthikumar 1992 and Suri, Sanders, and Kamath 1993). The approximations can be (and have been) used in what Suri et al. call *aggregate dynamic models* (ADMS) to perform *rough cut analyses*. Indeed, this paper is a minor revision of my earlier paper, Whitt (1985), previously cited in work on *queueing network analyzers* and *rapid modeling tools* (Segal and Whitt 1989; Suri and de Treville 1991, 1992; Whitt 1992, 1994; Suri, Sanders, and Kamath 1993). Tijms (1986) provides background on GI/G/m approximations in his fourth chapter.

Great progress has been made on numerical procedures for obtaining exact descriptions of GI/G/m models (Takahashi and Takami 1976; de Smit 1983a, 1983b; van Hoorn and Seelen 1984; Lucantoni and Ramaswami 1985; Ramaswami and Lucantoni 1985a, 1985b; Seelen 1986; Bertsimas 1988, 1990; Abate, Choudhury, and Whitt 1993). For some applications, these procedures eliminate the need for the kinds of approximations developed here. But even when exact numerical procedures are available, it is helpful to have simple approximations as concise summaries. For other applications, e.g., when GI/G/m models appear as submodels, simple closed-form analytic formulas are still useful. We use the exact numerical results to evaluate our approximations.

The approximations I present can be applied directly to a single GI/G/m queue, often yielding back-of-the-envelope formulas (Whitt 1992). However, their most common applications are as subroutines in algorithms for analyzing networks of queues. AT&T Bell Laboratories has applied some of the approximations here in their *Queueing Network Analyzer* (QNA) software package (Whitt 1983a, 1984c; Segal and Whitt 1989). The approximations here have also been used in software packages developed by others.

QNA describes open queueing network models with nonexponential service times, non-Poisson arrival processes, and non-Markovian (e.g., deterministic) routing, for which exact analytical techniques are unavailable. The approach to approximation is parametric-decomposition: the queues in the network are treated as independent GI/G/m models, each partially specified by the basic parameter 5-tuple $(\lambda, c_a^2, \tau, c_s^2, m)$ at that queue. The goal is to use the two arrival parameters (λ and c_a^2) at each queue to capture the main effects of the dependence among the queues and the actual properties of the arrival process at each queue. Solving the usual system of linear traffic-rate equations, an exact analysis, determines the arrival rates and solving another system of linear equations, which attempts to capture the significant variability effects, determines the approximate arrival variability parameters (Whitt 1982b, 1983a).

Interarrival times at each queue are not typically independent, but the two-parameter characterization is an approximation by a renewal process (having independent interarrival times). The idea is *not* to ignore the dependence among successive interarrival times, but to try to capture its essential properties with the variability parameter c_a^2. The dependence is represented in c_a^2 by deliberately making c_a^2 different from the SCV of an interarrival time. Methods for doing this, such as the asymptotic method, are described in Fendick and Whitt (1989), Fendick, Saksena, and Whitt (1991) and Whitt (1982b, 1983a, 1994).

Because the effect of dependence among the interarrival times on the queue depends on the *traffic intensity* $\rho \equiv \lambda\tau/m$, QNA allows c_a^2 to depend on ρ. Given the two arrival parameters λ and c_a^2, the QNA algorithm treats the different queues as independent, each with its own a renewal arrival process. Since each queue is regarded as a GI/G/m queue partially specified by the basic parameter 5-tuple $(\lambda, c_a^2, \tau, c_s^2, m)$, the approximations I present here may be applied directly. Moreover, because interarrival times at a queue in a network of queues are rarely independent (unless the arrival process is nearly Poisson) and because extra information about the arrival process at each queue is usually unavailable, the partially characterized GI/G/m queue is appropriate. Hence there are important examples of the GI/G/m model being meaningful while additional information about the underlying distributions is unavailable. See Whitt (1983a, 1984c, 1994), Bitran and Tirupati (1988), Segal and Whitt (1989), Suri, Sanders, and Kamath (1993) and references cited by them for further discussion of how to analyze networks. See Harrison and Nguyen (1990) and Dai, Nguyen, and Reiman (1993) for alternative approaches to queueing networks based on multidimensional reflected Brownian motion. From this point on, I will focus on a single GI/G/m queue.

Given the parameter 5-tuple $(\lambda, c_a^2, \tau, c_s^2, m)$, I assume that $\rho < 1$, so that a proper steady state exists (Asmussen 1987). I let $\tau = 1$ without loss of generality, using appropriate measuring units. This yields the basic parameter 4-tuple (ρ, c_a^2, c_s^2, m). Because there is a set of probability distributions consistent with two moments, there is a set of possible congestion measures associated with any basic parameter 4-tuple (ρ, c_a^2, c_s^2, m). I regard the approximate congestion measures as approximations for this set. The goal is not to attain an extremely close fit to a particular model based on specific distributions, but to approach typical values in the set of congestion measures. See Whitt (1984a, 1984b), Klincewicz and Whitt (1984) and Johnson and Taaffe (1991) for a discussion of this philosophy and an examination of some single-server queues from this perspective.

My experience has been that a congestion value near the middle of the set of possible values usually yields a satisfactory approximation (in the order of 10 percent relative error), *provided* that (1) the underlying probability distributions are not too irregular, (2) the variability parameters c_a^2 and c_s^2 (especially c_a^2) are not too large, and (3) the traffic intensity ρ is not too small. The violation of any of these conditions should be is a clear warning.

Unusually light traffic (low ρ) or high variability (high c^2) does not mean that alternative formulas based on the same parameter would be superior, but that the partial information (ρ, c_a^2, c_s^2, m) does not adequately determine the congestion measures. If one needs more accuracy, additional information about the distributions is needed. If the distributions are known to have unusual shape, refinements can be developed. To be more specific, distributions similar to the two-point extremal distributions described in Whitt (1984a) or the unimodel extremal distributions in

Klincewicz and Whitt (1984) are highly irregular. A tendency toward either of these extremes dictates the kind of correction to make.

Difficulties with variability parameters may begin when $c_a^2 = 5$ and may become serious when $c_a^2 \geq 15$. From Whitt (1984a, Corollary 1 to Theorem 2), the *maximum* relative error in the expected number in system in a partially specified GI/M/1 model is c_a^2 ($100c_a^2$ percent); according to Klincewicz and Whitt (1984), *typical* relative errors might be about $(0.04)\rho^{-1}c_a^2$, which for $\rho = 0.8$ and $c_s^2 = 1$ is 5 percent, but for $\rho = 0.2$ and $c_s^2 = 10$ is 200 percent. The definition of light traffic depends on the number of servers. For example, the traffic intensity becomes too small when $\rho \leq 0.3$ for $m = 1$ and $\rho \leq 0.6$ for $m = 20$. However, I tend not to be too concerned about relative errors in light traffic as long as the absolute errors are small. The discussion and numerical examples should help clarify these points. I primarily evaluate the approximations by comparing them to the standard special cases.

Section 5 of Whitt (1983a) contains approximations for some congestion measures of a partially specified GI/G/m queue, but there attention was primarily on approximations for the GI/G/1 model. For the single-server special case, I was able to rely heavily on the excellent work by Kraemer and Langenbach-Belz (1976). The previous approximations for multiserver queues in Section 5.2 of Whitt (1983a) are much less accurate than the single-server approximations in Section 5.1. By contrast, the new multiserver approximations I describe here are nearly as good as the single-server approximations. I now also provide an approximation for the delay and queue-length distributions for multiserver queues, whereas only the first two moments were described before. The main contributions here in relation to the literature on approximations for the GI/G/m queue (Kimura 1986; Tijms 1986; Allen 1990) are the approximations that go beyond the steady-state means.

My approximation for the delay distribution in the GI/G/m model is essentially the same as for the single-server queue in Whitt (1983a, Section 5.1), with the addition of new approximations for the expected waiting time and the probability of delay. My idea is to assign the specified probability mass at zero (the probability of no delay) and approximate the density of the conditional delay, given that the server is busy, by fitting mixtures or convolutions of two exponential distributions to the first two moments of the conditional delay. This tends to work well, because actual delay distributions often have approximately this form. The GI/G/m queueing system could be described as a smoothing operation: The descriptive characteristics (delay distribution, etc.) tend to be better behaved (more regular, i.e., closer to exponential) than the underlying interarrival-time and service-time distributions. Hence, queue behavior can be described surprisingly well given the partial information (ρ, c_a^2, c_s^2, m).

An alternative approach to approximating steady-state distributions (briefly discussed in Section 4.2) is simple exponential approximation using asymptotics: approximate the steady-state waiting-time tail probability $P(W > x)$ by $\alpha e^{-\eta x}$, where η and α are determined from the limit $e^{\eta x}P(W > x) \to \alpha$ as $x \to \infty$. The parameters η and α are called the *asymptotic decay rate* and *asymptotic constant*, respectively. Abate, Choudhury and Whitt (1994a, 1994b, 1994c), Abate and Whitt (1994) and Choudhury and Whitt (1994) discuss exponential approximations for steady-state distributions in the GI/G/m model based on asymptotics. The key quantity is the asymptotic decay rate η, which in general depends on *more* than the basic 4-tuple (ρ, c_a^2, c_s^2, m). However, my approximations here contribute significantly to the asymptotics approach by providing convenient approximations for the asymptotic constant α. Abate, Choudhury, and Whitt (1994a) suggest approximating α by ηEW. I focus on the conditional waiting-time tail probability $P(W > x | W > 0)$, and approximate α by $\eta EW/P(W > 0)$. I describe convenient approximations for EW and $P(W > 0)$.

My approach here is to build on the exact formulas for the Markovian M/M/m queue, in particular, the classical Erlang-C formula (2.3) below. I also offer an alternative closed-form expression for the M/M/m model; see Section 2.3. I exploit heavy-traffic theory to obtain first-order refinements (Borovkov 1965, 1967; Iglehart 1965; Kingman 1965; Iglehart and Whitt 1970; Köllerström 1974; Halfin and Whitt 1981; Whitt 1982a). An appropriate view of the total procedure may be as a heavy-traffic approximation. With judicious refinements, heavy-traffic approximations are effective over a wide range of parameter values. The most important heavy-traffic approximation for obtaining these new results is the approximation for the probability of delay in the GI/M/m model from Halfin and Whitt (1981).

We derive second-order refinements in the expected waiting time, by exploiting the excellent approximations for the M/D/m and D/M/m systems developed by Cosmetatos (1975). [However, even these need refinement in very light traffic; see (2.17) below]. Finally, I adjust these approximations after making comparisons with exact values, extensively using tables (Kühn 1976; Hillier and Yu 1981; Groenevelt, van Hoorn, and Tijms 1984; Seelen and Tijms 1984; Seelen, Tijms, and van Hoorn 1985; and de Smit 1983a, 1983b, and personal communication). Values for a few cases not covered by the tables were obtained from the Q-LIB program developed by Seelen, Tijms, and van Hoorn (private communication). Numerical procedures developed for the exact solution of the GI/PH/m queue developed by Lucantoni and Ramaswami (1985), Ramaswami and Lucantoni (1985a, 1985b) and Bertsimas (1988, 1990) could also be used. Seelen, Tijms, and van Hoorn (1985) would have sufficed for much of this study; the other tables were used primarily because their book was not available until my work was almost finished.

Data from de Smit (1983a, 1983b, and personal communication) were especially helpful, because they give an indication of the range of reasonable exact values consistent with the basic parameter 4-tuple (ρ, c_a^2, c_s^2, m). De Smit provided exact values for the principal congestion measures for several $GI/H_2/m$ models with different interarrival-time and service-time distributions but with the same basic parameter 4-tuple (ρ, c_a^2, c_s^2, m). Unlike Whitt (1984a, 1984b), we do not know the set of all possible values for these multiserver queues, but de Smit's cases indicate where the typical values lie. The size of this set in de Smit's data gives a good idea of the accuracy possible with these approximations based on a partial characterization of the distributions (Tables 8–10 and 18–20). The level if accuracy obtained with partial moment information is limited, but results show that reasonable, practical approximations are possible.

2. The Expected Waiting Time

Here I focus on the expected (steady-state) waiting time (before beginning service). I introduce some notation (Section 2.1), then relate the expected waiting time EW to four other mean values (Section 2.2). I will consider the classic M/M/m model (Section 2.3) and present the approximation formulas of Sakasegawa (1977) and Halfin and Whitt (1981), discuss heavy-traffic approximations (Section 2.4), and the Cosmetatos (1975) approximations for the D/M/m and M/D/m models (Section 2.5). Finally, I will develop an approximation for the general GI/G/m model (Section 2.6) and make numerical comparisons (Section 2.7).

2.1 *Basic Notation*

Following convention, let M, D, E_k, H_k and G denote the special distributions: exponential, deterministic, Erlang with k phases, hyperexponential (mixture of k exponentials) and general, respectively. Consider these distributions for the interarrival times and service times in the GI/G/m model. Always consider this model in equilibrium or in steady state.

Let W represent the waiting time before beginning service and let EW be its expected value. $EW(\mathrm{M/H_2/m})$ indicates EW in the $\mathrm{M/H_2/m}$ model. Let the *traffic intensity* be defined as usual by $\rho = \lambda\tau/m$. Assume that $\rho < 1$, so that the system is stable (a proper steady-state distribution exists for the sequence of waiting times). Let $EW(\rho,\ c_a^2,\ c_s^2,\ m)$ represent EW as a function of the four parameters $(\rho,\ c_a^2,\ c_s^2,\ m)$ with the understanding that $\tau = 1$ and $\lambda = m\rho$. Because

$$EW(\lambda,\ c_a^2,\ \tau,\ c_s^2,\ m) = \tau EW(\lambda\tau,\ c_a^2,\ 1,\ c_s^2,\ m)$$

is an exact relationship for any GI/G/m model with these parameters, it suffices to reduce the parameters from five to four and consider $EW(\rho,\ c_a^2,\ c_s^2,\ m)$.

The approximations here all depend on the basic 4-tuple $(\rho,\ c_a^2,\ c_s^2,\ m)$. In some cases extra information can be very useful. For example, Whitt (1984a, 1984b, 1989) shows how the third moments of the interarrival time can improve the approximations for EW when $m = 1$. We consider only the basic 4-tuple $(\rho,\ c_a^2,\ c_s^2,\ m)$ here though.

2.2 *Related Congestion Measures*

Along with the expected waiting time, I also consider four other related mean values. Let Q be the queue length (the number of customers waiting, not counting customers in service) at an arbitrary time (not at an arrival instant), let B be the number of busy servers at an arbitrary time, let N be the number of customers in the system at an arbitrary time, and let T be the customer sojourn time (waiting time plus service time). Clearly $N = B + Q$ and $T = W + V$ where V is a service time, so that

$$EN = EB + EQ \qquad \text{and} \qquad ET = EW + \tau. \tag{2.1}$$

Also, from Little's Law [the relation $L = \lambda W$ (with different notation), Heyman and Sobel (1982, Section 11.3), Franken, König, Arndt, and Schmidt (1981, Chapter 4) and Whitt (1991)],

$$EB = m\rho = \lambda\tau, \qquad EQ = \lambda EW, \qquad \text{and} \qquad EN = \lambda ET. \tag{2.2}$$

All the formulas in (2.1) and (2.2) are *exact*, even if the independence assumptions of the GI/G/m model are dropped. As a consequence, in a complicated open queueing network model, these relationships are valid without any approximation. In particular, ρ and EB depend only on the arrival rates and service rates, and if these are known, then ρ and EB in (2.2) are exact. Because EB often is a large part of EN by (2.1), any reasonable approximation of EQ often produces a very close approximation of EN:

Example 1. To illustrate the power of the exact relationships in (2.1) and (2.2) above, suppose that we are interested in EN (the expected number of customers in the system) in the $E_4/M/20$ model with $\rho = 0.90$. From Table 5, we see that the exact value of EQ is 2.55, while the previous QNA approximation (Whitt 1983a, Section 5.2) and the new approximation are 3.10 and 2.67, respectively. For EQ, then, the relative percentage errors ($100 \times |\text{exact} - \text{approx.}|/\text{exact}$) are 21.6 and 4.7%, respectively. (This comparison illustrates what can be gained with our new approximations.) However, we know EB exactly: by (2.2), $EB = m\rho = 18$. Hence, the exact value for EN is 20.55, while the approximate values are 21.10 and 20.67. For EN, the relative percentage errors in the two approximations for EN are thus 2.7 and 0.6%, respectively. Both approximations for EN are very close, primarily because of the exact relations in (2.1) and (2.2).

Given λ, τ and m, it suffices to determine one of EW, EN, EQ or ET in order to have all four according to (2.1) and (2.2). Hence, I develop an approximation for EW and then apply (2.1) and (2.2) to obtain corresponding approximations of EQ, EN and ET. We compare approximations with exact values of EQ and EW, where the only approximation is needed. Example 1 shows that the relative accuracy of the corresponding approximations for EN and ET is necessarily better, and often substantially so.

2.3 *The M/M/m Model*

I use the well-known exact values for the M/M/m model to construct my approximations (Halfin and Whitt 1981, Section 1). The Erlang delay formula or Erlang-C formula is a key quantity here, giving the probability all servers are busy,

$$P(B = m) = P(N \geq m) = [(m\rho)^m/m!(1 - \rho)]\zeta, \tag{2.3}$$

where

$$\zeta \equiv \left[(m\rho)^m/(m!(1 - \rho)) + \sum_{k=0}^{m-1} (m\rho)^k/k!\right]^{-1}. \tag{2.4}$$

Because the arrival process is Poisson (Wolff 1982), $P(W > 0) = P(N \geq m)$ in the M/M/m model. This relation remains true for all M/G/m models, but not other GI/G/m models, because N is the equilibrium number in the system at an arbitrary time (the steady-state distribution associated with the continuous-time process), whereas W is the waiting time at a customer arrival instant. Seelen, Tijms, and van Hoorn (1985) show the difference between $P(W > 0)$ and $P(N \geq m)$ for many GI/G/m models in their tables.

The Erlang-C formula (2.3) is significant for the expected queue length in the M/M/m model because

$$EQ = P(N \geq m)\rho/(1 - \rho); \tag{2.5}$$

[Halfin and Whitt 1981, (1.8)]. Combine (2.2) and (2.5) to obtain

$$EW = \tau P(N \geq m)/m(1 - \rho). \tag{2.6}$$

Formulas (2.5) and (2.6) show that it is natural to decompose the means EQ and EW into two parts:

$$EQ = P(N \geq m)E(Q|N \geq m) \tag{2.7}$$

and

$$EW = P(W > 0)E(W \mid W > 0), \tag{2.8}$$

where

$$E(Q \mid B \geq m) = \frac{\rho}{1-\rho} \quad \text{and}$$

$$E(W \mid W > 0) = \frac{E(Q \mid B \geq m)}{\lambda} = \frac{\tau}{m(1-\rho)}. \tag{2.9}$$

When m gets large, often $P(N \geq m) = P(W > 0)$ gets small, so that it is helpful to look at both components of (2.7) and (2.8). This is important in properly understanding GI/G/m steady-state behavior.

I construct the approximations for EW in GI/G/m models using (2.6) together with the exact value of the Erlang-C formula (2.3) for the M/M/m model. [Accurate calculation with (2.3) requires care.] However, to approximate EW in some applications closed-form approximate expression for the Erlang-C formula may be convenient. Halfin and Whitt (1981) developed a closed-form approximation for the Erlang-C formula, namely

$$P(N \geq m) \approx \xi \equiv \xi(\beta) = [1 + \sqrt{2\pi}\beta\Phi(\beta)\exp(\beta^2/2)]^{-1} \tag{2.10}$$

where $\beta = (1-\rho)m^{1/2}$ and $\Phi(t)$ is the cumulative distribution function (CDF) of the standard normal distribution having mean 0 and variance 1. Halfin and Whitt's (1981) Proposition 1 establishes that (2.10) is asymptotically correct in a certain case of heavy traffic, in particular as $\rho \rightarrow 1$ and $m \rightarrow \infty$ with $(1-\rho)m^{1/2} \rightarrow \beta$. In fact, approximation (2.10) performs well over a wide range of m and ρ (Table 13 and Table 1, where, combined with the exact relations (2.2) and (2.5), it yields an approximation for EQ). However, the accuracy of the approximation tends to degrade as m increases with fixed β or as $(1-\rho)m^{1/2}$ increases.

Formula (2.10) is essentially a closed form, but it does involve the normal CDF $\Phi(t)$. However, the normal CDF can be approximated very accurately by a rational approximation (Abramowitz and Stegun 1972, p. 299). Although (2.10) could be further refined to obtain an even better approximation, it is beyond the scope of this work. Using Formula (2.10) consistently slightly underestimates the exact value when m is sufficiently small and ρ is sufficiently large, while it consistently overestimates the exact value when m is sufficiently large and ρ is sufficiently small (Table 13). Because the latter discrepancy can be quite large (for example, $\rho = 0.70$ and $m = 100$), it would be useful to refine (2.10) in this region.

For large m and not too large ρ, it is natural to consider a normal approximation associated with the heavy-traffic limit theorem for GI/G/m systems in which $\lambda \rightarrow \infty$ and $m \rightarrow \infty$ with τ and ρ fixed (Iglehart 1965; Borovkov 1967; Whitt 1982a; Glynn and Whitt 1991; Whitt 1992). For an M/G/∞ system, this is simply a normal approximation for the exact Poisson distribution. The approximation is

$$P(N(\mathrm{M/M/m}) \geq m) \approx P(N(\mathrm{M/M/\infty}) \geq m) \approx P(N(\mathrm{M/M/\infty}) \geq m - 0.5)$$
$$\approx 1 - \Phi((m - m\rho - 0.5)/\sqrt{m\rho}), \tag{2.11}$$

where again Φ is the standard normal CDF. I discuss the normal approximation (2.11) further in Section 3. Obviously this is a poor method of approximation when

ρ is large and m is small but shows promise as a tool for constructing refined hybrid approximations when both ρ is small and m is large.

Sakasegawa (1977) proposed another closed-form approximation for EW(M/M/m):

$$EW(\mathrm{M/M/m}) \approx \tau(\rho^{(\sqrt{2(m+1)}-1)})/(m(1-\rho)). \tag{2.12}$$

Although formula (2.12) has been shown to perform quite well, it is not consistent with known heavy traffic limits: in particular, (2.12) yields $(1-\rho)EW \to 0$ (instead of 1) as $\rho \to 1$ with m fixed and (2.12) yields $(1-\rho)EW \to 0$ as $\rho \to 1$ and $m \to \infty$ with $(1-\rho)m^{1/2} \to \beta$ [instead of (2.10)]. Consistent with numerical experience, this analysis suggests that (2.12) and the related GI/G/m approximation obtained by combining (2.12) with (2.14) below will exceed the true values. I compare the approximations in (2.12) and (2.10) plus (2.6) to the exact M/M/m values in Table 1. [The approximations are actually applied to EQ, shown to be equivalent by (2.2).] Both approximations perform remarkably well except in very light traffic (Table 1).

2.4 *The Heavy-Traffic Approximation*

Heavy-traffic limit theorems for the general GI/G/m model show that W is exponentially distributed as $\rho \to 1$ (Borovkov 1965; Kingman 1965; Iglehart and Whitt 1970; Köllerström 1974). The *simple heavy-traffic approximation* for the mean is

$$EW(\rho, c_a^2, c_s^2, m) \approx \frac{\rho}{m(1-\rho)}\,\frac{(c_a^2 + c_s^2)}{2}, \tag{2.13}$$

TABLE 1

A Comparison of Approximations with Exact Values for the Expected Queue Length (Excluding Customers in Service), EQ, in the M/M/m Model

		Number of Servers, m					
Traffic Intensity, ρ	Method	1	2	4	8	20	100
0.50	Exact	0.50	0.33	0.17	0.059	0.0037	
	Sak	0.50	0.37	0.22	0.106	0.0224	
	Half-Whitt	0.50	0.37	0.22	0.101	0.0146	
0.70	Exact	1.63	1.35	1.00	0.63	0.29	0.0011
	Sak	1.63	1.39	1.08	0.73	0.33	0.0090
	Half-Whitt	1.57	1.32	1.01	0.68	0.27	0.0035
0.80	Exact	3.20	2.84	2.39	1.83	1.02	0.079
	Sak	3.20	2.89	2.47	1.94	1.18	0.052
	Half-Whitt	3.09	2.76	2.34	1.83	1.07	0.108
0.90	Exact	8.10	7.67	7.09	6.31	4.96	1.95
	Sak	8.10	7.72	7.17	6.40	5.05	2.24
	Half-Whitt	7.92	7.51	6.94	6.20	4.91	2.01
0.95	Exact	18.1	17.6	16.9	16.0	14.3	9.6
	Sak	18.1	17.6	17.0	16.1	14.3	9.7
	Half-Whitt	17.8	17.4	16.7	15.8	14.2	9.6
0.98	Exact	48.0	47.5	46.8	45.9	44.0	38.0
	Sak	48.0	47.6	46.9	45.9	43.8	37.5
	Half-Whitt	47.8	47.3	46.6	45.6	43.7	37.8

The exact values come from (2.5) and (2.3); "Sak" is the Sakesagawa (1977) approximation obtained by combining (2.2) and (2.12), while the "Half-Whitt" approximation is obtained by combining (2.2), (2.6), and (2.10). Blank spaces in Tables 1–29 occur where data were not available.

which is exact for M/G/1. Significantly, this simple heavy-traffic approximation (which might be used with unrefined direct diffusion approximations) is a very poor approximation when m is large and the probability of delay $P(W > 0)$ is *not* nearly 1. I will exploit more general heavy-traffic limits in which $m \to \infty$ as $\rho \to 1$ (Halfin and Whitt 1981).

Example 2. To see how poor the simple direct heavy-traffic approximation in (2.13) can be, consider the M/M/m model with $m = 20$ and $\rho = 0.8$. The traffic intensity $\rho = 0.8$ might be considered high enough for a heavy-traffic approximation, but $P(W > 0) = 0.26$ (Table 13). The exact value of the mean queue length in this case is $EQ = 1.02$, while the approximations based on (2.12) and (2.10) are 1.18 and 1.07 (Table 1). In contrast, the simple heavy-traffic approximation based on (2.13) and (2.2) is $EQ \approx \rho^2/(1-\rho) = 3.20$, just as in M/M/1. In this case, the simple direct heavy-traffic approximation in (2.13) is off by a factor of more than 3.

A natural refined heavy-traffic approximation exploiting the exact M/M/m results is

$$EW(\rho, c_a^2, c_s^2, m) \approx \left(\frac{c_a^2 + c_s^2}{2}\right) EW(\mathrm{M/M/m}), \qquad (2.14)$$

where it is understood that $\tau = 1$ and $\lambda = m\rho$ in both cases. Formula (2.14) is exact for the M/M/m model, so the difficulty with (2.13) shown in Example 2 is removed. Moreover, both (2.13) and (2.14) are asymptotically correct as $\rho \to 1$ in the sense that the ratio of the two sides approaches 1. Set $c_a^2 = 1$ and (2.14) also becomes a standard approximation for the M/G/m model (Lee and Longton 1959; Hokstad 1978; Nozaki and Ross 1978). Allen refers to (2.14) as the Allen-Cunneen approximation (Allen 1990, p. 341). For the M/G/m model, (2.14) is usually an excellent approximation, even given extra information about the service-time distribution. However, improvements can be made considering light-traffic limits (Boxma, Cohen, and Huffels 1979; Burman and Smith 1983).

QNA uses approximation (2.14) for EW (Whitt 1983a). I include (2.14) in the numerical comparisons for EQ and EW in Tables 1–7 and 11 because (2.14) was used before and since it is of general interest as the natural first-order approximation. I refer it to there as *Heavy*. [*Heavy* refers to the refined approximation (2.14) rather than the simple direct heavy-traffic approximation formula (2.13).]

Now the object is to improve on (2.14). Using numerical comparisons with exact values, I discovered that (2.14) tends to overestimate EW when $c_a^2 < c_s^2$ and to underestimate EW when $c_s^2 < c_a^2 \leq 1$. This phenomenon can easily be appreciated by comparing EQ in the D/M/m, M/D/m, and $\mathrm{E_2/E_2/m}$ models (Tables 2, 3 and 6). Because $c^2 = k^{-1}$ for E_k, $c_a^2 + c_s^2 = 1$ in all three models, and approximation (2.14) is identical for these three systems, while

$$EQ(\mathrm{D/M/m}) \leq EQ(\mathrm{E_2/E_2/m}) \leq EQ(2.14) \leq EQ(\mathrm{M/D/m}), \qquad (2.15)$$

where $EQ(2.14)$ refers to the value based on formulas (2.2) and (2.14). For example, the exact values when $\rho = 0.8$ and $m = 4$ are 0.87, 1.06, 1.19 and 1.24, respectively. Constructing refinements is not difficult since the deviations of (2.14) from the exact values are consistent for all values of m and ρ.

The numerical values also show that (2.14) is much closer to the M/D/m value than the D/M/m value. This is consistent with the rough, practical guideline that nonstandard distributions (departures from the M assumption) have a greater impact on the system when applied to the interarrival times than when applied to the service times (Whitt 1984b, Table III).

2.5 *The D/M/m and M/D/m Models*

I apply the very accurate approximations for the M/D/m and D/M/m systems developed by Cosmetatos (1975) to obtain a better approximation. His approximations are

$$EW(\mathrm{M/D/m}) \approx \phi_1(m,\rho)EW(\mathrm{M/M/m})EW(\mathrm{M/D/1})/EW(\mathrm{M/M/1})$$

$$\approx \phi_1(m,\rho)\left(\frac{c_a^2+c_s^2}{2}\right)EW(\mathrm{M/M/m}), \quad (2.16)$$

where $(c_a^2 + c_s^2)/2 = \frac{1}{2}$, $\phi_1(m, \rho) = 1 + \gamma(m, \rho)$,

$$\gamma(m,\rho) = \min\{0.24, (1-\rho)(m-1)((4+5m)^{1/2}-2)/(16m\rho)\}; \quad (2.17)$$

$$EW(\mathrm{D/M/m}) = \phi_2(m,\rho)EW(\mathrm{M/M/m})EW(\mathrm{D/M/1})/EW(\mathrm{M/M/1}) \quad (2.18)$$

where $\phi_2(m, \rho) = 1 - 4\gamma(m, \rho)$ for $\gamma(m, \rho)$ in (2.17).

I have modified the Cosmetatos (1975) approximations by inserting the minimum with 0.24 in (2.16). Without this modification, $\gamma(m, \rho) \to \infty$ as $\rho \to 0$ for any positive m. Moreover, $\phi_2(m, \rho)$ becomes negative for $\gamma(m, \rho) > 0.25$. Considering stochastic comparisons, $\gamma(m, \rho)$ should be less than 1 in (2.16) (Whitt 1983b). However, I only use the adjustment in extreme cases.

I apply (2.16) directly, but (2.18) involves the somewhat complicated EW(D/M/1) value. In fact, $EW(\mathrm{D/M/1}) = \lambda^{-1}\sigma$, where σ is the unique root in the interval (0, 1) of the equation $x - e^{-(1-x)/\rho} = 0$, so EW(D/M/1) is evaluated without great difficulty. However, I use the Kraemer and Langenbach-Belz (1976) approximation (Whitt 1983a), formula (45),

$$EW(\mathrm{D/M/1})/EW(\mathrm{M/M/1}) \approx 2^{-1}\exp(-2(1-\rho)/3\rho). \quad (2.19)$$

Then I combine (2.18) and (2.19) to obtain

$$EW(\mathrm{D/M/m}) \approx \phi_3(m,\rho)\left(\frac{c_a^2+c_s^2}{2}\right)EW(\mathrm{M/M/m}) \quad (2.20)$$

where $\phi_3(m, \rho) = \phi_2(m, \rho)\exp(-2(1 - \rho)/3\rho)$. The accuracy of these approximations for the M/D/m and D/M/m systems is excellent (Tables 2 and 3). Significant improvement over approximation (2.14) is apparent in the D/M/m case. For example, when $\rho = 0.70$ and $m = 2$ in the D/M/m model, the exact and new values of EW are both 0.46, whereas approximation (2.14) yields 0.67.

2.6 *The General GI/G/m Model*

A natural way to obtain approximations for general GI/G/m models, at least when $c_a^2 \le 1$ and $c_s^2 \le 1$, is to interpolate from the M/D/m, D/M/m and M/M/m cases that can be well treated (Cosmetatos 1982; Page 1982). An interesting variant has also recently been proposed by Kimura (1986). These approximations are evidently aimed primarily at GI/G/m models with $c_a^2 \leq 1$ and $c_s^2 \leq 1$, because they can produce very bad (negative) approximations when $c_a^2 > 1$ or $c_s^2 > 1$. Although high precision may not be possible when c_a^2 or c_s^2 is large, my goal is to produce reasonable values over the full range.

The idea is first to focus on the case $c_a^2 = c_s^2$, then treat the cases $c_a^2 > c_s^2$ and $c_s^2 > c_a^2$. When $c_a^2 \geq 1$, $c_s^2 \geq 1$ and c_a^2 and c_s^2 are close, the old approximation (2.14) is reasonable, so use it when $c_a^2 = c_s^2 \geq 1$. When $c_a^2 = c_s^2 < 1$, the numerical values clearly indicate that approximation (2.14) is too large. I tried various methods, eventually deciding to exploit the excellent approximations for EW in the M/D/m and D/M/m models in a simple way: When $c_a^2 = c_s^2 = 0.5$, I linearly interpolate between M/D/m and D/M/m, and fit a smooth curve through the correction functions in the other cases. Let

$$\phi_4(m, \rho) = \min\{1, (\phi_1(m, \rho) + \phi_3(m, \rho))/2\} \qquad (2.21)$$

and

$$\Psi(c^2, m, \rho) = \begin{cases} 1, & c^2 \geq 1 \\ \phi_4(m, \rho)^{2(1-c^2)}, & 0 \leq c^2 \leq 1. \end{cases} \qquad (2.22)$$

When $c_a^2 = c_s^2 = c^2$, we let

$$EW(\rho, c^2, c^2, m) \approx \Psi(c^2, m, \rho)\left(\frac{c_a^2 + c_s^2}{2}\right) EW(\mathrm{M/M/m}). \qquad (2.23)$$

Table 6 contains the outcome of this procedure for the E_2/E_2/m model, the principal model of interest. The approximation is clearly excellent, except in the regions where both m is large and ρ is small, where the actual value tends to be negligible.

I now treat the general case of the pair (c_a^2, c_s^2) with $c_a^2 \neq c_s^2$ by modifying the approximation for the case $[(c_a^2 + c_s^2)/2, (c_a^2 + c_s^2)/2]$ determined in (2.23). For this purpose, I started with the correction factors $\phi_1(m, \rho)$ for EW(M/D/m) in (2.19) and $\phi_3(m, \rho)$ for EW(D/M/m) in (2.26), and modified them to improve the fit. The final approximation is

$$EW(\rho, c_a^2, c_s^2, m) \approx \phi(\rho, c_a^2, c_s^2, m)\left(\frac{c_a^2 + c_s^2}{2}\right) EW(\mathrm{M/M/m}), \qquad (2.24)$$

where

$$\phi(\rho, c_a^2, c_s^2, m) = \begin{cases} \left(\dfrac{4(c_a^2 - c_s^2)}{4c_a^2 - 3c_s^2}\right)\phi_1(m, \rho) + \left(\dfrac{c_s^2}{4c_a^2 - 3c_s^2}\right)\Psi((c_a^2 + c_s^2)/2, m, \rho), & c_a^2 \geq c_s^2 \\ \left(\dfrac{c_s^2 - c_a^2}{2c_a^2 + 2c_s^2}\right)\phi_3(m, \rho) + \left(\dfrac{c_s^2 + 3c_a^2}{2c_a^2 + 2c_s^2}\right)\Psi((c_a^2 + c_s^2)/2, m, \rho), & c_a^2 \leq c_s^2 \end{cases} \qquad (2.25)$$

with $\Psi(c^2, m, \rho)$ in (2.22). Note that ϕ in (2.25) reduces to Ψ in (2.22) when $c_a^2 = c_s^2$, so that (2.24) then agrees with (2.23). Also ϕ in (2.25) reduces to ϕ_1 when $c_s^2 = 0$ and ϕ_3 when $c_a^2 = 0$. Hence, (2.24) coincides with the approximations in

Section 2.5 for $EW(\mathrm{M/D/m})$ and $EW(\mathrm{D/M/m})$. Formula (2.24) is also a continuous function of the parameters $(\rho,\ c_a^2,\ c_s^2)$. For inverse problems, any positive number can be realized by EW in (2.24) by changing only ρ or $c_a^2 + c_s^2$.

2.7 *Numerical Comparisons*

I present a representative set of tables comparing the approximations with exact values. Whitt (1985) presented additional tables. Before discussing these tables in detail, it is worth commenting on how we evaluate the quality of the approximations.

There are two standard ways to measure the quality of queueing approximations: absolute difference and relative percentage error. I contend that neither procedure alone is usually suitable over the entire range of values. I am usually quite satisfied if *either* the absolute difference is below a critical threshold *or* the relative percentage error is below another critical threshold. Thus a final *adjusted measure of error* (*AME*) might be

$$AME = \min\{A|exact - approx.|,\ 100(|exact - approx.|)/exact\}, \quad (2.26)$$

where A is a constant chosen in each instance to reflect the relative importance of absolute difference and the relative percentage of error.

I cannot choose a single constant A for the AME in advance for all descriptive characteristics, because the practical meaning of absolute differences changes. For one example, the expected waiting time EW depends on the time units. For another, the expected queue length EQ is always λ times the expected waiting time EW, by virtue of (2.2), so that even if I fix the time units by setting $\tau = 1$, $EQ = m\rho EW$. For large m and given ρ, EQ is much larger than EW. The constant A for EQ could be about m times the constant A for EW.

Although I do not display the calculations of any specific adjusted measures of error, my discussion explains the goals. Either the relative percent error or the absolute difference should be small. As a consequence, no great concern exists over large relative percentage errors in light traffic. For example, the case $m = 20$ and $\rho = 0.50$ is of relatively little concern, and when $m = 100$ and $\rho = 0.50$, the values are usually too small to record.

Tables 2–5 compare the new approximation of the expected queue length EQ, combining (2.2) and (2.24), with the old heavy-traffic approximation, combining (2.2) and (2.14), and exact values from Kühn (1976). The D/M/m and M/D/m approximations (Tables 2 and 3) reduce to the slightly modified Cosmetatos (1975) approximations in Section 2.5, known to be excellent.

I compare the same approximations with exact values of the expected queue length in the $\mathrm{H_2/M/m}$ model (Table 4). The H_2 interarrival-time distribution has density

$$f(x) = p_1\lambda_1 e^{-\lambda_1 x} + p_2\lambda_2 e^{-\lambda_2 x}, \qquad x \geq 0, \quad (2.27)$$

where $p_1, p_2, \lambda_1, \lambda_2 \geq 0$ and $p_1 + p_2 = 1$, so that there are three parameters. Given the mean λ^{-1} and c_a^2, there is thus one remaining degree of freedom, specified by the proportion of the total mean in the component with the smaller mean, and defined by

$$r = (p_1\lambda_1^{-1})/(p_1\lambda_1^{-1} + p_2\lambda_2^{-1}), \quad (2.28)$$

where $\lambda_1 > \lambda_2$ (Whitt 1984b, Section V). The H_2 interarrival-time distribution is specified by the traffic intensity ρ, $c_a^2 = 2.25$ and $r = \frac{1}{2}$. The case $r = \frac{1}{2}$ is often

TABLE 2

A Comparison of Approximations of the Expected Queue Length (Excluding Customers in Service) with Exact Values from Kühn (1976) for the D/M/m Model

		Number of Servers, m				
Traffic Intensity, ρ	Method	2	4	8	20	100
0.50	Exact	0.065	0.021	0.0027	0	0
	Heavy	0.17	0.09	0.030	0.002	0
	New	0.067	0.020	0	0	0
0.70	Exact	0.46	0.29	0.134	0.015	0
	Heavy	0.67	0.50	0.32	0.11	0.0005
	New	0.46	0.29	0.133	0.014	0
0.80	Exact	1.14	0.87	0.58	0.23	0.002
	Heavy	1.42	1.19	0.92	0.51	0.039
	New	1.14	0.87	0.58	0.22	0
0.90	Exact	3.4	3.1	2.6	1.8	0.41
	Heavy	3.8	3.5	3.2	2.5	0.98
	New	3.5	3.1	2.6	1.8	0.40
0.95	Exact	8.4	8.0	7.4	6.3	3.6
	Heavy	8.8	8.5	8.0	7.2	4.8
	New	8.4	8.0	7.3	6.2	3.4
0.98	Exact	23.3	22.9	22.2	20.9	16.9
	Heavy	23.8	23.4	22.9	22.0	19.0
	New	23.3	22.8	22.2	20.8	16.8

The New approximation is (2.2) plus (2.24). In this case (2.24) reduces to (2.20). Approximation Heavy is (2.2) and (2.14).

TABLE 3

A Comparison of Approximations of the Expected Queue Length (Excluding Customers in Service) with Exact Values from Kühn (1976) for the M/D/m Model

		Number of Servers, m				
Traffic Intensity, ρ	Method	2	4	8	20	100
0.50	Exact	0.177	0.099	0.037	0.0028	0
	Heavy	0.167	0.087	0.030	0.0019	0
	New	0.176	0.099	0.037	0.0028	0
0.70	Exact	0.69	0.53	0.35	0.134	0.0009
	Heavy	0.67	0.50	0.32	0.109	0.00054
	New	0.69	0.53	0.35	0.132	0.00083
0.80	Exact	1.44	1.24	0.97	0.57	0.054
	Heavy	1.42	1.19	0.92	0.51	0.039
	New	1.44	1.23	0.97	0.57	0.052
0.90	Exact	3.86	3.60	3.24	2.60	1.11
	Heavy	3.84	3.54	3.16	2.48	0.98
	New	3.86	3.60	3.25	2.61	1.11
0.95	Exact	8.8	8.5	8.1	7.3	5.1
	Heavy	8.8	8.5	8.0	7.2	4.8
	New	8.8	8.5	8.1	7.4	5.1
0.98	Exact	23.8	23.5	23.0	22.2	18.9
	Heavy	23.8	23.4	22.9	22.0	19.0
	New	23.8	23.5	23.0	22.2	19.5

The cases $m = 20$ and 100 with $\rho = 0.98$ come from Seelen, Tijms and van Hoorn (1985).

TABLE 4

A Comparison of Approximations of the Expected Queue Length (Excluding Customers in Service) with Exact Values from Kühn (1976) for the $H_2/M/m$ Model (Hyperexponential Interarrival-Time Distribution with Balanced Means) with $c_a^2 = 2.25$

		Number of Servers, m				
Traffic Intensity, ρ	Method	2	4	8	20	100
0.50	Exact	0.57	0.35	0.157	0.021	0
	Heavy	0.54	0.28	0.096	0.006	0
	New	0.57	0.31	0.116	0.009	0
0.70	Exact	2.27	1.83	1.30	0.59	0.015
	Heavy	2.19	1.63	1.03	0.35	0.002
	New	2.23	1.70	1.12	0.42	0.003
0.80	Exact	4.75	4.17	3.42	2.22	0.36
	Heavy	4.62	3.88	2.97	1.66	0.13
	New	4.67	3.99	3.13	1.83	0.16
0.90	Exact	12.7	11.9	10.9	9.1	4.6
	Heavy	12.5	11.5	10.3	8.1	3.2
	New	12.5	11.7	10.5	8.4	3.5
0.95	Exact	28.8	28.0	26.9	24.7	17.0
	Heavy	28.6	27.5	26.1	23.3	15.6
	New	28.6	27.7	26.4	23.8	16.5
0.98	Exact	77.5	76.7	75.4	73.0	65.2
	Heavy	77.2	76.1	74.5	71.4	61.7
	New	77.3	76.3	74.9	72.0	63.0

TABLE 5

A Comparison of Approximations of the Expected Queue Length (Excluding Customers in Service) with Exact Values from Kühn (1976) for the $E_4/M/m$ Model (Erlang Interarrival-Time Distribution with Four Phases) with $c_a^2 = 0.25$

		Number of Servers, m				
Traffic Intensity, ρ	Method	2	4	8	20	100
0.50	Exact	0.122	0.047	0.009	0.00011	0
	Heavy	0.213	0.113	0.038	0.0025	0
	New	0.140	0.065	0.018	0.0013	0
0.70	Exact	0.67	0.44	0.24	0.050	0
	Heavy	0.84	0.63	0.40	0.138	0.0006
	New	0.69	0.48	0.28	0.075	0.0004
0.80	Exact	1.55	1.23	0.86	0.39	0.0089
	Heavy	1.78	1.49	1.15	0.64	0.049
	New	1.58	1.28	0.92	0.45	0.025
0.90	Exact	4.52	4.08	3.51	2.55	0.71
	Heavy	4.80	4.43	3.95	3.10	1.23
	New	4.55	4.13	3.58	2.67	0.85
0.95	Exact	10.7	10.2	9.5	8.2	4.9
	Heavy	11.0	10.6	10.0	9.0	6.0
	New	10.7	10.2	9.6	8.4	5.1
0.98	Exact	29.4	28.8	28.1	26.6	22.1
	Heavy	29.8	29.3	28.6	27.5	23.8
	New	29.4	28.9	28.2	26.7	22.4

The case $m = 100$ and $\rho = 0.95$ comes from Seelen, Tijms and van Hoorn (1985).

referred to as *balanced means*; it produces an H_2 distribution approximately in the middle of the range for given first two moments. The fit in Table 4 is pretty good for both approximations, but they tend to understate the exact value when m is large and ρ is small. The fit is close considering the set of possible values associated with the parameter pair $(c_a^2, c_s^2) = (2.25, 1)$. This is demonstrated for related cases later (Tables 8–10). Whitt (1984a, 1984b), Klincewicz and Whitt (1984) and Johnson and Taaffe (1991) provide related background.

When evaluating the approximations, especially for large m, it is important to take into account that the approximation displayed here is for the expected queue length. By (2.2), $EW = (EQ)/(\rho m)$, so that for large m, EW is much less than EQ. For example, when $\rho = 0.80$ and $m = 100$ in the $H_2/M/m$ model, the exact value is $EQ = 0.36$ (Table 4), while $EW = 0.0045$, usually negligible for practical purposes. Hence, the large relative error in this case is not usually too important.

I also make numerical comparisons of the expected queue length in a GI/M/m model with $c_a^2 < 1$, in particular, the $E_4/M/m$ model (Table 5). The new approximation based on (2.24) also works quite well for these GI/M/m models with $0 < c_a^2 < 1$, clearly better than the old heavy-traffic approximate based on (2.14). Unlike the approximation in Table 4, the new approximation in Table 5 overestimates the exact values when both m is large and ρ is small. In all cases, the departure from an exponential interarrival-time distribution actually has a stronger effect than the approximation predicts in this region. In all cases, too however, the change from the heavy-traffic approximation (2.14) to (2.24) moves in the right direction.

I compare the approximations with exact value of expected queue length from tables by Hillier and Yu (1981) for $E_2/E_2/m$ systems (Table 6). Hillier and Yu only supplied exact values for $m = 2$, 4 and 8 because the cases in $m = 20$ and 100 were not available. The data for the other cases came from Seelen, Tijms, and van Hoorn (1985). The new approximation (2.24) again performs well.

De Smit (personal communication) ran a large set of numerical experiments based on his algorithm for $G/H_2/m$ models (de Smit 1983a, 1983b). I use these numerical results to compare approximations for the expected waiting time with exact values in $G/H_2/m$ models for the cases $m = 2$ and 4, and $c_s^2 = 9.0$ (Table 7). The exact values are for H_2 service-time distributions with balanced means, or $r = \frac{1}{2}$ for r in (2.28). The interarrival-time distribution is D, E_2 and H_2 $(r = \frac{1}{2})$ when $c_a^2 = 0.0$, 0.5, 2.0 and 9.0, respectively. In the cases $c_a^2 = c_s^2 = 9.0$, the new approximation (2.24) coincides with the heavy-traffic approximation (2.14). In other cases, the new approximation usually is better, although it is not uniformly better.

You can assess the set of possible exact values associated with given parameter values with the help of Tables 8–10. The values displayed there only constitute a subset of all possible values, so that the actual set is much larger. It is not difficult to show that the set of all values is connected, so that all values between the maximum and minimum (or supremum and infimum) are attainable. Hence, to describe the set it suffices to focus on the minimum and maximum values. The subset here is obtained by considering, for any value of $c^2 > 1$, H_2 distributions with three different values of r in (2.28), namely, $r = \frac{1}{9}$, $r = \frac{1}{2}$ and $r = \frac{8}{9}$. From numerical experience, I believe that this range shows the actual range of typical values. The values for $H_2/H_2/m$ models when $\rho = 0.80$ and $m = 2$ and 4 are in Table 8. The parameter r is allowed to vary for both the interarrival-time distribution and the service-time distribution, so that there are nine cases associated with each parameter pair (c_a^2, c_s^2)

TABLE 6

A Comparison of Approximations of the Expected Queue Length (Excluding Customers in Service) with Exact Values from Hillier and Yu (1981) and Seelen, Tijms and van Hoorn (1985) for the $E_2/E_2/m$ Model

		Number of Servers, m				
Traffic Intensity, ρ	Method	2	4	8	20	100
0.50	Exact	0.118	0.051	0.0123	0.0003	0
	Heavy	0.166	0.087	0.030	0.002	0
	New	0.121	0.060	0.0185	0.0014	0
0.70	Exact	0.58	0.40	0.23	0.058	
	Heavy	0.67	0.50	0.32	0.11	0.0005
	New	0.57	0.41	0.24	0.073	0.00041
0.80	Exact	1.30	1.06	0.78	0.38	0.02
	Heavy	1.42	1.19	0.92	0.51	0.039
	New	1.29	1.05	0.78	0.40	0.026
0.90	Exact	3.69	3.36	2.94	2.20	0.70
	Heavy	3.84	3.54	3.16	2.48	0.98
	New	3.67	3.35	2.92	2.21	0.76
0.95	Exact	8.63	8.26	7.76	6.81	4.24
	Heavy	8.79	8.47	8.02	7.18	4.81
	New	8.61	8.24	7.73	6.79	4.27
0.98	Exact	23.6	23.2	22.6	21.6	18.2
	Heavy	23.8	23.4	22.9	22.0	19.0
	New	23.6	23.2	22.6	21.5	18.1

(Table 8). As noted by Whitt (1984a, 1984b) and Klincewicz and Whitt (1984), the range (maximum minus minimum) for the single-server case, increases as c_a^2 or c_s^2 increases: greater variability not only means longer expected waiting times, but also less reliable approximations for partially specified systems. As Whitt (1984b, Table III) found, changing r_a has a much greater impact than changing r_s (Table 8). However, the range is not exceptionally large when $c_a^2 = 2.0$. When c_a^2 becomes 9.0 or higher, these two-moment approximations become relatively unreliable.

Table 9 shows the corresponding observed maxima and minima in GI/H_2/m models for $r_s = \frac{1}{9}, \frac{1}{2}$ and $\frac{8}{9}$ for interarrival times with $c_a^2 = 0$ (D) and $c_a^2 = 0.5$ (E_2) for three values of the traffic intensity. Table 10 then compares the heavy-traffic approximations (2.14) and the new approximation (2.24) for the expected waiting time with the sets of exact values for the cases in Tables 8 and 9 with $\rho = 0.80$. In each case the set of exact values is summarized by giving the minimum, median and maximum values. The new approximation values based on (2.24) are clearly an improvement over the heavy-traffic values based on (2.14) (Table 10). In general, the new approximations seem adequate, although changes to produce slightly further reductions in the approximations may provide better accuracy when $c_a^2 < 1$.

Table 11 compares the approximations for the expected queue length with exact values from Seelen, Tijms, and van Hoorn (1985) for a model with $c_a^2 > 1 > c_s^2$, in particular, for the H_2/D/m model with $c_a^2 = 2.0$. The new approximation in (2.24) produces slightly larger values than the old approximation (Table 11). Further modifications in this direction in this region for high m, but in the other direction for small m and ρ, could improve the new approximation.

TABLE 7

A Comparison of Approximations of the Expected Waiting Time (Excluding Service Time) with Exact Values from de Smit (1983a, 1983b, personal communication) for the G/H_2/m Model with $c_s^2 = 9.0$

Number of Servers, m	Traffic Intensity, ρ	Method	Arrival Process Variability Parameter			
			$c_a^2 = 0.0$	$c_a^2 = 0.5$	$c_a^2 = 2.0$	$c_a^2 = 9.0$
2	0.30	Exact	0.23	0.29	0.46	0.91
		Heavy	0.45	0.47	0.54	0.89
		New	0.25	0.28	0.39	0.89
	0.60	Exact	1.95	2.20	2.86	5.28
		Heavy	2.53	2.67	3.09	5.06
		New	1.96	2.13	2.65	5.06
	0.80	Exact	7.20	7.79	9.46	16.47
		Heavy	8.00	8.44	9.78	16.00
		New	7.20	7.69	9.16	16.00
	0.90	Exact	18.3	19.5	23.1	39.0
		Heavy	19.2	20.3	23.4	38.4
		New	18.3	19.4	22.7	38.4
4	0.30	Exact	0.014	0.022	0.050	0.143
		Heavy	0.060	0.063	0.073	0.037
		New	0.037	0.035	0.050	0.037
	0.60	Exact	0.47	0.56	0.82	1.82
		Heavy	0.81	0.85	0.99	1.61
		New	0.57	0.63	0.80	1.61
	0.80	Exact	2.72	3.00	3.79	7.15
		Heavy	3.35	3.54	4.10	6.71
		New	2.90	3.12	3.75	6.71
	0.90	Exact	8.08	8.69	10.5	18.3
		Heavy	8.86	9.35	10.8	17.7
		New	8.30	8.82	10.4	17.7

The interarrival-time distributions are deterministic (D) when $c_a^2 = 0.0$, Erlang (E_2) when $c_a^2 = 0.50$ and hyperexponential (H_2) with balanced means when $c_a^2 > 1$. The service-time distribution is hyperexponential with balanced means, having overall mean one.

3. The Probability of Delay

In this section I develop approximations for the probability of delay $P(W > 0)$. I begin in Section 3.1 by relating $P(W > 0)$ to $P(N \geq m)$ and by indicating how our approximation for $P(W > 0)$ can be used to generate an associated approximation for $P(N \geq m)$. In Section 3.2 I describe my basic strategy and give background on the GI/G/1 and GI/M/s/0 models. In Section 3.3 I develop an approximation for $P(W > 0$ in the GI/M/m model, and in Section 3.4 I extend it to GI/G/m. I make numerical comparisons with exact values in Section 3.5.

My approximations for the probability of delay yield new approximations for the expected delay, by incorporating approximations for the conditional expected delay given that a customer must wait, ED, proposed by Seelen and Tijms (1984). On the other hand, my two approximations yield alternative approximations for the expected conditioned delay by $ED = EW/P(W > 0)$.

3.1 *Two Related Congestion Measures*

Here I focus on probability of delay, $P(W > 0)$, or the probability that an arriving customer must wait before beginning service (in steady state). Distinguish this from

TABLE 8

The Range of Exact Values of the Expected Waiting Time (Excluding Service Time) for the $H_2/H_2/m$ Model with Traffic Intensity $\rho = 0.80$ and Mean Service Time 1, from de Smit (1983a, 1983b, personal communication)

	Arrival Parameters		Service-Time Parameters								
			$c_s^2 = 2.0$			$c_s^2 = 5.0$			$c_s^2 = 9.0$		
Number of Servers, m	c_a^2	r_a	$r_s = \frac{1}{9}$	$r_s = \frac{1}{2}$	$r_s = \frac{8}{9}$	$r_s = \frac{1}{9}$	$r_s = \frac{1}{2}$	$r_s = \frac{8}{9}$	$r_s = \frac{1}{9}$	$r_s = \frac{1}{2}$	$r_s = \frac{8}{9}$
2	2.0	$\frac{1}{9}$	3.78	3.75	3.72	6.42	6.27	6.05	9.93	9.57	9.09
		$\frac{1}{2}$	3.64	3.61	3.56	6.33	6.15	5.90	9.87	9.46	8.95
		$\frac{8}{9}$	3.32	3.28	3.22	6.11	5.91	5.60	9.71	9.26	8.67
	5.0	$\frac{1}{9}$	7.30	7.29	7.27	9.92	9.82	9.65	13.4	13.2	12.7
		$\frac{1}{2}$	6.55	6.49	6.43	9.39	9.13	8.84	13.0	12.5	11.9
		$\frac{8}{9}$	4.25	4.20	4.11	7.56	7.30	6.80	11.6	11.0	10.1
	9.0	$\frac{1}{9}$	12.1	12.1	12.1	14.7	14.6	14.5	18.1	18.0	17.6
		$\frac{1}{2}$	10.4	10.3	10.2	13.3	13.0	12.7	17.1	16.5	15.8
		$\frac{8}{9}$	4.8	4.8	4.7	8.6	8.3	7.7	13.1	12.5	11.3
4	2.0	$\frac{1}{9}$	1.62	1.60	1.56	2.71	2.58	2.30	4.15	3.84	3.19
		$\frac{1}{2}$	1.56	1.54	1.49	2.67	2.52	2.24	4.12	3.79	3.12
		$\frac{8}{9}$	1.42	1.39	1.33	2.57	2.40	2.09	4.05	3.69	2.97
	5.0	$\frac{1}{9}$	3.30	3.29	3.27	4.35	4.28	4.08	5.78	5.57	5.07
		$\frac{1}{2}$	2.95	2.92	2.88	4.10	3.95	3.71	5.59	5.26	4.68
		$\frac{8}{9}$	1.87	1.83	1.76	3.26	3.06	2.67	4.93	4.50	3.67
	9.0	$\frac{1}{9}$	5.61	5.62	5.61	6.61	6.59	6.47	8.00	7.90	7.50
		$\frac{1}{2}$	4.81	4.77	4.74	5.99	5.80	5.61	7.51	7.15	6.66
		$\frac{8}{9}$	2.15	2.11	2.03	3.78	3.56	3.10	5.68	5.21	4.24

TABLE 9

The Range of Exact Values of the Expected Waiting Time (Excluding Service Time) for the $G/H_2/m$ Model, from de Smit (1983a, 1983b, personal communication)

			Service-Time Parameters								
			$c_s^2 = 2.0$			$c_s^2 = 5.0$			$c_s^2 = 9.0$		
Number of Servers, m	Arrival Variability Parameter, c_a^2	Traffic Intensity, ρ	$r_s = \frac{1}{9}$	$r_s = \frac{1}{2}$	$r_s = \frac{8}{9}$	$r_s = \frac{1}{9}$	$r_s = \frac{1}{2}$	$r_s = \frac{8}{9}$	$r_s = \frac{1}{9}$	$r_s = \frac{1}{2}$	$r_s = \frac{8}{9}$
2	0.0	0.6	0.39	0.37	0.34	1.17	1.05	0.81	2.25	1.95	1.39
		0.8	1.55	1.52	1.49	4.14	3.96	3.77	7.64	7.20	6.78
		0.9	4.01	3.98	3.95	10.33	10.13	9.97	18.8	18.3	18.0
	0.5	0.6	0.60	0.58	0.55	1.40	1.28	1.05	2.49	2.20	1.65
		0.8	2.09	2.07	2.03	4.71	4.54	4.32	8.21	7.79	7.34
		0.9	5.19	5.16	5.13	11.5	11.3	11.2	20.0	19.5	19.2
4	0.0	0.6	0.100	0.090	0.069	0.33	0.25	0.13	0.65	0.47	0.20
		0.8	0.61	0.58	0.53	1.67	1.50	1.17	3.10	2.72	1.96
		0.9	1.80	1.76	1.70	4.69	4.48	4.09	8.55	8.08	7.18
	0.5	0.6	0.174	0.161	0.139	0.42	0.34	0.21	0.74	0.56	0.28
		0.8	0.85	0.82	0.77	1.92	1.77	1.45	3.36	3.00	2.27
		0.9	2.36	2.33	2.27	5.26	5.07	4.69	9.13	8.69	7.82

The cases of D, E_2 and M arrival processes. The mean service time is 1 in each case.

TABLE 10

A Comparison of Approximations of the Expected Waiting Time (Excluding Service Time) with the Exact Values from Tables 8 and 9 for the G/H_2/m Model with Traffic Intensity $\rho = 0.80$

	Variability Parameters		Exact Values				
Number of Servers, m	c_a^2	c_s^2	Min	Median	Max	Heavy	New
2	0.0	2.0	1.49	1.52	1.55	1.78	1.60
	0.5		2.03	2.07	2.09	2.22	2.09
	2.0		3.22	3.61	3.78	3.56	3.56
	9.0		4.7	10.3	12.1	9.78	9.90
	0.0	9.0	6.78	7.20	7.64	8.00	7.20
	0.5		7.34	7.79	8.21	8.44	7.69
	2.0		8.67	9.46	9.93	9.78	9.16
	9.0		11.3	16.5	18.1	16.0	16.0
4	0.0	2.0	0.53	0.58	0.61	0.75	0.64
	0.5		0.77	0.82	0.85	0.93	0.86
	2.0		1.33	1.54	1.62	1.49	1.49
	9.0		2.03	4.77	5.61	4.10	4.23
	0.0	9.0	1.96	2.72	3.10	3.35	2.90
	0.5		2.27	3.00	3.36	3.54	3.12
	2.0		2.97	3.79	4.15	4.10	3.75
	9.0		4.24	7.15	8.00	6.71	6.71

The interarrival-time distribution is deterministic (D) when $c_a^2 = 0.0$, Erlang (E_2) when $c_a^2 = 0.5$ and hyperexponential (H_2) when $c_a^2 > 1$.

TABLE 11

A Comparison of Approximations of the Expected Queue Length (Excluding Customers in Service) with Exact Values for the H_2/D/m Model (Having Hyperexponential Interarrival Times with Balanced Means) with $c_a^2 = 2.0$, from Seelen, Tijms and van Hoorn (1985)

		Number of Servers, m					
Traffic Intensity, ρ	Method	1	2	4	8	20	100
0.50	Exact	0.41	0.32	0.22	0.113	0.021	0
	Heavy	0.50	0.33	0.17	0.059	0.0037	0
	New	0.50	0.35	0.20	0.074	0.0055	0
0.70	Exact	1.48	1.33	1.13	0.86	0.47	0.023
	Heavy	1.63	1.35	1.00	0.63	0.22	0.0011
	New	1.63	1.38	1.06	0.70	0.26	0.0017
0.80	Exact	3.02	2.85	2.58	2.22	1.57	0.364
	Heavy	3.20	2.84	2.39	1.83	1.02	0.079
	New	3.20	2.88	2.47	1.95	1.15	0.103
0.90	Exact	7.9	7.7	7.4	6.9	6.0	3.4
	Heavy	8.1	7.7	7.1	6.3	5.0	2.0
	New	8.1	7.7	7.2	6.5	5.2	2.2
0.95	Exact	17.9	17.6	17.3	16.7	15.6	12.2
	Heavy	18.1	17.6	16.9	16.0	14.4	9.6
	New	18.1	17.6	17.1	16.3	14.7	10.3
0.98	Exact	47.8	47.6	47.2	46.6	45.4	41.4
	Heavy	48.0	47.5	46.8	45.9	44.0	38.0
	New	48.0	47.6	47.0	46.1	44.4	39.0

the probability that all servers are busy at an arbitrary time, $P(N \geq m)$, the steady-state probability for the associated continuous-time stochastic process. Because Poisson arrivals see times averages, (Wolff 1982), these two congestion measures coincide in M/G/m models, so that our approximations for $P(W > 0)$ are then equally good for $P(N \geq m)$, but for non-Poisson arrival processes this is not the case. Otherwise, I focus on $P(W > 0)$.

Kraemer and Langenbach-Belz (1976) developed a good approximation for $P(W > 0)$ when $m = 1$ [Whitt 1983a, Formula (48)]. Also $P(N \geq m) = \rho$ (exact) when $m = 1$. In cases $m = 1$, $m = 10$ and $m = 100$, $P(W > 0)/P(N \geq m)$ does not change greatly with m (Table 12), so that the previously developed approximation for $P(W > 0)$ when $m = 1$ can yield an approximation for the ratio $P(W > 0)/P(N \geq m)$ for GI/G/m models when $m \geq 1$. Hence, with this ratio approximation, my approximation for $P(W > 0)$ yields an additional approximation for $P(N \geq m)$ in GI/G/m models.

3.2 *The Basic Strategy*

As in Section 2, my strategy is to build on exact results for the M/M/m queue, namely, the Erlang-C formula in (2.3). I can also exploit the more elementary closed-form approximations (2.10) and (2.12) plus (2.6). Table 13 compares approximation (2.10) with (2.3) and the normal approximation (2.11). Obviously the normal approximation only tends to be even nearly reasonable for small ρ.

We have accumulated considerable experience with M/G/m models, and it is known that the Erlang-C formula in (2.3) is usually an excellent approximation for

TABLE 12

The Exact Values of the Ratio P (W > 0)/P (N ≥ m) for Several GI/G/m Queues from Seelen, Tijms and van Hoorn (1985)

ρ	c_a^2	m	$c_s^2 = 0.0$	$c_s^2 = 1.0$	$c_s^2 = 2.5$
$\rho = 0.5$	$c_a^2 = 0.25$	1	0.326	0.602	0.674
		10	0.25	0.602	0.673
	$c_a^2 = 2.0$	1	1.23	1.18	1.16
		10	1.20	1.19	1.18
	$c_a^2 = 4.0$	1	1.43	1.27	1:32
		10	1.39	1.38	1.35
$\rho = 0.7$	$c_a^2 = 0.25$	1	0.586	0.790	0.827
		10	0.519	0.790	0.835
	$c_a^2 = 2.0$	1	1.14	1.11	1.09
		10	1.13	1.11	1.10
	$c_a^2 = 4.0$	1	1.26	1.21	1.18
		10	1.23	1.21	1.21
$\rho = 0.9$	$c_a^2 = 0.25$	1	0.818	0.931	0.948
		10	0.816	0.936	0.953
		100	0.828	0.932	
	$c_a^2 = 2.0$	1	1.05	1.03	1.03
		10	1.04	1.03	1.03
		100	1.03	1.04	
	$c_a^2 = 4.0$	1	1.08	1.06	1.06
		10	1.06	1.05	1.06
		100	1.06	1.06	

TABLE 13

A Comparison of Exact Values of the Probability of Delay, P (W > 0), in the M/M/m Model in (2.3) with Approximations Based on Heavy-Traffic Limits

		Number of Servers, m					
Traffic Intensity, ρ	Method	1	2	4	8	20	100
0.50	Exact	0.50	0.33	0.174	0.059	0.0037	0
	Half-Whitt	0.51	0.37	0.223	0.101	0.0146	0
	Normal	0.50	0.31	0.144	0.040	0.0013	0
0.70	Exact	0.70	0.58	0.43	0.27	0.094	0.0005
	Half-Whitt	0.67	0.56	0.43	0.29	0.117	0.0015
	Normal	0.59	0.47	0.34	0.21	0.071	0.0002
0.80	Exact	0.80	0.71	0.60	0.46	0.26	0.020
	Half-Whitt	0.77	0.69	0.58	0.46	0.27	0.027
	Normal	0.63	0.53	0.43	0.33	0.19	0.015
0.90	Exact	0.90	0.85	0.79	0.70	0.55	0.22
	Half-Whitt	0.88	0.83	0.77	0.69	0.55	0.22
	Normal	0.66	0.59	0.52	0.46	0.36	0.16
0.95	Exact	0.950	0.926	0.89	0.84	0.76	0.506
	Half-Whitt	0.939	0.914	0.88	0.83	0.75	0.504
	Normal	0.678	0.614	0.56	0.51	0.45	0.32
0.98	Exact	0.980	0.970	0.956	0.936	0.897	0.78
	Half-Whitt	0.975	0.965	0.951	0.931	0.892	0.77
	Normal	0.686	0.629	0.584	0.548	0.509	0.44

The Half-Whitt approximation is (2.10), while the normal approximation is (2.11).

nonexponential service-time distributions. Hence, I use this same approximation for the M/G/m case. I do not present tables in this case, because the good quality of the approximations is relatively well known.

Much work has yet to be done on approximations for either $P(W > 0)$ or $P(N \geq m)$ in GI/G/m models with non-Poisson arrivals. In the way of background, it is useful to consider what is known about the GI/G/1 special case and the associated GI/M/m loss system.

Kraemer and Langenbach-Belz (1976) provided the good approximation for $P(W > 0)$ in the GI/G/1 model used by Whitt (1983a). Whitt (1984d) also studied the probability of delay in the GI/G/1. As should be expected, their analysis supports having approximations with the property that $P(W_1 > 0) \leq P(W_2 > 0)$ for two systems that differ only in the arrival squared coefficients of variation c_{ai}^2 when $c_{a1}^2 \leq c_{a2}^2$. However, Whitt (1984d, Theorem 2) showed that $P(W > 0)$ in an H_2/G/1 model tends to be *decreasing* in c_s^2, with his Example 5 showing that the difference between $P(W > 0)$ in the H_2/D/1 and H_2/M/1 models with the same interarrival-time distribution is very small. Significantly, Kraemer and Langenbach-Belz's GI/G/1 approximation is consistent with these theoretical results.

We have also accumulated considerable experience with the GI/M/m loss system, and the blocking probability (call congestion) can be quite different from the probability all servers are busy $P(N = m)$ (time congestion). Fredericks (1983) and Sanders and van Doorn (1985) have developed good approximations in the case of overflow arrival processes. In heavy traffic,

$$P(Blocking) \approx zP(N = m) \tag{3.1}$$

where z is the peakedness factor of the arrival process, which approaches $(1 + c_a^2)/2$ as λ increases; see (3.6) below and Whitt (1984e, 1992). Moreover, a normal distribution approximation is often appropriate. However, it seems difficult to develop good approximations for the GI/M/m loss system solely in terms of the basic parameter 5-tuple $(\lambda,\ c_a^2,\ \tau,\ c_s^2,\ m)$. The upper boundary evidently makes the loss model more difficult analyze and to approximate. The behavior of the loss model seems to depend more on the full distributions of the interarrival times and service times.

3.3 *GI/M/m Models*

The key theoretical result supporting my approximation for GI/M/m models is the heavy-traffic limit theorem for the probability of delay in these models (Halfin and Whitt 1981, Theorem 4). Halfin and Whitt show that approximation (2.10) for M/M/m models is again asymptotically correct for GI/M/m models as $m \to \infty$ and $\rho \to 1$ with $(1 - \rho)m^{1/2} \to \beta$ if β in (2.10) is replaced by $\beta_G = 2\beta/(1 + c_a^2)$. For GI/M/m models, let

$$HW(c_a^2) = \xi(2(1 - \rho)m^{1/2}/(1 + c_a^2)) \tag{3.2}$$

with $\xi(\beta)$ defined in (2.10).

An obvious way to use (3.2) with the exact M/M/m formula (2.3) is in the ratio

$$P(W(c_a^2) > 0) \approx \min\{1, (HW(c_a^2)/HW(1))P(W(\text{M/M/m}) > 0)\}, \tag{3.3}$$

where $W(c_a^2)$ denotes the waiting time in the GI/M/m model as a function of c_a^2. Using the ratio approximation in (3.3) instead of (3.2), I obtain the exact value from the M/M/m model when $c_a^2 = 1$.

I compared the approximations (3.2) and (3.3), plus a number of related approximations for a family of GI/M/m models. The best approximation of this kind was a ratio approximation using a lower bound for $HW(c_a^2)$ (Halfin and Whitt 1981, Remark 1, p. 575). This approximation, the lower-bound or LB-ratio, is defined by

$$P(W(c_a^2) > 0)$$
$$\approx \min\left\{1, \frac{1 - \Phi(2(1 - \rho)m^{1/2}/(1 + c_a^2))}{1 - \Phi((1 - \rho)m^{1/2})} P(W(\text{M/M/m}) > 0)\right\} \tag{3.4}$$

where Φ is again the standard normal CDF. As can be seen from the D/M/m case (Table 14), the LB-ratio approximation in (3.4) usually performs well, much better than the M/M/m exact values for the same cases (Table 13).

However, when $c_a^2 < 1$, the LB-ratio approximation [as well as all the others based on (2.10)] significantly underestimates the delay probability when both m is large and ρ is small. Hence, I modify (3.4) to obtain an improvement in this region by using the normal approximation in (2.11). For this purpose, we use the normal approximation in (2.11). The normal approximation for GI/G/m is $P(N \geq m) \approx 1 - \Phi(\gamma)$, where Φ is the normal CDF,

$$\gamma \equiv \gamma(m, \rho, z) = (m - m\rho - 0.5)/(m\rho z)^{1/2} \tag{3.5}$$

$$z = 1 + (c_a^2 - 1)\tau^{-1} \int_0^\infty [1 - G(x)]^2 dx, \tag{3.6}$$

TABLE 14

A Comparison of the Approximations with the Exact Values of the Probability of Delay, $P(W > 0)$, for Several GI/G/4 Models with $\rho = 0.90$

Model	Exact Value	New Approx	M/M/m Approx
M/M/4	0.79	0.79	0.79
D/M/4	0.67	0.65	0.79
H_2/M/4, $c_a^2 = 2.25$	0.85	0.85	0.79
E_4/M/4	0.71	0.70	0.79
G/M/4, $c_a^2 = 0.56$	0.75	0.75	0.79
M/D/4	0.78	0.79	0.79
M/H_2/4, $c_s^2 = 2.25$	0.79	0.79	0.79
M/G/4, $c_s^2 = 0.75$	0.79	0.79	0.79
D/H_2/4, $c_s^2 = 2.0$	0.70	0.71	0.79
E_2/H_2/4, $c_s^2 = 2.0$	0.76	0.76	0.79
H_2/H_2/4, $c_s^2 = 2.0$	0.84	0.83	0.79
H_2/H_2/4, $c_s^2 = 2.0$	0.93	0.90	0.79
D/H_2/4, $c_s^2 = 9.0$	0.74	0.75	0.79
E_2/H_2/4, $c_s^2 = 9.0$	0.77	0.77	0.79
H_2/H_2/4, $c_s^2 = 9.0$	0.84	0.81	0.79
H_2/H_2/4, $c_s^2 = 9.0$	0.92	0.86	0.79
H_2/D/4, $c_a^2 = 2.0$	0.86	0.86	0.79
H_2/E_2/4, $c_a^2 = 2.0$	0.85	0.84	0.79
H_2/D/4, $c_a^2 = 4.0$	0.91	0.89	0.79
E_2/E_2/4	0.73	0.72	0.79
G/E_2/4, $c_a^2 = 0.1$	0.64	0.60	0.79

and G is the service-time CDF. In the case of exponential service times with mean 1 that we are considering, $z = (c_a^2 + 1)/2$. For M/G/m models, (3.5) reduces to (2.11). I show this normal approximation for G/M/m models (Tables 14 and 15). Again, this is a poor approximation unless both m is large and ρ small. However, it clearly helps with a large m and a small ρ in the D/M/m model (Table 14).

The specific approximation I use for GI/M/m models depends on the number of servers m and the normal argument $\gamma = (m - m\rho - 0.5)/\sqrt{m\rho z}$ in (3.5). In particular, my approximation is

$$P(W(\mathrm{GI/M/m}) > 0)$$

$$\approx \begin{cases} (3.4), & \text{if} \quad m \leq 6 \text{ or if } \gamma \leq 0.5 \text{ or if } c_a^2 \geq 1 \\ c_a^2(3.4) + (1 - c_a^2)(3.5), & \text{if} \quad m \geq 7,\ \gamma \geq 1.0 \text{ and } c_a^2 < 1 \\ 2(1 - c_a^2)(\gamma - 0.5)(3.5) + (1 - [2(1 - c_a^2)(\gamma - 0.5)])(3.4), & \\ \quad \text{if} \quad m \geq 7,\ c_a^2 < 1 \text{ and } 0.5 < \gamma < 1, 0, & \end{cases} \tag{3.7}$$

where the equation numbers (3.4) and (3.5) are used in (3.7) instead of the actual values. Approximation (3.7) is the *New* approximation for GI/M/m models in Tables 14 to 15. Formula (3.7) usually reduces to the LB-ratio approximation in (3.4). However, in some regions I also apply the normal approximation in (3.5). The idea is to use the normal approximation when three conditions hold together: (1) the number of servers m is large, (2) the arrival variability parameter c_a^2 is small, and

TABLE 15

A Comparison of Approximations of the Probability of Delay, P (W > 0), with Exact Values from Kühn (1976) in the D/M/m Model

		Number of Servers, m					
Traffic Intensity, ρ	Method	1	2	4	8	20	100
0.50	Exact	0.20	0.103	0.033	0.0043	0.000020	0
	Half-Whitt	0.22	0.101	0.027	0.0026	0.000004	0
	LB-Ratio	0.26	0.109	0.025	0.0018	0.000001	0
	Normal	0.50	0.240	0.067	0.0067	0.000011	0
	New	0.26	0.109	0.025	0.0067	0.000011	0
0.70	Exact	0.47	0.35	0.22	0.102	0.0161	0
	Half-Whitt	0.43	0.29	0.16	0.055	0.0041	0
	LB-Ratio	0.50	0.34	0.18	0.061	0.0038	0
	Normal	0.63	0.45	0.28	0.128	0.0188	0
	New	0.50	0.34	0.18	0.128	0.0188	0
0.80	Exact	0.63	0.53	0.41	0.27	0.105	0.00108
	Half-Whitt	0.58	0.46	0.32	0.18	0.045	0.00003
	LB-Ratio	0.66	0.52	0.37	0.21	0.051	0.00003
	Normal	0.68	0.54	0.41	0.27	0.108	0.00102
	New	0.66	0.52	0.37	0.22	0.108	0.00102
0.90	Exact	0.81	0.75	0.67	0.56	0.39	0.088
	Half-Whitt	0.77	0.69	0.58	0.46	0.27	0.027
	LB-Ratio	0.82	0.75	0.65	0.52	0.31	0.031
	Normal	0.72	0.62	0.53	0.44	0.31	0.078
	New	0.82	0.75	0.65	0.52	0.31	0.078
0.95	Exact	0.90	0.87	0.82	0.76	0.65	
	Half-Whitt	0.88	0.83	0.77	0.69	0.55	0.22
	LB-Ratio	0.91	0.87	0.82	0.74	0.60	0.26
	Normal	0.74	0.66	0.59	0.52	0.44	0.26
	New	0.91	0.87	0.82	0.74	0.60	0.26
0.98	Exact	0.960	0.947	0.927	0.900	0.85	0.69
	Half-Whitt	0.951	0.931	0.903	0.865	0.79	0.58
	LB-Ratio	0.964	0.948	0.925	0.892	0.83	0.64
	Normal	0.753	0.679	0.618	0.568	0.51	0.42
	New	0.964	0.948	0.925	0.892	0.83	0.64

(3) the normal argument γ in (3.5) is relatively large. Just like formula (2.24) for EW, formula (3.7) is a continuous function of the parameters λ, c_a^2 and τ.

3.4 *The General GI/G/m Model*

My first idea for extending the approximation (3.7) for GI/M/m models to GI/G/m models with nonexponential service-time distributions was to use exactly the same formula, independent of the service-time variability parameter c_s^2. This was based on the intuition that the probability of delay should depend much more on the interarrival-time distribution than on the service-time distribution, consistent with the M/G/m experience. However, some additional refinement turned out to be beneficial.

My second idea was to draw on the GI/G/∞ case, where there are concrete results for general service times. In particular, the key parameter z in (3.6) as a function of c_a^2 and c_s^2 can be approximated well by

$$z \approx (c_a^2 + c_s^2)/(1 + c_s^2). \tag{3.8}$$

Formula (3.8) is exact when either $c_a^2 = 1$ or $c_s^2 = 1$, reducing to the previously discussed M/G/m and GI/M/m cases. Formula (3.8) is also exact when $c_s^2 = 0$.

Formula (3.8) suggests extending the Half-Whitt approximation (2.10) by replacing β with $(1 + c_s^2)\beta/(c_a^2 + c_s^2)$, which agrees with (3.2) when $c_s^2 = 1$. [However, the heavy-traffic limit theorem supporting (2.10) and (3.2) does not extend to support this heuristic approximation.] Similarly, from (3.8) I suggest using $1 - \Phi((1 + c_s^2)(1 - \rho)m^{1/2}/(c_a^2 + c_s^2))$ in the numerator of the LB-ratio approximation in (3.4). Hence I suggest using $(c_a^2 + c_s^2)/(1 + c_s^2)$ for z in (3.6) for the normal approximation (3.5). Indeed, I use this refinement to the normal distribution in (3.5) within the context of (3.7). I also use a convex combination of the old LB-ratio in (3.2) with the new LB-ratio, weighting the new by ρ^2.

The final new approximation for the general partially specified GI/G/m model is thus

$$P(W(\rho, c_a^2, c_s^2, m) > 0) \approx \min\{\pi, 1\}, \tag{3.9}$$

where

$$\pi = \begin{cases} \pi_1, & \text{if} \quad m \leq 6 \text{ or } \gamma \leq 0.5 \text{ or } c_a^2 \geq 1 \\ \pi_2, & \text{if} \quad m \geq 7 \text{ and } \gamma \geq 1.0 \text{ and } c_a^2 < 1 \\ \pi_3, & \text{if} \quad m \geq 7 \text{ and } c_a^2 < 1 \text{ and } 0.5 < \gamma < 1, \end{cases} \tag{3.10}$$

and

$$\pi_1 = \rho^2\pi_4 + (1 - \rho^2)\pi_5$$

$$\pi_2 = c_a^2\pi_1 + (1 - c_a^2)\pi_6$$

$$\pi_3 = 2(1 - c_a^2)(\gamma - 0.5)\pi_2 + (1 - [2(1 - c_a^2)(\gamma - 0.5)])\pi_1$$

$$\pi_4 = \min\left\{1, \frac{1 - \Phi((1 + c_s^2)(1 - \rho)m^{1/2}/(c_a^2 + c_s^2))}{1 - \Phi((1 - \rho)m^{1/2})} P(W(\mathrm{M/M/m}) > 0)\right\}$$

$$\pi_5 = \min\left\{1, \frac{1 - \Phi(2(1 - \rho)m^{1/2}/(1 + c_a^2))}{1 - \Phi((1 - \rho)m^{1/2})} P(W(\mathrm{M/M/m}) > 0)\right\}$$

$$\pi_6 = 1 - \Phi((m - m\rho - 0.5)/\sqrt{m\rho z}) \tag{3.11}$$

for γ in (3.5) with z in (3.8). Formula π_5 is the LB-ratio approximation for the GI/M/m model in (3.4), while π_4 is the modification of it above based on z in (3.8). Formula π_1 is the convex combination of the LB-ratio approximations; it is the final approximation in most cases. Formula π_6 is the normal approximation in (3.5) with the approximation (3.8) used for z in (3.6). Finally, (3.9) coincides with (3.7) except that π_1 is substituted for π_5 and π_6 is substituted for the previous GI/M/m normal approximation. As a consequence, (3.9) reduces to the GI/M/m approximation (3.7) when $c_s^2 = 1$. Formula (3.9) reduces to the exact M/M/m value in (2.8) when $c_a^2 = 1$.

Although (3.9) is rather cumbersome, it presents no problem for the computer. The approximation π_1 in (3.11) might be used instead for analytical manipulations. In most cases (3.9) reduces to π_1. For even greater simplicity, π_4 might be used.

Future work might aim for an approximation of the form (3.4) where the numerator of the LB ratio is $1 - \Phi((1-\rho)\sqrt{m}x)$ for $x \equiv x(\rho, c_a^2, c_s^2, m)$; in (3.4) $x = 2/(1 + c_a^2)$. The idea would be to incorporate the convex combinations in π_1, π_2 and π_3 in (3.11) in the function x inside the normal CDF. It would be nice to produce a cleaner formula.

3.5 *Numerical Comparisons*

Now I compare the approximations with the exact values of the probability of delay (Tables 14–22). These numerical comparisons show that the new approximation (3.9) performs remarkably well. For a quick overview, I compare the new approximation to the exact values for several GI/G/m models in the special case of $m = 4$ and $\rho = 0.90$ (Table 14). The new approximation is obviously much better than the exact M/M/m formula, but even the M/M/m formula is adequate in most cases.

There is significant relative error in light traffic. Because precision in light traffic is not of major practical importance for most applications, I do not even test cases with $\rho < 0.50$, and I did not try to exploit light-traffic methods. However, this remains a promising direction for future research. In light traffic the range of exact values consistent with the partial information provided by the parameters, c_a^2 and c_s^2 often is so great to rule out accurate approximation (with the criterion of relative error) for the partially specified model. According to Whitt's (1984) Theorems 1 and 2, the maximum relative error is at least $c_a^2/(1 + c_a^2)\rho$ in the GI/M/1 model. For $\rho = 0.5$ and $c_a^2 = 1$, it is 100%, and for $\rho = 0.1$ and $c_a^2 = 1$, 500%.

Typical levels of ρ in real systems sharply increase as m increases (Whitt 1992). For example, the probability of delay is negligible (0.0037) in an M/M/20 system with $\rho = 0.50$. An M/M/20 system with $\rho = 0.90$ is less congested than an M/M/2 system with $\rho = 0.70$; for the M/M/20 system with $\rho = 0.90$, $EW = 0.28$ and $P(W > 0) = 0.55$; for the M/M/2 system with $\rho = 0.70$, $EW = 0.96$ and $P(W > 0) = 0.58$.

I compare the Half-Whitt approximation in (2.10) for $P(W > 0)$ and the normal approximation in (2.11) [which coincides with π_6 in (3.11)] with the exact values for the M/M/m model (Table 13). For most cases (except large m together with small ρ), the Half-Whitt approximation (2.10) could reasonably be substituted for the exact M/M/m formula $EW(\text{M/M/m})$ in π_4 and π_5 in (3.11). It should not be difficult to further refine (2.10) for this purpose. Of course, the final approximation (3.9) produces the exact value for the M/M/m model.

I next compare the approximations with exact values from Kühn (1976) for various GI/M/m models, in particular, D/M/m and H_2/M/m with $c_a^2 = 2.25$ and balanced means (Tables 15 and 16). I display the direct Half-Whitt approximation (3.2), the LB-ratio (3.4), and the normal approximation (3.5) in addition to the new approximation (3.9), which coincides with (3.7) for the GI/M/m case. The new approximation (3.9) agrees with the LB-ratio except when both m is large and ρ is small: they differ only when both $m \geq 8$ and $\rho \leq 0.80$ and in the case $m = 100$ and $\rho = 0.90$ (Table 15). In these cases, the new approximation (3.9) obviously is much better than the LB-ratio (3.4).

I have compared the approximations with the exact values for several M/G/m values, but have omitted the tables. In these cases the approximation is the exact M/

TABLE 16

A Comparison of Approximations of the Probability of Delay, P (W > 0), with Exact Values from Kühn (1976) for the H_2/M/m Model (Hyperexponential Interarrival-Time Distribution with Balanced Means) with $c_a^2 = 2.25$

		Number of Servers, m				
Traffic Intensity, ρ	Method	2	4	8	20	100
0.50	Exact	0.45	0.28	0.123	0.016	
	Half-Whitt	0.56	0.42	0.280	0.109	
	New	0.46	0.30	0.144	0.024	
	Normal	0.35	0.20	0.085	0.009	
0.70	Exact	0.69	0.55	0.39	0.18	0.0045
	Half-Whitt	0.71	0.61	0.49	0.30	0.0390
	New	0.68	0.56	0.41	0.21	0.0110
	Normal	0.47	0.37	0.26	0.12	0.0028
0.80	Exact	0.80	0.70	0.57	0.37	0.060
	Half-Whitt	0.80	0.72	0.63	0.47	0.146
	New	0.79	0.70	0.58	0.40	0.094
	Normal	0.52	0.45	0.37	0.25	0.044
0.90	Exact	0.901	0.85	0.78	0.65	0.32
	Half-Whitt	0.895	0.85	0.80	0.70	0.42
	New	0.894	0.84	0.78	0.66	0.37
	Normal	0.570	0.52	0.47	0.39	0.22
0.95	Exact	0.951	0.925	0.887	0.82	
	Half-Whitt	0.947	0.925	0.895	0.84	0.67
	New	0.947	0.921	0.885	0.82	0.62
	Normal	0.590	0.548	0.511	0.46	0.36
0.98	Exact	0.981	0.970	0.955	0.924	0.83
	Half-Whitt	0.978	0.969	0.957	0.933	0.85
	New	0.979	0.968	0.953	0.924	0.83
	Normal	0.601	0.566	0.538	0.507	0.45

M/m value (2.3). The results confirm that the M/M/m formula is a good approximation for other M/G/m models.

Paralleling Tables 7–10, Tables 17–20 contain comparisons of approximations with exact values for G/H_2/m models using de Smit's (1983a, 1983b, and personal communication) data. Table 17 shows the LB-ratio approximation π_5 in (3.4) and (3.11) and the new ratio approximation π_4 in (3.11) as well as the new approximation (3.9), which incorporates the convex combination π_1 in (3.11). For example, the D/H_2/2 case with $\rho = 0.60$ and $c_s^2 = 9.0$ (Table 17) helps motivate the convex combination π_1 in (3.11): the exact value is 0.31, while $\pi_4 = 0.20$, $\pi_5 = 0.42$ and $\pi_1 = 0.28$ (new).

Tables 18–20 add perspective to the approximations for GI/H_2/m models partially specified by the basic parameters ρ, c_a^2, c_s^2, and m. There the range of exact values is given for H_2 distributions with $r = \frac{1}{9}$, $\frac{1}{2}$ and $\frac{8}{9}$ [for r in (2.28)] for each c^2 in the case $\rho = 0.80$. From this perspective, the new approximation for the probability of delay is quite satisfactory (Table 20). The approximation falls in the interval of exact values in every case and toward the middle when the interval is rather wide ($c_s^2 = 9.0$). Whitt's (1983) old M/M/m approximation is not good enough, and it is difficult to do much better than the new approximation given the available information (Table 20).

TABLE 17

A Comparison of Approximations of the Probability of Delay, $P\ (W > 0)$, with Exact Values from de Smit (1983a, 1983b, personal communication) for the $G/H_2/m$ Model (Hyperexponential Service Times with Balanced Means) with $c_s^2 = 9.0$

Number of Servers, m	Traffic Intensity, ρ	M/M/m Delay Probab	Method	Arrival Variability Parameter			
				$c_a^2 = 0.0$	$c_a^2 = 0.5$	$c_a^2 = 2.0$	$c_a^2 = 9.0$
2	0.3	0.138	Exact	0.059	0.090	0.19	0.30
			LB-Ratio	0.021	0.080	0.22	0.36
			New Ratio	0.117	0.128	0.16	0.25
			New	0.029	0.085	0.21	0.35
	0.6	0.45	Exact	0.31	0.39	0.53	0.69
			LB-Ratio	0.20	0.35	0.56	0.71
			New Ratio	0.42	0.43	0.48	0.59
			New	0.28	0.38	0.53	0.67
	0.8	0.71	Exact	0.62	0.68	0.77	0.88
			LB-Ratio	0.52	0.65	0.78	0.87
			New Ratio	0.69	0.70	0.73	0.80
			New	0.63	0.68	0.75	0.83
	0.9	0.85	Exact	0.80	0.84	0.89	0.94
			LB-Ratio	0.75	0.82	0.89	0.94
			New Ratio	0.84	0.85	0.86	0.90
			New	0.82	0.84	0.87	0.91
4	0.3	0.037	Exact	0.0085	0.024	0.061	0.14
			LB-Ratio	0.0012	0.014	0.080	0.18
			New Ratio	0.0275	0.032	0.047	0.10
			New	0.0035	0.016	0.077	0.17
	0.6	0.29	Exact	0.19	0.25	0.37	0.57
			LB-Ratio	0.07	0.19	0.40	0.59
			New Ratio	0.25	0.27	0.32	0.44
			New	0.14	0.22	0.37	0.54
	0.8	0.60	Exact	0.51	0.57	0.67	0.82
			LB-Ratio	0.37	0.51	0.68	0.81
			New Ratio	0.57	0.58	0.62	0.71
			New	0.50	0.56	0.64	0.75
	0.9	0.79	Exact	0.74	0.77	0.84	0.92
			LB-Ratio	0.65	0.74	0.84	0.91
			New Ratio	0.77	0.78	0.80	0.85
			New	0.75	0.77	0.81	0.86

The interarrival times are deterministic (D) when $c_a^2 = 0.0$, Erlang (E_2) when $c_a^2 = 0.5$ and hyperexponential (H_2) with balanced means when $c_a^2 > 1$. The LB ratio is π_5 in (3.4) and (3.11). The new ratio is π_4 in (3.11) and the new is (3.9).

Table 21 compares the M/M/m and new approximations with the exact values for the probability of delay in the $H_2/D/m$ model with $c_a^2 = 4.0$. As in Table 11, the exact values come from Seelen, Tijms and van Hoorn (1985). Unlike Table 11 for EW, the changes in $P(W > 0)$ from old to new in Table 21 are quite dramatic.

Table 22 compares the approximations with the exact values of the probability of delay, as well as the mean and the SCV of Q, in the $E_{10}/E_2/m$ model with $c_a^2 = 0.1$. Again the exact values come from Seelen, Tijms and van Hoorn (1985). As before, the accuracy of the approximation for $P(W > 0)$ is good.

TABLE 18

The Range of Exact Values of the Probability of No Delay, P (W = 0), for the $H_2/H_2/m$ Model with Traffic Intensity $\rho = 0.80$, from de Smit (1983a, 1983b, personal communication)

	Arrival Parameters		Service-Time Parameters								
			$c_s^2 = 2.0$			$c_s^2 = 5.0$			$c_s^2 = 9.0$		
Number of Servers, m	c_a^2	r_a	$r_s = \frac{1}{9}$	$r_s = \frac{1}{2}$	$r_s = \frac{8}{9}$	$r_s = \frac{1}{9}$	$r_s = \frac{1}{2}$	$r_s = \frac{8}{9}$	$r_s = \frac{1}{9}$	$r_s = \frac{1}{2}$	$r_s = \frac{8}{9}$
2	2.0	$\frac{1}{9}$	0.219	0.213	0.211	0.233	0.215	0.210	0.241	0.216	0.210
		$\frac{1}{2}$	0.228	0.223	0.219	0.247	0.230	0.220	0.256	0.232	0.220
		$\frac{8}{9}$	0.245	0.243	0.241	0.256	0.249	0.244	0.264	0.251	0.245
	5.0	$\frac{1}{9}$	0.126	0.120	0.117	0.148	0.123	0.117	0.161	0.125	0.117
		$\frac{1}{2}$	0.145	0.138	0.130	0.183	0.156	0.133	0.206	0.165	0.133
		$\frac{8}{9}$	0.203	0.202	0.201	0.220	0.214	0.209	0.233	0.223	0.212
	9.0	$\frac{1}{9}$	0.081	0.076	0.074	0.099	0.078	0.074	0.113	0.080	0.074
		$\frac{1}{2}$	0.099	0.092	0.085	0.138	0.112	0.088	0.167	0.124	0.088
		$\frac{8}{9}$	0.184	0.183	0.183	0.199	0.195	0.192	0.213	0.204	0.197
4	2.0	$\frac{1}{9}$	0.319	0.313	0.310	0.333	0.314	0.309	0.340	0.313	0.308
		$\frac{1}{2}$	0.329	0.322	0.318	0.347	0.327	0.317	0.356	0.328	0.317
		$\frac{8}{9}$	0.347	0.344	0.342	0.359	0.348	0.342	0.367	0.350	0.342
	5.0	$\frac{1}{9}$	0.201	0.193	0.190	0.225	0.198	0.190	0.240	0.199	
		$\frac{1}{2}$	0.222	0.210	0.201	0.268	0.230	0.203	0.295	0.239	0.204
		$\frac{8}{9}$	0.291	0.289	0.288	0.312	0.303	0.294	0.328	0.313	0.297
	9.0	$\frac{1}{9}$	0.136	0.129	0.127	0.159	0.133	0.127	0.174	0.134	0.127
		$\frac{1}{2}$	0.157	0.146	0.136	0.210	0.169	0.139	0.246	0.182	0.139
		$\frac{8}{9}$	0.264	0.263	0.263	0.283	0.277	0.271	0.301	0.289	0.276

TABLE 19

The Range of Exact Values of the Probability of No Delay, P (W = 0), for the $G/H_2/m$ Model, from de Smit (1983a, 1983b, personal communication)

			Service-Time Parameters								
	Arrival Variability Parameter,	Traffic Intensity,	$c_s^2 = 2.0$			$c_s^2 = 5.0$			$c_s^2 = 9.0$		
No. of Servers	c_a^2	ρ	$r_s = \frac{1}{9}$	$r_s = \frac{1}{2}$	$r_s = \frac{8}{9}$	$r_s = \frac{1}{9}$	$r_s = \frac{1}{2}$	$r_s = \frac{8}{9}$	$r_s = \frac{1}{9}$	$r_s = \frac{1}{2}$	$r_s = \frac{8}{9}$
2	0.0	0.6	0.72	0.75	0.77	0.65	0.70	0.76	0.62	0.69	0.75
		0.8	0.41	0.43	0.45	0.36	0.40	0.45	0.34	0.38	0.44
		0.9	0.22	0.23	0.24	0.19	0.21	0.24	0.18	0.20	0.24
	0.5	0.6	0.63	0.63	0.64	0.60	0.61	0.64	0.59	0.61	0.63
		0.8	0.34	0.34	0.35	0.32	0.33	0.35	0.31	0.32	0.35
		0.9	0.17	0.18	0.18	0.16	0.17	0.18	0.16	0.17	0.18
4	0.0	0.6	0.85	0.87	0.88	0.80	0.83	0.88	0.78	0.81	0.87
		0.8	0.53	0.55	0.57	0.48	0.51	0.56	0.46	0.49	0.55
		0.9	0.29	0.30	0.32	0.25	0.27	0.31	0.24	0.26	0.30
	0.5	0.6	0.78	0.78	0.79	0.75	0.76	0.78	0.74	0.75	0.78
		0.8	0.45	0.46	0.47	0.43	0.44	0.46	0.42	0.43	0.46
		0.9	0.24	0.24	0.25	0.23	0.23	0.25	0.22	0.23	0.24

The cases of D, E_2, and M arrival processes.

TABLE 20

A Comparison of Approximations of the Probability of Delay, $P(W > 0)$, with the Exact Values from de Smit (1983a, 1983b, personal communication) for the $G/H_2/m$ Model with Traffic Intensity $\rho = 0.80$ and $m = 2$ and 4

	Variability Parameters		Exact Values				
Number of Servers, m	c_a^2	c_s^2	Min	Median	Max	M/M/m	New
2	0.0	2.0	0.55	0.57	0.59	0.71	0.58
	0.5		0.65	0.66	0.66	0.71	0.66
	2.0		0.75	0.78	0.79	0.71	0.77
	9.0		0.82	0.91	0.93	0.71	0.86
	0.0	9.0	0.56	0.62	0.66	0.71	0.63
	0.5		0.65	0.68	0.69	0.71	0.68
	2.0		0.74	0.77	0.79	0.71	0.75
	9.0		0.79	0.88	0.93	0.71	0.83
4	0.0	2.0	0.43	0.45	0.47	0.60	0.44
	0.5		0.53	0.54	0.55	0.60	0.53
	2.0		0.65	0.68	0.69	0.60	0.67
	9.0		0.74	0.85	0.87	0.60	0.80
	0.0	9.0	0.45	0.51	0.54	0.60	0.50
	0.5		0.54	0.57	0.58	0.60	0.56
	0.2		0.63	0.67	0.69	0.60	0.64
	9.0		0.70	0.82	0.87	0.60	0.75

The interarrival-time distribution is deterministic (D) when $c_a^2 = 0.0$, Erlang (E_2) when $c_a^2 = 0.5$ and hyperexponential (H_2) when $c_a^2 > 1$.

TABLE 21

A Comparison of Approximations of the Probability of Delay, $P(W > 0)$, with Exact Values for the Model $H_2/D/m$ Having Hyperexponential Interarrival Times with Balanced Means with $c_a^2 = 4.0$, from Seelen, Tijms and van Hoorn (1985)

		Number of Servers, m					
Traffic Intensity, ρ	Method	1	2	4	8	20	100
0.50	Exact	0.72	0.58	0.42	0.25	0.074	0
	M/M/m	0.50	0.33	0.17	0.059	0.0047	0
	New	0.69	0.55	0.39	0.23	0.062	0
0.70	Exact	0.88	0.81	0.71	0.58	0.36	0.047
	M/M/m	0.70	0.58	0.43	0.27	0.094	0.0005
	New	0.84	0.76	0.66	0.53	0.35	0.058
0.80	Exact	0.93	0.89	0.82	0.72	0.55	0.18
	M/M/m	0.80	0.71	0.60	0.46	0.26	0.020
	New	0.90	0.85	0.78	0.69	0.54	0.24
0.90	Exact	0.97	0.948	0.91	0.86	0.76	0.48
	M/M/m	0.900	0.85	0.79	0.70	0.55	0.22
	New	0.956	0.93	0.89	0.85	0.76	0.53
0.95	Exact	0.985	0.975	0.958	0.932	0.88	0.71
	M/M/m	0.950	0.926	0.891	0.84	0.76	0.51
	New	0.979	0.966	0.948	0.92	0.87	0.73
0.98	Exact	0.994	0.990	0.983	0.973	0.950	0.88
	M/M/m	0.980	0.970	0.956	0.936	0.897	0.775
	New	0.992	0.987	0.980	0.969	0.949	0.884

TABLE 22

Comparison of Approximations of Several Congestion Measures in the $E_{10}/E_2/m$ Model Having $c_a^2 = 0.1$ and $c_s^2 = 0.5$ with Exact Values from Seelen, Tijms and van Hoorn (1985)

		Number of Servers, m					
		2		4		20	
Traffic Intensity, ρ	Performance Measure	Exact	Approx	Exact	Approx	Exact	Approx
0.50	$P(W > 0)$	0.086	0.117	0.025	0.030	0.00001	
	EQ	0.034	0.056	0.010	0.025	0	0.0004
	$c^2(Q)$	34.6	23.7	119.	95.0	716.0	
0.70	$P(W > 0)$	0.32	0.31	0.19	0.16	0.011	0.010
	EQ	0.26	0.31	0.16	0.21	0.010	0.027
	$c^2(Q)$	6.2	6.5	10.8	13.7	203.0	249.0
0.80	$P(W > 0)$	0.50	0.48	0.37	0.32	0.084	0.079
	EQ	0.66	0.72	0.51	0.58	0.13	0.19
	$c^2(Q)$	3.2	3.6	4.6	5.9	23.0	27.0
0.90	$P(W > 0)$	0.73	0.71	0.64	0.60	0.35	0.25
	EQ	2.06	2.14	1.84	1.93	1.07	1.20
	$c^2(Q)$	1.8	2.0	2.1	2.5	4.5	7.3
0.95	$P(W > 0)$	0.86	0.85	0.81	0.78	0.62	0.54
	EQ	5.00	5.09	4.74	4.85	3.73	3.90
	$c^2(Q)$	1.33	1.41	1.46	1.61	2.16	2.79
0.98	$P(W > 0)$	0.942	0.937	0.920	0.910	0.832	0.797
	EQ	14.0	14.1	13.7	13.8	12.5	12.7
	$c^2(Q)$	1.12	1.15	1.17	1.22	1.37	1.53

4. The Waiting-Time Distribution

In this section I obtain approximations for the variances of the steady-state waiting time W and the steady-state sojourn time T and even their full distributions. In Section 4.1 I review the approximation procedure in Whitt (1983a, Section 5), applied again here. In Section 4.2 I briefly discuss an alternative approach based on asymptotics, for which the approximations here for EW and $E(W \mid W > 0)$ can be used. In Section 4.3 I make numerical comparisons.

4.1 *The Conditional Waiting Time*

To obtain approximations for the variance of the waiting time, Var (W), and the waiting-time CDF, $P(W \leq x)$, I adopt the identical procedure used for the GI/G/1 queue in Whitt (1983a, Section 5.1). I focus on the conditional wait given that the server is busy, $D = (W \mid W > 0)$. Clearly,

$$ED = EW/P(W > 0), \tag{4.1}$$

so that I obtain an approximation for ED by combining the two previous approximations for EW in (2.24) and $P(W > 0)$ in (3.9). Following Whitt's (1983a) formula (50), I introduce the following approximation for c_D^2, the squared coefficient of variation (SCV) of D,

$$c_D^2 = 2\rho - 1 + 4(1 - \rho)d_s^3/3(c_s^2 + 1)^2, \tag{4.2}$$

where $d_s^3 = E(V^3)/(EV)^3$ with V a service time. Because the third moment $E(V^3)$

is not available in the partial model specification, I use approximations based on H_2 and E_k distributions. In particular, let

$$2 \qquad d_s^3 = \begin{cases} 3c_s^{\boxed{3}}(1 + c_s^2), & c_s^2 \geq 1 \\ (2c_s^2 + 1)(c_s^2 + 1), & c_s^2 < 1. \end{cases} \tag{4.3}$$

Formula (4.2) is the exact formula for the M/G/1 model, used as an approximation for the M/G/m model, as suggested by Hokstad (1978). I then use c_D^2 as an approximation for the GI/G/m model as well as the M/G/m model. The idea is that the conditional delay should depend much more on the service-time distribution than the interarrival-time distribution. Seelen and Tijms (1984) provided additional support for this approximation principle.

I obtain approximations for the second moments and variances in a straightforward manner from (4.1) to (4.3).

$$\begin{aligned} \mathrm{Var}\,(D) &= (ED)^2 c_D^2 = (EW)^2 c_D^2 / P(W > 0)^2 \\ E(D)^2 &= \mathrm{Var}\,(D) + (ED)^2 \\ c_W^2 &= \frac{E(W^2)}{(EW)^2} - 1 = \frac{c_D^2 + 1 - P(W > 0)}{P(W > 0)} \\ \mathrm{Var}\,(W) &= (EW)^2 c_W^2 \\ E(W^2) &= \mathrm{Var}\,(W) + (EW)^2. \end{aligned} \tag{4.4}$$

I then obtain an approximate waiting-time distribution by having an atom $P(W = 0)$ at zero, again using (3.9), and fitting a density to D given ED and c_D^2 in (4.1) and (4.2), just as described in Whitt [1983a, formulas (55) to (61)].

I easily obtain the sojourn-time congestion measures from (4.4). Because the sojourn time is the sum of independent waiting and service times,

$$ET = EW + \tau \qquad \text{and} \qquad \mathrm{Var}\,T = \mathrm{Var}\,W + \tau^2 c_s^2. \tag{4.5}$$

Then $E(T^2) = \mathrm{Var}\,T + (ET)^2$ and $c_T^2 = \mathrm{Var}\,T/(ET)^2$.

4.2 *Asymptotics*

An alternative approach, not carefully examined here, is to approximate the tail probabilities by a simple exponential distribution

$$P(W > x) \approx \alpha e^{-\eta x}, \tag{4.6}$$

where η and α are obtained from the limit

$$e^{\eta x} P(W > x) \to \alpha \qquad \text{as} \qquad x \to \infty. \tag{4.7}$$

The asymptotics in (4.7) is known to hold in considerable generality. It has been established for the PH/PH/m, GI/PH/m and PH/G/m cases (where PH means phase-type) by Takahashi (1981), Neuts and Takahashi (1981), and Abate, Choudhury and Whitt (1994c), respectively. (The PH/G/m case requires regularity conditions, the major one being that the service-time distribution should have a finite moment generating function.) Moreover, (4.7) is conjectured to hold more generally, and analogs have been established for the other steady-state random variables T, Q and N. Tijms (1986) also discussed the asymptotics, and Seelen, Tijms and van

Hoorn (1985) give the asymptotic parameters η and α (for delayed customers) in their tables.

The asymptotic decay rate η in (4.7) is obtained as the root of the transform equation

$$Ee^{\eta[(V/m)-U]} = 1, \tag{4.8}$$

where V is a service time (with mean 1) and U is an interarrival time. Moreover, approximations for η in terms of the first few moments of U and V are developed in Abate, Choudhury and Whitt (1994a), Abate and Whitt (1994) and Choudhury and Whitt (1994).

The exponential approximation (4.6) based on the small-tail asymptotics (4.7) is remarkably accurate when for x is not too small. However, the true asymptotic parameters η and α in (4.7) and (4.8) depend on more than the partial information (ρ, c_a^2, c_s^2, m). In terms of the parameter 4-tuple (ρ, c_a^2, c_s^2, m), a simple heavy-traffic approximation for η is

$$\eta \approx \frac{2m(1-\rho)}{c_a^2 + c_s^2}. \tag{4.9}$$

Approximation (4.9) can be improved significantly, though, by incorporating third moments.

Abate, Choudhury and Whitt (1992a) proposed an associated approximation for the asymptotic constant α:

$$\alpha \approx \eta EW. \tag{4.10}$$

When m is large, focusing on the conditional probability distribution $P(W > x \mid W > 0)$ may be desirable; then instead of EW in (4.10) we can use $ED = EW/P(W > 0)$.

TABLE 23

A Comparison of Approximations for ED and c_D^2 with Exact Values from Kühn (1976) for the M/D/m Model

		Number of Servers, m									
		2		4		8		20		100	
Traffic Intensity, ρ	Method	ED	c_D^2	ED	c_D^2	ED	c_D^2	ED	c_D^2	ED	c_D^2
0.50	Exact	0.55	0.62	0.30	0.62	0.17	0.67	0.081	0.78		
	Approx	0.52	0.67	0.28	0.67	0.16	0.67	0.062	0.67	0	0.67
0.70	Exact	0.87	0.75	0.46	0.71	0.25	0.70	0.113	0.72	0.029	0.86
	Approx	0.86	0.80	0.44	0.80	0.23	0.80	0.101	0.80	0.020	0.80
0.80	Exact	1.29	0.83	0.67	0.79	0.35	0.75	0.153	0.73	0.038	0.80
	Approx	1.27	0.87	0.65	0.87	0.34	0.87	0.140	0.87	0.031	0.87
0.90	Exact	2.54	0.91	1.29	0.88	0.66	0.85	0.28	0.81	0.063	0.76
	Approx	2.51	0.93	1.27	0.93	0.64	0.93	0.27	0.93	0.057	0.93
0.95	Exact	5.03		2.54		1.28		0.52	0.89	0.113	0.82
	Approx	5.01	0.97	2.52	0.97	1.26	0.97	0.52	0.97	0.106	0.97
0.98	Exact	12.53		6.29		3.16		1.27		0.26	
	Approx	12.51	0.99	6.27	0.99	3.14	0.99	1.27	0.99	0.26	0.99

The values of ED for $\rho = 0.95$ and 0.98 come from Seelen, Tijms and van Hoorn (1985).

TABLE 24

A Comparison of Approximations with the Exact Values of ED and c_D^2 in the M/H$_2$/m Model (Hyperexponential Distribution Having Balanced Means) with $c_s^2 = 2.25$, from Groenevelt, van Hoorn and Tijms (1984)

		Number of Servers, m							
		2		4		8		20	
Traffic Intensity, ρ	Method	ED	c_D^2	ED	c_D^2	ED	c_D^2	ED	c_D^2
0.50	Exact	1.50	1.43	0.68	1.42	0.31	1.33	0.112	1.18
	Approx	1.44	1.38	0.69	1.38	0.33	1.38	0.132	1.38
0.70	Exact	2.60	1.28	1.23	1.31	0.58	1.32	0.209	1.27
	Approx	2.54	1.23	1.24	1.23	0.60	1.23	0.225	1.23
0.80	Exact	3.96	1.19	1.91	1.22	0.92	1.26	0.34	1.27
	Approx	3.91	1.15	1.92	1.15	0.94	1.15	0.36	1.15
0.90	Exact	8.03	1.10	3.95	1.12	1.93	1.15	0.74	1.19
	Approx	7.98	1.08	3.96	1.08	1.96	1.08	0.77	1.08
0.95	Exact	16.2	1.05	8.02	1.06	3.97	1.08	1.55	1.11
	Approx	16.1	1.04	8.03	1.04	3.99	1.04	1.58	1.04
0.99	Exact	81.2	1.01	40.5	1.01	20.2	1.02	8.05	1.02
	Approx	81.1	1.01	40.5	1.01	20.2	1.01	8.08	1.01

4.3 *Numerical Comparisons*

I compare approximations for ED and c_D^2 in Section 4.1 with the exact values in the M/D/m and M/H$_2$/m ($c_s^2 = 2.25$ and balanced means) using data from Kühn (1976) in the first case (Table 23) and Groenevelt, van Hoorn and Tijms (1984) in the second case (Table 24). These cases show that the approximations for c_D^2 are usually adequate, but its accuracy degrades as the number m of servers increases. Because $ED = (EW)/P(W > 0)$, the performance of the approximations for ED is easy to predict given the performance of the approximations for EW and $P(W > 0)$ (Tables 1–22). Seelen, Tijms and van Hoorn (1985) provide tables of exact values for ED.

I also compare c_W^2 (as well as c_Q^2 and c_N^2) in G/H$_2$/m models with exact values from de Smit (1983a, 1983b, and personal communication) (Table 25). Finally, I compare the approximate percentiles (90, 95, 98 and 99%) of the waiting-time distribution with exact values from de Smit (Table 26). I derived the approximate values were obtained by interpolating crudely from 20 values displayed in the QNA output. When the three key variables EW, $P(W > 0)$ and c_D^2 are approximated reasonably well, the approximate waiting-time distribution seems to be remarkably accurate. On the basis of comparisons of examples from Seelen, Tijms and van Hoorn (1985) and de Smit (1983a, 1983b, and personal communication), we conclude that fitting the delay distribution with mixtures and convolutions of exponential distributions is very appropriate, especially when the interarrival-time and service-time distributions are of this form. For example, de Smit (1983a) shows that the conditional delay distribution is actually a mixture of exponentials when the service-time distribution is a mixture of exponentials.

5. The Queue Length and the Number in System

In this section I develop approximations for the queue length Q (excluding customers in service) and the number in system N, describing the system at an arbitrary

TABLE 25

A Comparison of Approximations of the Squared Coefficients of Variation of Waiting Time W, Queue Length Q and Number in System N in the GI/H_2/m Model Having c_s^2 = 2.0 and Balanced Means with Exact Values, from de Smit (1983a, 1983b, personal communication)

			Arrival Process Variability Parameter							
			$c_a^2 = 0.0$		$c_a^2 = 0.5$		$c_a^2 = 2.0$		$c_a^2 = 9.0$	
Number of Servers, m	Traffic Intensity, ρ	Congest Measure	Exact	Approx	Exact	Approx	Exact	Approx	Exact	Approx
2	0.30	$c^2(W)$	93.0	102.0	31.0	29.0	11.4	10.4	6.1	5.9
		$c^2(Q)$	109.0	199.0	47.0	57.0	22.1	21.1	14.4	12.3
		$c^2(N)$	1.3	1.4	1.6	1.6	2.38	2.52	3.3	4.5
	0.60	$c^2(W)$	7.9	8.4	5.3	5.2	3.19	3.15	1.85	2.21
		$c^2(Q)$	8.1	11.1	6.1	7.0	4.37	4.37	3.20	3.15
		$c^2(N)$	1.1	1.4	1.3	1.4	1.70	1.54	2.13	1.83
	0.80	$c^2(W)$	2.8	2.7	2.31	2.22	1.75	1.78	1.28	1.47
		$c^2(Q)$	2.8	3.1	2.47	2.60	2.07	2.11	1.72	1.76
		$c^2(N)$	1.2	1.3	1.25	1.30	1.36	1.34	1.47	1.44
	0.90	$c^2(W)$	1.7	1.6	1.54	1.49	1.32	1.35	1.12	1.22
		$c^2(Q)$	1.7	1.8	1.60	1.63	1.45	1.47	1.31	1.34
		$c^2(N)$	1.1	1.2	1.15	1.17	1.18	1.20	1.21	1.23
4	0.30	$c^2(W)$	895.0	1366.0	162.0	165.0	34.0	30.0	13.0	13.0
		$c^2(Q)$	1091.0	2659.0	253.0	322.0	69.0	60.0	31.0	26.0
		$c^2(N)$	0.6	0.6	0.71	0.70	1.08	1.26	1.7	3.1
	0.60	$c^2(W)$	16.0	21.0	9.6	10.0	5.0	4.8	2.5	2.9
		$c^2(Q)$	17.0	27.0	11.1	13.2	6.8	6.5	4.1	4.1
		$c^2(N)$	0.40	0.52	0.57	0.61	0.97	0.87	1.7	1.41
	0.80	$c^2(W)$	3.9	3.9	3.1	3.0	2.2	2.2	1.5	1.7
		$c^2(Q)$	3.9	4.5	3.3	3.5	2.6	2.6	1.9	2.0
		$c^2(N)$	0.68	0.82	0.83	0.86	1.08	1.00	1.39	1.26
	0.90	$c^2(W)$	2.05	1.94	1.80	1.74	1.49	1.50	1.19	1.30
		$c^2(Q)$	2.05	2.10	1.86	1.88	1.62	1.63	1.38	1.43
		$c^2(N)$	0.89	0.94	0.96	0.98	1.08	1.06	1.19	1.19

The interarrival-time distribution is deterministic (D) when $c_a^2 = 0$, Erlang (E_2) when $c_a^2 = 0.50$ and hyperexponential (H_2) with balanced means when $c_a^2 > 1$.

time in steady state. Approximations for the means EQ and EN in the GI/G/m model are already established by (2.1) and (2.2), using the approximation for EW in (2.24). In Section 5.1 I develop an approximation for $P(Q > 0)$. In Section 5.2 I develop approximations for the second moments, variances and squared coefficients of variation of Q and N. In Sections 5.3 and 5.4 I develop approximations for the distributions of Q and N, and in Section 5.5 I make numerical comparisons. Finally, in Section 5.6 I briefly discuss approximations for the distribution of N when there is a finite waiting room.

The approximations in this section are more exploratory than the approximations in the previous sections, so that they are likely to be less accurate. And there are fewer exact values for numerical comparisons.

5.1 *The Probability That the Queue Is Not Empty*

I now build on the previous approximation to obtain an approximation for $P(Q > 0)$, the probability that the queue is not empty at an arbitrary time in steady state. Because $P(Q > 0) = P(N \geq m + 1)$, the previous approximation for $P(N \geq m)$ in Section 3.1 is a different quantity. Of course, we should have $P(Q > 0) = P(N \geq m + 1) < P(N \geq m)$.

TABLE 26

A Comparison of the Approximate Percentiles of the Waiting-Time Distribution for the $G/H_2/m$ Model Having $c_s^2 = 2.0$ and Balanced Means with Exact Values from de Smit (1983a, 1983b, personal communication)

			Percentiles of Waiting-Time Distribution							
			90%		95%		98%		99%	
Number of Servers, m	Arrival Variability Parameter, c_a^2	Traffic Intensity, ρ	Exact	Approx	Exact	Approx	Exact	Approx	Exact	Approx
2	0.0	0.6	1.3	1.4	2.4	2.7	4.0		5.2	
	0.5		2.0	1.9	3.3	3.2	5.1	5.0	6.5	
	2.0		3.7	3.4	5.4	5.0	7.8	7.4	9.5	9.1
	0.0	0.8	4.8	4.8	6.8	6.8	9.5	9.6	11.6	11.6
	0.5		6.1	5.4	8.5	8.2	12.0	11.0	14.0	14.0
	2.0		9.8	9.0	13.0	13.0	18.0	18.0	21.0	21.0
	0.0	0.9	10.8	10.2	15.0	14.0	20.0	19.0	23.0	23.0
	0.5		13.5	13.1	18.0	18.0	24.0	24.0	29.0	29.0
	2.0		22.0	20.0	28.0	27.0	38.0	38.0	45.0	44.0
4	0.0	0.6	0.16	0.05	0.60	0.82	1.28		1.85	
	0.5		0.51	0.56	1.06	1.2	1.87		2.53	
	2.0		1.3	1.1	2.1	1.9	3.3	2.9	4.1	
	0.0	0.8	1.9	2.0	3.0	3.2	4.3	4.7	5.3	5.8
	0.5		2.6	2.6	3.8	3.8	5.4	5.5	6.6	6.6
	2.0		4.5	3.9	6.2	5.9	8.5	8.2	10.3	10.0
	0.0	0.9	5.0	5.0	6.9	6.8	9.4	9.4	11.3	11.2
	0.5		6.4	6.1	8.7	8.5	11.7	11.0	14.0	13.8
	2.0		10.0	10.0	14.0	13.0	18.0	18.0	22.0	22.0

The interarrival-time distribution is D when $c_a^2 = 0$, E_2 when $c_a^2 = 0.5$ and H_2 with balanced means when $c_a^2 > 1$.

My approximation for $P(Q > 0)$ is based on an *exact* expression for $P(Q > 0)$ given the CDF's of an interarrival time U and a waiting time W. In particular,

$$P(Q > 0) = \lambda E(\min\{U, W\}) = \lambda \int_0^\infty P(U \geq t)P(W \geq t)\mathrm{d}t$$

$$= \lambda P(W > 0) \int_0^\infty P(U \geq t)P(D \geq t)\mathrm{d}t. \quad (5.1)$$

Brumelle (1972, Theorems 2 and 3) showed that formula (5.1) can be deduced from the fundamental queueing relation $H = \lambda G$.

I apply (5.1) by approximating the two component CDF's $P(U \leq t)$ and $P(D \leq t)$ by convenient CDF's involving exponentials obtained by matching the first two moments. I follow essentially the same procedure as for $P(D \leq t)$ in Section 4.1, described by Whitt (1983a, p. 2805). For ease of calculation, in Case 4 I use a shifted-exponential distribution (Whitt 1982b, p. 138) when $0.01 \leq c^2 < 0.501$ and a deterministic distribution when $c^2 < 0.01$. I can easily carry out the integration for each of the $5 \times 5 = 25$ cases, so that I obtain closed-form expressions for $P(Q > 0)$ in terms of the parameters of the approximating distributions, (λ, c_a^2) and (ED, c_D^2).

Because $E(Q|Q \geq 1) = (EQ)/P(Q > 0)$ is necessarily greater than or equal to 1, we must have $P(Q > 0) \leq EQ$. Because $EB = m\rho = mP(Q > 0 + \sum_{k=1}^{m} kP(N$

$= k)$, $P(Q > 0) \leq \rho$. Consequently, in the final approximation formula, I replace the approximation based on (5.1) with $\min \{EQ, \rho, P(Q > 0)\}$.

A more elementary, but cruder, heuristic that could be used instead of (5.1) is

$$\begin{aligned} P(Q > 0) &\approx P(N(\mathrm{M/M/m}) \geq m + 1)P(W > 0)/P(W(\mathrm{M/M/m}) > 0) \\ &= [P(N(\mathrm{M/M/m}) \geq m + 1)/P(N(\mathrm{M/M/m}) \geq m)] \\ &\quad \times P(W(\rho, c_a^2, c_s^2, m) > 0) \\ &= \rho P(W(\rho, c_a^2, c_s^2, m) > 0), \qquad (5.2) \end{aligned}$$

which is exact for M/M/m models. Approximation (5.2) may seem quite crude because it does not take account of the fact that W is the delay at an arrival epoch while Q is the queue length at an arbitrary time. However, there is actually additional strong theoretical support for (5.2). As a consequence of $H = \lambda G$ (Franken et al. 1981, (4.3.5), Exercise 11–21 of Heyman and Sobel 1982), formula (4.2) is *exact* for all G/M/m models, even with a nonrenewal arrival process. Formula (5.1) reduces to (5.2) in the case of exponential services times, using integration by parts and the basic GI/M/m equation $\sigma = \int_0^\infty e^{-m(1-\sigma)/\tau} \mathrm{d}P(U \leq t)$.

For nonexponential services times, the exact formula (5.1) is to be preferred, but (5.2) is a useful quick approximation.

5.2 *Second-Moment Characteristics*

I rely heavily on the relatively well-justified approximate mean values EQ and EN in developing the second-moment approximations. In particular, let

$$\mathrm{Var}\,(Q) = c_Q^2(EQ)^2, \qquad \mathrm{Var}\,(N) = c_N^2(EN)^2$$
$$E(Q^2) = \mathrm{Var}\,(Q) + (EQ)^2 \quad \text{and} \quad E(N^2) = \mathrm{Var}\,(N) + (EN)^2 \qquad (5.3)$$

and focus on developing approximations for c_Q^2 and c_N^2.

5.2.1 THE QUEUE LENGTH Q. Just as I introduced the conditional delay D in Section 4.1, let C be the conditional queue length given that the queue is nonempty, i.e., $C = (Q \mid Q > 0)$. Paralleling (4.4), c_C^2 and c_Q^2 are related by

$$c_C^2 = P(Q > 0)c_Q^2 - 1 + P(Q > 0)$$

and

$$c_Q^2 = (c_C^2 + 1 - P(Q > 0))/P(Q > 0). \qquad (5.4)$$

As with c_D^2 in (4.2), I approximate c_C^2 by the M/G/m formula. I again combine an exact relation with previous approximations. In particular, we use EW, $P(W > 0)$ and c_D^2. The exact M/G/m formula is

$$E(Q^2) - EQ = \lambda^2 E(W^2) \qquad (5.5)$$

or, equivalently,

$$c_Q^2 = (1/EQ) + c_W^2; \qquad (5.6)$$

see Brumelle's (1972) Corollary to Theorem 4. From (5.5) or (5.6), I get

$$c_C^2 = \frac{1}{EC} - 1 + \frac{P(Q > 0)}{P(W > 0)}(c_D^2 + 1). \tag{5.7}$$

Naturally, I require that $c_C^2 \geq 0$. For example, for the special M/M/m case, $EC = (1 - \rho)^{-1}$, $c_D^2 = 1$, $P(Q > 0)/P(W > 0) = \rho$ and (5.7) yields $c_C^2 = \rho$. To apply (5.7) to the general GI/G/m model with parameters (ρ, c_a^2, c_s^2, m), let $c_C^2(\rho, c_a^2, c_s^2, m) = c_C^2(\rho, 1, c_s^2, m)$ where $c_c^2(\rho, 1, c_s^2, m)$ is given by (5.7). This means that c_D^2 is given by (4.2),

$$EC = P(Q > 0)EQ = \lambda P(Q > 0)EW$$

with $EW = EW(\text{M/M/m})(1 + c_s^2)/2$, $P(W > 0) = P(W(\text{M/M/m}) > 0)$ and $P(Q > 0) = P(Q(\rho, 1, c_s^2, m) > 0)$ as given by Section 5.1. In other words, c_C^2 is given the M/G/m formula for all associated GI/G/m systems. Then c_Q^2 is given by (5.4), where now $P(Q > 0) = P(Q(\rho, c_a^2, c_s^2, m) > 0)$; c_a^2 affects $P(Q(\rho, c_a^2, c_s^2, m) > 0)$ but not $c_C^2(\rho, c_a^2, c_s^2, m)$.

Finally, I obtain $E(Q^2)$ and Var (Q) by combining (5.3) and (5.4). Similarly,

$$E(C^2) = \text{Var}\,(C) + (EC)^2, \qquad \text{Var}\,(C) = (EC)^2 c_C^2$$

and

$$EC = \max\,\{1, (EQ)/P(Q > 0)\}. \tag{5.8}$$

5.2.2 The Number in System N. My approximation method for c_N^2 starts with a preliminary approximation for $E(N^2)$:

$$E(N^2) \approx S(\rho, c_a^2, c_s^2, m) \equiv P(Q(\rho, c_a^2, c_s^2, m) > 0)(m^2 + 2mEC + E(C^2)) + P(Q(\rho, c_a^2, c_s^2, m) = 0)(\min\,\{m^2, (\rho m)^2 + \rho m z\}), \tag{5.9}$$

where $P(Q(\rho, c_a^2, c_s^2, m) > 0)$ comes from Section 5.1, $EC = (EQ)/P(Q > 0)$, $E(C^2) = (c_C^2 + 1)(EC)^2$ with c_C^2 coming from Section 5.2.1, and $z = (c_a^2 + c_s^2)/(1 + c_s^2)$ as in (3.8). Formula (5.9) is a convex combination of approximations that are usually good when $P(Q > 0)$ is near one or zero. In particular, if $P(Q > 0) \approx 1$, then $N \approx m + C$, so that $EN^2 \approx E((m + C)^2) = m^2 + 2mEC + E(C^2)$. On the other hand, $P(Q > 0) \approx 0$ means that the finite-server model is well approximated by the infinite-server model, so that I use the heavy-traffic approximation in (3.5) for the GI/G/∞ model. Thus I regard N as distributed approximately as $N(\rho m, \rho m z)$ for z in (3.8), so that $E(N^2) \approx (\rho m)^2 + \rho m z$. The minimum is introduced as a correction primarily for small m. Given that $Q = 0$, $N \leq m$ so that $N^2 \leq m^2$. For large m, $(\rho m)^2 + \rho m z$ is typically less than m^2, but often not for small m, e.g., $m = 1$.

I obtain the actual approximation for c_N^2 using the M/M/m exact value and a ratio involving (5.9). In particular,

$$c_N^2(\rho, c_a^2, c_s^2, m) + 1 \approx \frac{EN^2(\text{M/M/m})S(\rho, c_a^2, c_s^2, m)}{[EN(\rho, c_a^2, c_s^2, m)]^2 S(\rho, 1, 1, m)} \tag{5.10}$$

where $EN^2(\text{M/M/m}) = (c_N^2(\text{M/M/m}) + 1)(EN(\text{M/M/m}))^2$ is the exact M/M/m value [Halfin and Whitt 1981, formula (1.8)],

$$c_N^2(\text{M/M/m}) = \frac{\rho m(1 + \delta) + (1 - \rho)^{-2}(\delta\rho + \delta(1 - \delta)\rho^2)}{[\rho m + (1 - \rho)^{-1}\delta\rho]^2}, \tag{5.11}$$

with $\delta \equiv P(N \geq m)$ the Erlang-C formula in (2.3). In (5.10), $S(\rho, c_a^2, c_s^2, m)$ is given by (5.9), and

TABLE 27

A Comparison of the Approximation of c_Q^2, the Squared Coefficient of Variation of the Queue Length, with the Exact Value for the $E_2/E_2/m$ Model from Seelen, Tijms and van Hoorn (1985)

Traffic Intensity, ρ	Method	Number of Servers, m 1	2	4	8	20
0.50	Exact	7.67	13.5	33.1	146.2	
	Approx	7.47	11.5	30.0	159.8	12,000
0.70	Exact	3.05	4.13	6.44	12.4	54.1
	Approx	2.96	4.01	6.62	13.7	69.8
0.80	Exact	2.06	2.52	3.37	5.09	11.8
	Approx	2.02	2.53	3.51	5.65	14.8
0.90	Exact	1.43	1.59	1.84	2.27	3.40
	Approx	1.41	1.60	1.90	2.41	3.85
0.95	Exact	1.20	1.26	1.36	1.52	
	Approx	1.18	1.27	1.39	1.57	2.01
0.98	Exact	1.07	1.10	1.13	1.18	1.30
	Approx	1.07	1.10	1.14	1.20	1.33

$$\begin{aligned} S(\rho, 1, 1, m) &= \rho\delta\left(m^2 + 2m\frac{EQ}{\rho\delta} + \frac{E(Q^2)}{\rho\delta}\right) \\ &\quad + (1-\rho\delta)(\min\{m^2, (\rho m)^2 + \rho m\}) \\ &= \rho\delta m^2 + 2m\delta\rho(1-\rho)^{-1} + (1+\rho)(\delta\rho)(1-\rho)^{-2} \\ &\quad + (1-\rho\delta)(\min\{m^2, (\rho m)^2 + \rho m\}) \qquad (5.12) \end{aligned}$$

with $EQ \equiv EQ(\rho, 1, 1, m) = \delta\rho/(1-\rho)$ as in (2.5) and $E(Q^2) \equiv E(Q^2)(\rho, 1, 1, m) = (EQ)^2(c_Q^2 + 1)$ with $EQ = \delta\rho/(1-\rho)$ as above and $c_Q^2 \equiv c_Q^2(\rho, 1, 1, m) = (\rho + 1 - \rho\delta)/\rho\delta$ as determined by (5.6).

Finally, I obtain $E(N^2)(\rho, c_a^2, c_s^2, m)$, not directly from (5.9), but by combining (5.3) and (5.10).

TABLE 28

A Comparison of the Approximation for c_Q^2, the Squared Coefficient of Variation of the Queue Length, with the Exact Value for the M/D/m Model from Seelen, Tijms and van Hoorn (1985)

Traffic Intensity, ρ	Method	Number of Servers, m 1	2	4	8	20	100
0.50	Exact	6.3	9.7	19.0	56.6		
	Approx	4.3	7.0	14.3	44.2	713.7	
0.70	Exact	2.80	3.54	5.1	8.6	26.9	
	Approx	2.18	2.87	4.2	7.2	22.8	4853
0.80	Exact	1.96	2.30	2.90	4.07	8.1	
	Approx	1.64	1.98	2.55	3.63	7.3	106.7
0.90	Exact	1.39	1.51	1.71	2.03	2.85	8.9
	Approx	1.27	1.40	1.60	1.91	2.71	8.4
0.95	Exact	1.18	1.23	1.31	1.43		3.03
	Approx	1.13	1.18	1.27	1.39	1.67	2.99
0.98	Exact	1.07	1.09	1.11	1.16	1.24	1.58
	Approx	1.05	1.07	1.10	1.14	1.24	1.59

5.2.3 NUMERICAL COMPARISONS. I compare approximations for c_Q^2 and c_N^2 with exact values (Tables 25, 27 and 28). The squared coefficients of variation (c_W^2, c_Q^2 and c_N^2) in moderate-to-heavy traffic ($\rho \geq 0.60$) are reasonably accurate, but the approximations for c_N^2 obviously degrade seriously in light traffic.

5.3 *The Queue-Length Distribution*

I now develop an approximation for the queue-length distribution. The approximation procedure is a discrete analog of the procedure for the delay distribution in Section 4.1. In particular, we fit a discrete distribution to $P(Q = k)$ using the approximations already developed for $P(Q > 0)$ in Section 5.1, $EC = \max\{1, EQ/P(Q > 0)\}$, and c_C^2 in Section 5.2.1. The approximate distribution has mass $P(Q = 0)$ at 0. For $k > 0$, set $P(Q = k) = P(Q > 0)P(C = k)$. I devote the rest of this section to approximating the distribution of C on the positive integers.

As do Klincewicz and Whitt (1984), I use geometric distributions as the building blocks; they are the discrete analogs of exponential distributions. I use geometric distributions on the positive integers and on the nonnegative integers. A geometric probability mass function (PMF) on the *positive integers* $\{1, 2, 3, \ldots\}$ is

$$f(k) = p(1 - p)^{k-1}, \qquad k \geq 1, \tag{5.13}$$

with tail probabilities

$$1 - F(k) = (1 - \rho)^k, \tag{5.14}$$

mean $1/p$, variance $(1 - p)/p^2$ and $c^2 = 1 - p$. A geometric PMF on the *nonnegative integers* is

$$f(k) = p(1 - p)^k, \qquad k \geq 0, \tag{5.15}$$

with tail probabilities

$$1 - F(k) = (1 - p)^{k+}{}_1, \qquad k \geq 0, \tag{5.16}$$

mean $(1 - p)/p$, variance $(1 - \rho)/p^2$ and $c^2 = 1/(1 - p)$. The single parameter p characterizes both geometric distributions.

For $P(C = k)$, there are four cases.

Case 1. $c_C^2 > 1 - (EC)^{-1} + 0.02$.

Let C be distributed as the mixture of two geometric distributions on the positive integers with balanced means, with the three parameters p_1, p_2 and γ chosen to match EC and c_C^2 and to have balanced means; i.e.,

$$P(C = k) = \gamma p_1(1 - p_1)^{k-1} + (1 - \gamma)p_2(1 - p_2)^{k-1}, \qquad k \geq 1,$$

$$P(C > k) = \gamma(1 - p_1)^k + (1 - \gamma)(1 - p_2)^k, \tag{5.17}$$

where $p_1 = m_1^{-1} > p_2 = m_2^{-1}$, $(EC)/2 = \gamma m_1 = (1 - \gamma)m_2$ and

$$\gamma = [1 + (1 - 2[c^2 + 1 + (EC)^{-1}]^{-1})^{1/2}]/2. \tag{5.18}$$

Notice that γ in (5.18) has a solution in the interval $(\frac{1}{2}, 1)$ provided that $c^2 > (EC - 1)/EC$. However, it is possible to have $p_1 = 2\gamma/EC > 1$ when EC is relatively small, or when $EC < 2$. If $p_1 > 1$ in this calculation, skip to Case 2 and use a simple geometric distribution.

Case 2. $|c_C^2 - 1 + (EC)^{-1}| \leq 0.02$.

Let $P(C = k)$ have a geometric distribution as in (5.13) with $p = 1/EC$, which has $c^2 = (EC - 1)/EC$, therefore agreeing closely with c_C^2.

Case 3. $[(EC)^2 - 1]/2(EC)^2 < c_C^2 \leq 1 - (EC)^{-1} - 0.02$.

Let C be distributed as the *convolution* of two geometric distributions, the first on the nonnegative integers with mean $m_1 = (1 - p_1)/p_1 \geq 0$ and the second on the positive integers with mean $m_2 = p_2^{-1} \geq 1$. Do not use two geometric distributions on the positive integers as in (5.13) because then the support of the convolution would be $\{k : k \geq 2\}$. Similarly, do not use two geometric distributions on the nonnegative integers, because then the support of the convolution would be $\{k : k \geq 0\}$. Hence, use one of each.

Let the PMF and tail probabilities of the new approximate distribution be

$$P(C = k) = \sum_{j=0}^{k-1} p_1(1 - p_1)^j p_2(1 - p_2)^{k-j-1}, \qquad k \geq 1,$$

$$P(C > k) = 1 - \sum_{j=0}^{k} P(C = j) \tag{5.19}$$

where, in terms of $x = EC$ and $c^2 = c_C^2$, the means of the component distributions are

$$m_1 = ((x - 1) - [(x - 1)^2 - 2x^2(1 - c^2 - x^{-1})]^{1/2})/2$$

$$m_2 = ((x + 1) + [(x - 1)^2 - 2x^2(1 - c^2 - x^{-1})]^{1/2})/2. \tag{5.20}$$

Then $p_1 = (m_1 + 1)^{-1}$ and $p_2 = m_2^{-1}$. In order to have $(x - 1)^2 - 2x^2(1 - c^2 - x^{-1})$ nonnegative, we must have $(x^2 - 1)/2x^2 \leq c^2 \leq 1 - x^{-1}$, which determines the region specified.

Example 3. We now illustrate the procedure in Case 3. If $EC = 4$ and $c_C^2 = 0.50$, then (5.20) yields $m_1 = 1$ and $m_2 = 3$, so that $p_1 = \frac{1}{2}$ and $p_2 = \frac{1}{3}$ and the associated variance is $(1 - p_1)/p_1^2 + (1 - p_2)/p_2^2 = 2 + 6 = 8$, as desired.

Case 4. $c_C^2 \leq [(EC)^2 - 1]/2(EC)^2$.

In this case, act as if $c_C^2 = [(EC)^2 - 1]/2(EC)^2$ and apply Case 3. In particular, use (5.19) with $p_1 = (m_1 + 1)^{-1} = p_2 = m_2^{-1}$ where $m_1 = (EC - 1)/2$ and $m_2 = (EC + 1)/2$.

5.4 *The Distribution for the Number in System*

A simple approximation for the PMF $P(N = k)$ is

$$P(N = k) = \begin{cases} P(Q = k - m), & k \geq m + 1 \\ p(k) & 0 \leq k \leq m, \end{cases} \tag{5.21}$$

where $p(k) = q(k)/\sum_{j=0}^{m} q(j)$ with $q(j) = \alpha^j e^{-\alpha}/j!$, i.e., $p(k)$ is a truncated Poisson distribution with intensity α. I want to consider the truncated Poisson approximation due to the exact behavior of infinite-server models: N has a Poisson distribution for M/G/∞ systems. N is asymptotically normally distributed in heavy traffic for GI/G/∞ systems, so that a Poisson distribution should be a reasonable discrete analog. Hence, I expect this approximation to perform relatively well when the number of servers m is large. A natural two-parameter alternative would be the binomial distribution.

Complete approximation of $P(N = k)$ by specifying the Poisson intensity α, which we can do by matching the exact value of the expected number of busy servers:

$$EB = m\rho = \sum_{k=0}^{m} kP(N = k) + mP(Q > 0), \tag{5.22}$$

leading to the formula

$$\sum_{k=0}^{m} kp(k) = m[\rho - P(Q > 0)]. \tag{5.23}$$

I choose the Poisson intensity α to satisfy (5.23), using the approximation for $P(Q > 0)$ in Section 5.1. [Section 5.1 shows that both the exact formula and the approximation satisfy $P(Q > 0) \leq \rho$.]

The left sides of (5.22) and (5.23) in turn are $\alpha[1 - B(m, \alpha)]$ where $B(m, \alpha)$ is the Erlang-B formula associated with the M/M/m loss system, i.e., the left side is the carried load (Cooper 1982, p. 89). Thus calculation (5.23) is equivalent to finding the offered lead α in an M/M/m loss system given the carried lead $m[\rho - P(Q > 0)]$. Jagerman (1984, pp. 1289 and 1303) describes an appropriate computational procedure.

Of course, this approximation for the PMF $P(N = k)$ yields approximations for associated characteristics of N. By (5.21)–(5.23), this procedure is consistent with my previous approximation for the mean EN, but it yields new candidate approximations for c_N^2 in (5.10) and $P(N \geq m)$ in Section 3.1. However, I regard the approach here as less accurate.

5.5 *Numerical Comparisons*

I compare the approximations for the queue-length distribution with exact values from Hillier and Yu (1981) (Table 29). The cases considered are D/M/8 and M/E$_2$/8 with $\rho = 0.7$ and $\rho = 0.9$. For these cases, in which both $c_a^2 \leq 1$ and $c_s^2 \leq 1$, the approximations appear to be remarkably accurate. (The calculation is exact for M/M/m models.) In general, the accuracy improves as the traffic intensity increases. The weakest part of the approximation scheme seems to be the initial values, e.g., the probabilities $P(Q = 0)$ and $P(Q = 1)$. [However, recall that (5.1) is exact for G/M/m models, so that for the D/M/m cases in Table 29 all error in $P(Q > 0)$ is attributable to error in $P(W > 0)$.] Both the probability mass function $P(Q = k)$ and the cumulative distribution function $P(Q \leq k)$ are quite accurate for larger values of k. Overall, the approximations seem to be sufficiently accurate for practical engineering purposes.

5.6 *Finite Buffers*

I obtain a simple approximation for the distribution of the number in system with a finite buffer (and stipulate that arriving customers who find a full system are lost without affecting future arrivals) from the approximate distribution of N in Section 5.4 by conditioning or, equivalently, by truncating and renormalizing (Whitt 1984e). Let $N(\rho, c_a^2, c_s^2, m, K)$ denote the number in system in the GI/G/m/K model with K extra waiting spaces at an arbitrary time in steady state. The proposed approximation is then

$$P(N(\rho, c_a^2, c_s^2, m, K) = k) = \frac{P(N(\rho, c_a^2, c_s^2, m, \infty) = k)}{P(N(\rho, c_a^2, c_s^2, m, \infty) \leq K)}. \tag{5.24}$$

TABLE 29

A Comparison of Approximate Queue-Length Distributions in GI/G/8 Models with Exact Values from Hillier and Yu (1981)

The D/M/8 Model

	$\rho = 0.7$				$\rho = 0.9$			
	$P(Q = k)$		$P(Q \leq k)$		$P(Q = k)$		$P(Q \leq k)$	
k	Exact	Approx	Exact	Approx	Exact	Approx	Exact	Approx
0	0.928	0.910	0.928	0.910	0.495	0.535	0.495	0.535
1	0.038	0.046	0.966	0.956	0.097	0.081	0.593	0.616
2	0.018	0.021	0.984	0.977	0.079	0.066	0.672	0.682
3	0.008	0.010	0.992	0.987	0.063	0.054	0.735	0.738
4	0.004	0.005	0.996	0.992	0.051	0.045	0.786	0.782
5	0.002	0.003	0.998	0.995	0.041	0.037	0.827	0.820
10					0.014	0.014	0.941	0.927
15					0.005	0.006	0.980	0.970
20					0.002	0.002	0.993	0.991

The M/E$_2$/8 Model

	$\rho = 0.7$				$\rho = 0.9$			
	$P(Q = k)$		$P(Q \leq k)$		$P(Q = k)$		$P(Q \leq k)$	
k	Exact	Approx	Exact	Approx	Exact	Approx	Exact	Approx
0	0.823	0.811	0.823	0.811	0.385	0.369	0.385	0.369
1	0.061	0.072	0.885	0.882	0.075	0.083	0.461	0.452
2	0.041	0.044	0.926	0.927	0.068	0.072	0.528	0.524
3	0.027	0.027	0.953	0.955	0.060	0.063	0.589	0.587
4	0.017	0.017	0.970	0.972	0.053	0.054	0.642	0.641
5	0.011	0.011	0.981	0.982	0.046	0.047	0.688	0.688
10	0.0011	0.0011	0.998	0.998	0.023	0.023	0.845	0.845
15	0.00011	0.00012	0.9998	0.9998	0.012	0.012	0.923	0.923
20		0.00001			0.0057	0.0057	0.962	0.962

Formula (5.24) is exact for M/M/m/K models and M/G/m/0 models, but not elsewhere. See Yao and Buzacott (1985a, 1985b) and Berger and Whitt (1992) for related work.

6. Conclusions

I have developed approximations for most of the standard steady-state congestion measures describing the GI/G/m model. As indicated in Section 2.2, the quality of the approximations is not the same for all these congestion measures. Some congestion measures, such as the expected number of busy servers *EB* in (2.2), are exact. Other approximate congestion measures, such as the expected number of customers in the system *EN* in (2.2), tend to be extremely accurate, primarily because a large component is exact. Still other approximate congestion measures, such as the expected waiting time *EW* in Section 2, are fairly accurate because they are relatively robust and extensively studied. Finally, some approximate congestion measures, such as the probability that the queue is not empty at an arbitrary time, are less reliable, because they have not yet been studied sufficiently and because such descriptions evidently depend more critically on the missing information (the distributions beyond the first two moments). Nevertheless, very detailed approximations with seemingly little theoretical basis, such as the queue-length probability mass function, are often remarkably accurate, as was indicated in Section 5.5.

Because exact numerical procedures have been developed for a large class of GI/G/m queues, the approximations developed here are not necessarily needed. Nevertheless, relatively concise formulas are helpful for understanding and for using GI/G/m models as submodels in larger models. The methods for developing the approximations may also be useful for related problems for which exact analysis is not yet possible.[1]

[1] I am especially grateful to my colleague A. T. Seery for running the QNA programs on many cases in 1982–1985 to help develop and evaluate these approximations, to J. H. A. de Smit of Twente University, Enschede, the Netherlands, for providing extensive tables of exact values based on his algorithm for the $GI/H_2/m$ model in 1982, and to H. Tijms and M. H. van Hoorn of the Free University, Amsterdam, for providing additional exact values from the Q-LIB program developed by van Hoorn and Seelen (1984). Finally, I am grateful to Mark Spearman for suggesting that I prepare a revision of Whitt (1985) for this journal.

References

ABATE, J., G. L. CHOUDHURY, AND W. WHITT (1994a), "Exponential Approximations for Tail Probabilities in Queues, I: Waiting Times," *Operations Research*, to appear.

——— (1994b), "Exponential Approximations for Tail Probabilities in Queues, II: Sojourn Time and Workload," *Operations Research*, to appear.

——— (1994c), "Asymptotics for Steady-State Tail Probabilities in Structured Markov Queueing Models," *Stochastic Models*, 10, 1, 99–143.

——— (1993), "Calculation of the GI/G/1 Waiting-Time Distribution and its Cumulants from Pollaczek's Formulas," *Archiv für Elektronik und Übertragungstechnik*, 47, 311–321.

ABATE, J. AND W. WHITT (1994), "A Heavy-Traffic Expansion for Asymptotic Decay Rates of Tail Probabilities in Multi-Channel Queues," *Operations Research Letters*, to appear.

ABRAMOWITZ, M. AND I. A. STEGUN (1972), *Handbook of Mathematical Functions*, 10th Printing, National Bureau of Standards, Washington, D.C.

ALLEN, A. O. (1990), *Probability, Statistics and Queueing Theory, with Computer Science Applications*, 2nd ed., Academic Press, Boston.

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

BERGER, A. W. AND W. WHITT (1992), "The Brownian Approximation for Rate-Control Throttles and the G/G/1/C Queue," *Journal of Discrete Event Dynamic Systems*, 2, 7–60.

BERTSIMAS, D. (1988), "An Exact FCFS Waiting Time Analysis for a General Class of G/G/s Queueing Systems," *Queueing Systems: Theory and Applications*, 3, 305–320.

——— (1990), "An Analytic Approach to a General Class of G/G/s Queueing Systems," *Operations Research*, 38, 139–155.

BITRAN, G. R. AND D. TIRUPATI (1988), "Multiproduct Queueing Networks with Deterministic Routing: Decomposition Approach and the Notion of Interference," *Management Science*, 34, 75–100.

BOROVKOV, A. A. (1965), "Some Limit Theorems in the Theory of Mass Service, I," *Theory of Probability and its Applications*, 10, 375–400.

——— (1967), "On the Limit Laws for Service Processes in Multi-Channel Systems," *Siberian Mathematics Journal*, 8, 746–763.

BOXMA, O. J., J. W. COHEN, AND N. HUFFELS (1979), "Approximations of the Mean Waiting Time in an M/G/s Queueing System," *Operations Research*, 27, 1115–1127.

BRUMELLE, S. L. (1972), "A Generalization of $L = \lambda W$ to Moments of Queue Length and Waiting Times," *Operations Research*, 20, 1127–1136.

BURMAN, D. Y. AND D. R. SMITH (1983), "A Light-Traffic Theorem for Multi-Server Queues," *Mathematics of Operations Research*, 8, 15–25.

BUZACOTT, J. A. AND J. G. SHANTHIKUMAR (1992), *Stochastic Models of Manufacturing Systems*, Prentice-Hall, Englewood Cliffs, NJ.

CHOUDHURY, G. L. AND W. WHITT (1994), "Heavy-Traffic Asymptotic Expansions for the Asymptotic Decay Rates in the BMAP/G/1 Queue," *Stochastic Models*, 10, 2, to appear.

COOPER, R. B. (1981), *Introduction to Queueing Theory*, 2nd ed., North-Holland Publishers, New York.

COSMETATOS, G. P. (1975), "Approximate Explicit Formulae for the Average Queueing Time in the Process (M/D/r) and (D/M/r)," *INFOR*, 13, 328–331.

COSMETATOS, G. P. (1982), "On the Implementation of Page's Approximation for Waiting Times in General Multi-Server Queues," *Journal of Operational Research Society*, 33, 1158–1159.

DAI, J. G., V. NGUYEN AND M. I. REIMAN (1993), "Sequential Bottleneck Decomposition: An Approximation Method for Open Queueing Networks," *Operations Research*, to appear.

DE SMIT, J. H. A. (1983a), "The Queue GI/M/s with Customers of Different Types or the Queue GI/$H/_m/s$," *Advances in Applied Probability*, 15, 392–419.

——— (1983b), "A Numerical Solution for the Multi-Server Queue with Hyperexpotential Service Times," *Operations Research Letters*, 2, 217–224.

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," *Proceedings of the IEEE*, 77, 171–194.

FENDICK, K. W., V. R. SAKSENA AND W. WHITT (1991), "Investigating Dependence in Packet Queues with the Index of Dispersion for Work," *IEEE Transactions of Communications*, 39, 1231–1244.

FRANKEN, P., D. KÖNIG, U. ARNDT, AND V. SCHMIDT (1981), *Queues and Point Processes*, Akademie-Verlag, Berlin.

FREDERICKS, A. A. (1983), "Approximating Parcel Blocking via State Dependent Birth Rates," *Proceedings Tenth Int. Teletraffic Congress*, Montreal, Paper 5.3.2.

GLYNN, P. W. AND W. WHITT (1991), "A New View of the Heavy-Traffic Limit for Infinite-Server Queues," *Advances in Applied Probability*, 23, 188–209.

GROENEVELT, H., M. H. VAN HOORN, AND H. C. TIJMS (1984), "Tables for M/G/c Queueing Systems with Phase-Type Service," *European Journal of Operations Research*, 16, 257–269.

HALFIN, S. AND W. WHITT (1981), "Heavy-Traffic Limits for Queues with Many Exponential Servers," *Operations Research*, 29, 567–588.

HARRISON, J. M. AND V. NGUYEN (1990), "The QNET Method for Two-Moment Analysis of Open Queueing Networks," *Queueing Systems: Theory and Applications*, 6, 1–32.

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

HILLIER, F. S. AND O. S. YU (1981), *Queueing Tables and Graphs*, North-Holland Publishers, New York.

HOKSTAD, P. (1978), "Approximations for the M/G/n Queue," *Operations Research*, 26, 510–523.

IGLEHART, D. L. (1965), "Limit Diffusion Approximations for the Many-Server Queue and the Repairman Problem," *Journal of Applied Probability*, 2, 429–441.

——— AND W. WHITT (1970), "Multiple Channel Queues in Heavy Traffic, II: Sequences, Networks and Batches," *Advances in Applied Probability*, 2, 355–369.

JAGERMAN, D. L. (1984), "Methods in Traffic Calculations," *AT&T Bell Laboratories Technical Journal*, 63, 1283–1310.

JOHNSON, M. A. AND M. R. TAAFFE (1991), "An Investigation of Phase-Distribution Moment-Matching Algorithms for Use in Queueing Models," *Queueing Systems: Theory and Applications*, 8, 129–148.

KIMURA, T. (1986), "A Two-Moment Approximation for the Mean Waiting Time in the GI/G/s Queue," *Management Science*, 32, 751–763.

KINGMAN, J. F. C. (1965), "The Heavy Traffic Approximation in the Theory of Queues," in *Proceedings of Symposium on Congestion Theory*, W. Smith and W. Wilkinson (eds.), University of North Carolina Press, Chapel Hill, 137–159.

KLINCEWICZ, J. G. AND W. WHITT (1984), "On Approximations for Queues, II: Shape Constraints," *AT&T Bell Laboratories Technical Journal*, 63, 139–161.

KÖLLERSTRÖM, J. (1974), "Heavy Traffic Theory for Queues with Several Servers: I," *Journal of Applied Probability*, 11, 544–552.

KRAEMER, W. AND M. LANGENBACH-BELZ (1976), "Approximate Formulae for the Delay in the Queueing System GI/G1," *Proceedings of Eighth International Teletraffic Congress*, Melbourne, 235, 1–8.

KÜHN, P. (1976), *Tables on Delay Systems*, Institute of Switching and Data Technics, University of Stuttgart, Germany.

Lee, A. M. and P. A. Longton (1959), "Queueing Processes Associated with Airline Passengers Check-In," *Operations Research Quarterly*, 10, 56–71.

Lucantoni, D. M. and V. Ramaswami (1985), "Efficient Algorithms for Solving The Non-Linear Matrix Equations Arising in Phase Type Queues," *Stochastic Models*, 1, 29–52.

Neuts, M. F. (1986), "The Caudal Characteristic Curve of Queues," *Advances in Applied Probability*, 18, 221–254.

——— and Y. Takahashi (1981), "Asymptotic Behavior of the Stationary Distributions in the GI/PH/c Queue with Heterogeneous Servers," *Zeitschrift für Wahrscheinlichkeitsthorie und Verwandte Gebiete*, 57, 441–452.

Nozaki, S. A. and S. M. Ross (1978), "Approximations in Finite Capacity Multi-Server Queues with Poisson Arrivals," *Journal of Applied Probability*, 15, 826–834.

Page, E. (1982), "Tables of Waiting Times for M/M/n, M/D/n and D/M/n and Their Use to Give Approximate Waiting Times in More General Queues," *Journal of the Operational Research Society*, 33, 453–473.

Ramaswami, V. and D. M. Lucantoni (1985a), "Stationary Waiting Time Distribution in Queues with Phase Type Service and In Quasi-Birth-And-Death Processes," *Stochastic Models*, 1, 125–136.

——— (1985b), "Algorithms for the Multi-Server Queue with Phase Type Service," *Stochastic Models*, 1, 393–417.

Sakasegawa, H. (1977), "An Approximation Formula $L_q = \alpha\beta^\rho/(1-\rho)$," *Annals of the Institute for Statistical Mathematics*, 29, 67–75.

Sanders, B. and E. A. Van Doorn (1985), "Estimating Time Congestion from Traffic Parameters," Memorandum 509, Department of Applied Mathematics Twente University of Technology, Enschede, The Netherlands.

Seelen, L. P. (1986), "An Algorithm for Ph/Ph/c Queues," *European Journal of the Operations Research Society*, 23, 118–127.

Seelen, L. P. and H. C. Tijms (1984), "Approximations for the Conditional Waiting Times in the GI/G/c Queue," *Operations Research Letters*, 3, 183–190.

——— (1985), "Approximations to the Waiting Time Percentiles in the M/G/c Queue," in *Teletraffic Issues in an Advanced Information Society*, *ITC*-11, M. Akiyama (ed.), Elsevier Science, Amsterdam, 53–57.

——— and M. H. Van Hoorn (1985), *Tables for Multi-Server Queues*, North-Holland, Amsterdam.

Segal, M. and W. Whitt (1989), "A Queueing Network Analyzer for Manufacturing," in *Teletraffic Science for New Cost-Effective Systems, Networks and Services*, *ITC*-12, M. Bonatti (ed.), Elsevier-Science, Amsterdam, 1146–1152.

Suri, R. and S. De Treville (1991), "Full Speed Ahead: "A Look at Rapid Modeling Technology in Operations Management," *OR/MS Today*, June, 34–42.

——— (1992), "Rapid Modeling: The Use of Queueing Models to Support Time-Based Competitive Manufacturing," in *Proceedings German/US Conference on Recent Development in Operations Research*, G. Fandel (ed.), Springer, Berlin.

Suri, R., J. L. Sanders, and M. Kamath (1993), "Performance Evaluation of Production Networks," Chapter 5 in *Handbooks in OR & MS*, *Vol.* 4, S. C. Graves et al. (eds.), Elsevier Science Publishers, Amsterdam, 199–286.

Takahashi, Y. (1981), "Asymptotic Exponentiality of the Tail of the Waiting-Time Distribution in a PH/PH/c Queue," *Advances in Applied Probability*, 13, 619–630.

——— and Y. Takami (1976), "A Numerical Method for the Steady-State Probabilities of a GI/G/s Queueing system in a General Class," *Journal of the Operations Research Society of Japan*, 19, 147–157.

Tijms, H. (1986), *Stochastic Modelling and Analysis: A Computational Approach*, Wiley, New York.

Van Hoorn, M. H. and L. P. Seelen (1984), "Q-Lib, A Program Library for Multi-Server Queues," unpublished paper and program, Dept. of Actuarial Sciences and Econometrics, Free University, Amsterdam.

Whitt, W. (1982a), "On the Heavy-Traffic Limit Theorem for GI/G/∞ Queues," *Advances in Applied Probability*, 14, 171–190.

——— (1982b), "Approximating a Point Process by a Renewal Process, I: Two Basic Methods," *Operations Research*, 30, 125–147.

——— (1983a), "The Queueing Network Analyzer," *Bell System Technical Journal*, 62, 2779–2815.

——— (1983b), "Comparison Conjectures for the M/G/s Queue," *Operations Research Letters*, 2, 203–209.

——— (1984a), "On Approximations for Queues, I: Extremal Distributions," *AT&T Bell Laboratories Technical Journal*, 63, 115–138.

——— (1984b), "On Approximations for Queues, III: Mixtures of Exponential Distributions," *AT&T Bell Laboratories Technical Journal*, 63, 163–175.

——— (1984c), "Open and Closed Models for Networks of Queues," *AT&T Bell Laboratories Technical Journal*, 63, 1911–1979.

——— (1984d), "Minimizing Delays in the GI/G/1 Queue," *Operations Research* 32, 41–51.

——— (1984e), "Heavy-Traffic Approximations for Service Systems with Blocking," *AT&T Bell Laboratories Technical Journal*, 63, 689–708.

——— (1985), *Approximations for the GI/G/m Queue*, AT&T Bell Laboratories, Holmdel, NJ.

——— (1989), "An Interpolation Approximation for the Mean Workload in a GI/G/1 Queue," *Operations Research*, 37, 936–952.

——— (1991), "A Review of $L = \lambda W$ and Extensions," *Queueing Systems: Theory and Applications*, 9, 235–268.

——— (1992), "Understanding the Efficiency of Multi-Server Service Systems," *Management Science*, 38, 708–723.

——— (1994), "Towards Better Multi-Class Parametric-Decomposition Approximations for Open Queueing Networks," *Annals of Operations Research*, 48, to appear.

WOLFF, R. W. (1982), "Poisson Arrivals See Time Averages," *Operations Research*, 30, 223–231.

YAO, D. D. W. AND J. A. BUZACOTT (1985a), "Queueing Models for a Flexible Machining Station, Part I: The Diffusion Approximation," *European Journal Operations Research*, 19, 233–240.

——— (1985b), "Queueing Models for a Flexible Machining Station, Part II: The Method of Coxian Phases," *European Journal Operations Research*, 19, 241–252.