MANAGEMENT SCIENCE
Vol. 51, No. 2, February 2005, pp. 221–235
ISSN 0025-1909 | EISSN 1526-5501 | 05 | 5102 | 0221

informs

DOI 10.1287/mnsc.1040.0302

Engineering Solution of a Basic Call-Center Model

Ward Whitt
Department of Industrial Engineering and Operations Research, Columbia University,
New York, New York 10027, ww2040@columbia.edu

An algorithm is developed to rapidly compute approximations for all the standard steady-state performance measures in the basic call-center queueing model $M/GI/s/r + GI$, which has a Poisson arrival process, independent and identically distributed (IID) service times with a general distribution, s servers, r extra waiting spaces and IID customer abandonment times with a general distribution. Empirical studies of call centers indicate that the service-time and abandon-time distributions often are not nearly exponential, so that it is important to go beyond the Markovian $M/M/s/r + M$ special case, but the general service-time and abandon-time distributions make the realistic model very difficult to analyze directly. The proposed algorithm is based on an approximation by an appropriate Markovian $M/M/s/r + M(n)$ queueing model, where $M(n)$ denotes state-dependent abandonment rates. After making an additional approximation, steady-state waiting-time distributions are characterized via their Laplace transforms. Then the approximate distributions are computed by numerically inverting the transforms. Simulation experiments show that the approximation is quite accurate. The overall algorithm can be applied to determine desired staffing levels, e.g., the minimum number of servers needed to guarantee that, first, the abandonment rate is below any specified target value and, second, that the conditional probability that an arriving customer will be served within a specified deadline, given that the customer eventually will be served, is at least a specified target value.

Key words: call centers; contact centers; queues; multiserver queues; queues with customer abandonment; multiserver queues with customer abandonment; staffing; staffing call centers; birth-and-death processes; numerical transform inversion

History: Accepted by Wallace J. Hopp, stochastic models and simulation; received December 9, 2003. This paper was with the author 1 month for 2 revisions.

1. Introduction

In this paper, we aim to contribute to the better design and management of telephone call centers and their generalizations to include new media such as e-mail and chat. The research effort is important because call centers are a growing part of the economy and because call centers are quite complicated (see Gans et al. 2003 for background). One reason that call centers are complicated is that they often involve multiple sites with multiple groups of agents having different skills, serving multiple classes of customers with different needs. Another reason call centers are complicated is that waiting customers may abandon. Moreover, the probability distributions of both the service times and abandonment times often are not nearly exponential, making it inappropriate to directly apply a simple Markovian model (see Bolotin 1994, Brown et al. 2005).

We focus on the problem of nonexponential service-time and abandonment-time distributions. In this paper, we only consider a single call center with a single group of agents, serving a single group of callers, but we hope to show in future work that our approach to the single-site, single-class problem will help analyze the more general multisite, multiclass problem. Assuming that waiting customers cannot see the queue, it is natural to assume that the customer abandonment times are IID (independent and identically distributed) with a general distribution. In this single-site, single-class setting with invisible queues, it is commonly agreed that a good model is the $M/GI/s/r + GI$ queue, which has a Poisson arrival process (the M), IID service times with a general distribution (the first GI), s servers, r extra waiting spaces, IID customer abandonment times with a general distribution (the final GI) and the first-come–first-served service discipline. This model ignores the time dependence almost always found in call arrival processes, but the time dependence often tends to be not too important over short time intervals, such as 15–60 minutes.

A serious problem is that the $M/GI/s/r + GI$ queue is extremely difficult to analyze. In the special case of the $M/M/s/r + M$ queue, where the service-time and abandon-time distributions are exponential, the number of customers in the system over time is a birth-and-death process, so the model is relatively tractable (see Palm 1937, Ancker and Gafarian 1963, Whitt 1999, Garnett et al. 2002). However, even in the $M/M/s/r + M$ model, computing waiting-time distributions is somewhat complicated. Since the Laplace transforms of waiting times are not difficult

to construct in the $M/M/s/r+M$ model, numerical transform inversion is an effective approach there, as pointed out in Whitt (1999). We will use numerical transform inversion again here to calculate our approximate waiting-time distributions for the $M/GI/s/r+GI$ model.

Important work on non-Markovian generalizations of the $M/M/s/r+M$ queue have been done previously (see Baccelli and Hebuterne 1981; Brandt and Brandt 1999, 2002; Mandelbaum and Zeltyn 2004; and references therein). In particular, much is now known about the $M/M/s/r+GI$ model. However, there still seems to be a need for an effective algorithm for the $M/GI/s/r+GI$ queue. For other studies of customer abandonment behavior, see Mandelbaum and Shimkin (2000) and Zohar et al. (2002).

Our goal in this paper is to develop an efficient algorithm for calculating effective approximations for all standard steady-state performance measures in the $M/GI/s/r+GI$ queue for distributions and parameters commonly occurring in call centers. In particular, we are particularly interested in the case in which there is ample waiting room (r might be taken to be ∞), the number of servers is relatively large (e.g., $s=100$ or even $s=1{,}000$) and there is nonnegligible, but not excessively large, customer abandonment (e.g., 1–10%). We want to allow realistic nonexponential service-time and abandon-time distributions. For example, as observed in Brown et al. (2005), the service-time distribution might be lognormal with a squared coefficient of variation (SCV, variance divided by the square of the mean) between 1 and 2.

Our approach involves two approximations: First, we approximate the given $M/GI/s/r+GI$ model by a Markovian $M/M/s/r+M(n)$ model, which has IID exponential service times with the given service-time mean and state-dependent abandonment rates. Most of the novelty lies in the state-dependent abandonment rates. Second, we develop an approximate solution for all the performance measures in the approximating $M/M/s/r+M(n)$ model. Just like for the $M/M/s/r+M$ model, the steady-state distribution of the number of customers in the $M/M/s/r+M(n)$ system at an arbitrary time is easy to compute exactly, because the process is a birth-and-death process. The second approximation appears when we describe the experience of individual customers, e.g., when we compute the probability that an entering customer eventually is served or the conditional waiting-time distribution given that a customer eventually will be served.

Our two approximations satisfy an important consistency condition: The approximations are all exact for the special case of the $M/M/s/r+M$ model, which is sometimes referred to as the Erlang-A model. Indeed, the computational effort required for our algorithm is essentially the same as for the Erlang-A model, which is covered as a special case. The algorithm is very fast, so that it easily can be applied to determine appropriate staffing levels in $M/GI/s/r+GI$ systems. It can also serve as a component analysis tool in more complex systems.

We should also mention that Brandt and Brandt (2002) previously proposed a state-dependent Markovian approximation for abandonments in the $M(n)/M(n)/s+GI$ model, but their approximation is quite different, as we explain at the end of §3. Their primary focus is on the exact analysis of the $M(n)/M(n)/s+GI$ model (for which they have considerable success), rather than on simple engineering approximations.

Here is how the rest of this paper is organized: In §2, we start by presenting simulation results to show that it can be important to go beyond the corresponding Erlang-A model, obtained by using exponential service-time and abandon-time distributions with the given means. In §3, we introduce the state-dependent Markovian approximation for the abandonments. In §4, we present more simulation results to show that the Markovian approximations for abandonments are effective for the $M/M/s/r+GI$ model, which has exponential service times. In §5, we discuss the simple exponential approximation for the more general GI service times. In §6, we present additional simulation results to show that the $M/M/s/r+M(n)$ approximation is effective for the $M/GI/s/r+GI$ model. In §7, we derive the steady-state performance measures in the $M/M/s/r+M(n)$ model, most of which require additional approximations. In §8, we discuss fitting the model parameters to call-center data. Finally, in §9, we draw conclusions. Additional material appears in an online supplement (Whitt 2004b).

2. The Need to Go Beyond the Erlang-A Model

A natural first approximation to try for the $M/GI/s/r+GI$ queueing model is the more elementary Erlang-A model, $M/M/s/r+M$, where we obtain both the exponential time-to-abandon distribution and the exponential service-time distribution by using exponential distributions with the same means as the given general distributions. Our problem is interesting, in large part, because that natural simple approximation procedure often performs badly. In some cases, however, the Erlang-A model describes call-center performance quite well (see Brown et al. 2005). Certainly, the Erlang-A model is superior to the commonly used Erlang-C model ($M/M/s/\infty$).

To see that the Erlang-A model does not provide a consistently good approximation for the $M/GI/s/r+GI$ model, consider the $M/E_2/100/200+E_2$ model

with arrival rate $\lambda = 102$, individual mean service time $\mu^{-1} = 1$ and expected time to abandon of 1, where both the service time and the time to abandon have an Erlang-E_2 distribution, which is the sum of two IID exponentials. Since an E_k distribution has SCV $1/k$, the E_2 distributions here have mean 1, $SCV = 1/2$ and variance 1/2. In all our examples, we let the mean service time be 1. That is without loss of generality, because we are free to choose measuring units for time.

In Table 1 we compare simulations of the $M/E_2/100/200 + E_2$ and $M/M/100/200 + M$ models with the same arrival rate, mean service time and mean time to abandon. In this example and throughout this paper, we choose the waiting room size r sufficiently large so that blocking is negligible and so not a factor. All simulation experiments reported in this paper are based on 10 independent replications of runs each having five million arrivals. The independent replications make it possible to reliably estimate confidence intervals using the t-statistic. For all estimates, we show the half width of 95% confidence intervals.

Table 1 A Comparison of Steady-State Performance Measures for the $M/E_2/100/200 + E_2$ and $M/M/100/200 + M$ Models

Performance measure	$M/E_2/100/200 + E_2$		$M/M/100/200 + M$	
	Sim.	Approx.	Sim.	Exact
$P(W = 0)$	0.217	0.250	0.4092	0.4083
	±0.0021	—	±0.0013	—
$P(A)$	0.0351	0.0381	0.0498	0.0499
	±0.00029	—	±0.00020	—
$E[Q]$	11.52	11.41	5.073	5.092
	±0.075	—	±0.024	—
$\mathrm{Var}(Q)$	112.0	121.9	44.4	44.6
	±0.71	—	±0.30	—
$E[N]$	109.9	109.5	102.0	102.0
	±0.092	—	±0.036	—
$E[W \mid S]$	0.1115	0.1102	0.0489	0.0490
	±0.00071	—	±0.00023	—
$\mathrm{Var}(W \mid S)$	0.0101	0.0119	0.00418	0.0042
	±0.000061	—	±0.000027	—
$E[W \mid A]$	0.1508	0.1521	0.0665	0.0666
	±0.00042	—	±0.00021	—
$\mathrm{Var}(W \mid A)$	0.0067	0.0079	0.0031	0.0031
	±0.000044	—	±0.000018	—
$P(W \le 0.1 \mid S)$	0.510	0.528	0.7994	0.7986
	±0.0030	—	±0.0012	—
$P(W \le 0.1 \mid A)$	0.305	0.316	0.7678	0.7671
	±0.0014	—	±0.0013	—
$P(W \le 0.2 \mid S)$	0.795	0.786	0.9648	0.9644
	±0.0023	—	±0.00057	—
$P(W \le 0.2 \mid A)$	0.740	0.726	0.9705	0.9702
	±0.0019	—	±0.00054	—

Note. The two models have common arrival rate $\lambda = 102$, mean service time $\mu^{-1} = 1$ and mean time to abandon 1.0. The half width of the 95% confidence interval is given for each simulation estimate.

To define the performance measures we examine, let S be the event that a typical customer who enters who system (is not blocked) eventually will be served; let A be the event that a typical customer who enters the system abandons before starting service; let W be the steady-state waiting time (before beginning service or abandoning, whichever happens first) for a typical entering customer (conditional on the arrival not being blocked); let N be the steady-state number of customers in the system at an arbitrary time; and let $Q \equiv \max\{0, N - s\}$ be the steady-state queue length at an arbitrary time.

The performance measures we examine are: $P(W = 0)$, the probability an entering customer will not have to wait before beginning service; $P(A)$, the probability an entering customer will eventually abandon; $E[Q]$ and $\mathrm{Var}(Q)$, the mean and variance of the queue length at an arbitrary time; $E[N]$, the expected number of customers in the system at an arbitrary time; $E[W \mid S]$ and $\mathrm{Var}(W \mid S)$, the conditional mean and variance of the waiting time of an entering customer, given that the entering customer eventually will be served; $E[W \mid A]$ and $\mathrm{Var}(W \mid A)$, the conditional mean and variance of the waiting time of an entering customer, given that the entering customer eventually will abandon; $P(W \le t \mid S)$, the conditional probability that an entering customer waits less than time t, given that the customer eventually will be served; and $P(W \le t \mid A)$, the conditional probability that an entering customer waits less than time t, given that the customer eventually will abandon. We usually consider $t = 0.1$ and $t = 0.2$, corresponding to 10% and 20% of a mean service time. If the mean service time is 200 seconds, then $t = 0.1$ corresponds to 20 seconds; then the performance target of answering 80% of all answered calls within 20 seconds translates into $P(W \le 0.1 \mid S) \ge 0.8$.

In Table 1, we also display the numerical approximation results for the two models. The extremely close agreement between simulation results and numerical results for the $M/M/s/r + M$ model is to be expected because the formulas are exact in that case. Having both simulation and exact numerical results for the $M/M/s/r + M$ model provides an important check on both programs. For the $M/E_2/s/r + E_2$ model, the numerical results reveal the quality of the proposed approximations in that case. We regard the quality of the new approximation as excellent, even though one might want to do even better. For example, there is a 15% error in the approximation for the probability of no delay, $P(W = 0)$. However, there would be an 88% error if we used the Erlang-A model instead.

The simulation results in Table 1 show that performance in the $M/E_2/100/200 + E_2$ model is not too close to performance in the corresponding $M/M/100/200 + M$ model. For example, the mean queue length

with the Erlang distributions is 11.5, while it is 5.1 with the exponential distributions. Perhaps contrary to intuition, from the perspective of queue length and waiting time, the performance in the model with the less variable Erlang-(E_2) distributions is *significantly worse* than in the corresponding model with exponential (M) distributions. The E_2 distribution produces fewer abandonments than an exponential time-to-abandon distribution, and thus bigger queues and bigger delays.

It is also useful to see how the models compare from a decision perspective. Suppose that our goal is to determine an appropriate staffing level. Suppose that we want to determine the minimum number of servers so that the abandonment probability is less than 0.05 and the conditional probability of having to wait less than 0.1, given that the customer eventually will be served, is at least 0.80 (corresponding to the classic 80/20 rule mentioned above when the average call holding time is 200 seconds). Suppose that we fix the arrival rate at $\lambda = 100$ and let the remaining parameters be as above. For the $M/E_2/s/200 + E_2$ model, we find that the required number of servers is $s = 104$, whereas for the $M/M/s/200 + M$ model the required number is 99, a 5% difference. If we use the $M/M/s/200+M$ model and let the number of servers be 99, then in the actual $M/E_2/s/200 + E_2$ model the conditional probability of having to wait less than 0.1 mean service times, given that the customer eventually will be served, is only 0.58 instead of 0.80. Moreover, the mean queue length is 9.9 instead of 4.7, the value with $s = 104$. In contrast, our proposed approximation yields exactly the required number of servers for this $M/E_2/s/200 + E_2$ example.

Among all distributions on the positive real line, an Erlang-E_2 distribution is not too radically different from an exponential distribution. The Erlang-A model provides an even worse approximation for the $M/GI/s/r+GI$ model in other cases. For example, see the results for the $M/M/s/r + LN$ model in Table 4 below.

3. Markovian Approximation for Abandonments

The main new idea in this paper is to develop a state-dependent Markovian approximation for abandonments. With invisible queues, it is natural to assume at the outset that waiting customers have IID times to abandon with a general cdf F having a density f, with the clock starting the instant the customer joins the queue. As an approximation, we propose having a state-dependent Markovian approximation for abandonments. Specifically, we will assume that a customer who is jth from the end of a queue will abandon at rate α_j, independent of the rest of the history up to that point. We will first develop a way to define suitable infinitesimal rates α_j and then develop a way to approximately analyze the queue with those state-dependent rates.

The model with state-dependent Markovian abandonment rates arises naturally when customers are provided information about system state, as discussed in Whitt (1999). It is significant that we are *not* discussing that situation here. We are intending the state-dependent Markovian abandonments to serve as an approximation for the GI case that arises naturally with invisible queues, where customers are not given state information. Thus, from a direct modelling perspective, it is natural to expect that our approach might not work at all. If it does, in fact, work, then we may be able to apply the general Markovian $M(n)/M(n)/s/r + M(n)$ model with state-dependent rates to many call-center situations, both when state information is provided and when it is not.

When trying to understand the behavior of the $M/GI/s/r + GI$ model, an important initial insight is that, in contrast to single-server queues, waiting times in multiserver queues with a large number of servers tend to be quite small relative to the mean service times. This phenomenon is well known in call centers, and is reflected by the classical 80/20 rule. Since the mean length of the calls themselves tends to be 200 or 400 seconds or even longer, that implies that the waiting times tend to be only 10% or 5% of a mean service time or even less. Often, about half of the customers do not have to wait at all, even though there may be a 5% abandonment rate.

The tendency for waiting times in multiserver queues to be relatively small is also supported by the heavy-traffic limit theorems for multiserver queues in which the number of servers, s, increases along with the traffic intensity, ρ, so that

$$(1-\rho)\sqrt{s} \to \xi \quad \text{as } s \to \infty \tag{3.1}$$

for some constant ξ. In that limiting regime, the probability of delay approaches a proper limit strictly between 0 and 1 (see Halfin and Whitt 1981; Puhalskii and Reiman 2000; Garnett et al. 2002; Whitt 2002, Chapter 10; 2004a, 2005a; Jelenkovic et al. 2004; Mandelbaum and Zeltyn 2004). For our purposes, the important limit is for the waiting times; in the limit as $s \to \infty$, the waiting times are asymptotically negligible; specifically, they are of order $O(1/\sqrt{s})$. Since waiting times tend to be relatively small, we see that what matters about the time-to-abandon cdf F is its behavior for small-time arguments, not its moments or tail behavior.

If we knew that a customer had been waiting for time t, then the appropriate infinitesimal rate of abandonment for that customer at that time would be

given by the time-to-abandon hazard (or failure rate) function

$$h(t) = \frac{f(t)}{F^c(t)}, \quad t \geq 0, \tag{3.2}$$

where $f(t)$ is the density and $F^c(t) \equiv 1 - F(t)$ is the complementary cdf (ccdf) associated with the time-to-abandon cdf F.

To understand abandonment behavior, the key quantity is the hazard function h in (3.2) for relatively small-time arguments. Our experience indicates that performance is significantly affected by the form of the abandon-time hazard function h for small values of t, but the performance evidently is not too sensitive to the fine detail. Thus it may suffice to work with the first few terms of the Taylor series expansion about 0, e.g., by letting $h(t) \approx h(0) + h'(0)t + h''(0)t^2/2$. That has the advantage that it may be easier to fit to data. It may even suffice to work with the first nonzero term in this approximation. The main point is to use the approximate form of the hazard function for small-time arguments. In the process of doing this research, we discovered that similar ideas also have been advanced by Mandelbaum and Zeltyn (2004).

Given the hazard function h or an approximation to it, our goal is to produce, as an approximation, abandonment rates that depend on a customer's position in queue and the length of that queue. However, if the state is a customer's position in queue and the length of that queue, then we clearly do not know how long the customer has been waiting. What we propose to do, then is to estimate how long the customer has been waiting, given the available state information.

Suppose that we look at the number of customers in the system at an arbitrary time in steady state. Suppose that all s servers are busy and that there are k customers waiting in the queue. Given that information, we want to estimate how long each of the k customers in queue have been waiting. Suppose that we focus on the customer that is jth from the *end* of the queue, where $1 \leq j \leq k$. If there were no abandonments, then there would have been exactly $j-1$ arrivals since the customer in question arrived, and we would be in the middle of another interarrival time. Assuming that abandonments are relatively rare compared to service completions, we estimate that there have been j new arrival events since the customer who is jth from the end of the queue arrived. (This assumption is reasonable because we are aiming our approximation for the case of approximately 5% abandonments. Experience indicates that the approximation performs reasonably well even in the case of 20% abandonments, but it breaks down in extreme overload, e.g., in case of 50% abandonments.)

We now need to estimate the expected time between successive arrival events. A simple rough estimate for the average time between arrival events is $1/\lambda$, the reciprocal of the exogenous arrival rate. Thus, we propose as approximate state-dependent Markovian abandonment rates

$$\alpha_j \equiv h(j/\lambda), \quad 1 \leq j \leq k, \tag{3.3}$$

where λ is the exogenous arrival rate (not counting retrials) and h is the time-to-abandon hazard rate function in (3.2). The associated total abandonment rate from the queue in that state would be

$$\delta_k \equiv \sum_{j=1}^{k} \alpha_j = \sum_{j=1}^{k} h(j/\lambda). \tag{3.4}$$

In making the definitions above, we assume that the time-to-abandon cdf F has a density and that the density is relatively smooth. If the density were not smooth, we might instead let

$$\alpha_j \equiv \lambda \int_{(j-1)/\lambda}^{j/\lambda} h(t)\, dt, \quad 1 \leq j \leq k. \tag{3.5}$$

Then the approximate total abandonment rate would be

$$\delta_k \equiv \lambda \int_0^{k/\lambda} h(t)\, dt = -\lambda \log_e F^c(k/\lambda). \tag{3.6}$$

We close this section by briefly discussing the state-dependent Markovian approximation for GI abandonments in the $M(n)/M(n)/s+GI$ model developed by Brandt and Brandt (2002). Instead of developing an approximating rate α_j for the jth customer from the *end* of a queue of length k, they develop an approximate abandonment rate β_j for the jth customer from the *front* of the queue, which is based on detailed analysis of the $M(n)/M(n)/s+GI$ model. Moreover, they do not attempt to develop further approximations to describe customer experience with such state-dependent abandonment rates, as we do in §7. Brandt and Brandt (2002) focus much more on exact analysis.

4. Testing the Approximation for $M/M/s/r+GI$

In this section, we present simulation results to show that the Markovian approximation for abandonments proposed in §3 is effective for the $M/M/s/r+GI$ model, which has exponential service times. By separately considering the case of exponential service times, we separately evaluate the approximations for the abandon times and the service times. As we will demonstrate in §6, our experience indicates that the cruder service-time approximation causes greater errors. Before proceeding, it should be noted that many exact results can be computed for the $M/M/s/\infty+GI$ model, as shown by Brandt (1999, 2002) and Mandelbaum and Zeltyn (2004). These papers should be consulted for additional insights.

Table 2 A Comparison of Approximations for Steady-State Performance Measures with Simulations in Two Models with Exponential Service Times, Arrival Rate $\lambda = 102$, and Mean Abandon Time 1

Performance measure	$M/M/100/200 + E_2$		$M/M/100/200 + LN(1,1)$		$M/M/100/200 + M$
	Sim.	Approx.	Sim.	Approx.	Exact
$P(W = 0)$	0.246	0.250	0.242	0.247	0.408
	±0.0020	—	±0.0026	—	—
$P(A)$	0.0378	0.0381	0.0376	0.0379	0.0499
	±0.00032	—	±0.00032	—	—
$E[Q]$	11.75	11.41	11.42	11.02	5.09
	±0.075	—	±0.071	—	—
$Var(Q)$	129.2	121.9	115.6	107.2	44.6
	±0.94	—	±0.46	—	—
$E[N]$	109.9	109.5	109.6	109.1	102.0
	±0.091	—	±0.092	—	—
$E[W \mid S]$	0.1133	0.1102	0.1094	0.1058	0.0490
	±0.00072	—	±0.00067	—	—
$Var(W \mid S)$	0.0119	0.0113	0.0104	0.0097	0.0042
	±0.000083	—	±0.000042	—	—
$E[W \mid A]$	0.1628	0.1521	0.1788	0.1642	0.0666
	±0.00063	—	±0.00026	—	—
$Var(W \mid A)$	0.0079	0.0076	0.0054	0.0054	0.0031
	±0.000061	—	±0.000024	—	—
$P(W \le 0.1 \mid S)$	0.520	0.528	0.518	0.527	0.799
	±0.0026	—	±0.0028	—	—
$P(W \le 0.1 \mid A)$	0.273	0.316	0.140	0.204	0.767
	±0.0019	—	±0.00064	—	—
$P(W \le 0.2 \mid S)$	0.775	0.786	0.792	0.807	0.964
	±0.0023	—	±0.0018	—	—
$P(W \le 0.2 \mid A)$	0.688	0.726	0.644	0.706	0.970
	±0.0027	—	±0.00066	—	—

Note. The two models have Erlang-E_2 and lognormal $LN(1,1)$ abandon-time distributions.

From these papers, we see that we could instead use the more complicated *exact* solution of the $M/M/s/r + GI$ model to approximate performance in the $M/GI/s/r + GI$ model. However, we believe that there is not great incentive for doing so, because the approximation for the $M/M/s/r + GI$ model is remarkably accurate and because most of the error in approximating the $M/GI/s/r + GI$ model that we really want to consider is because of the service-time approximation.

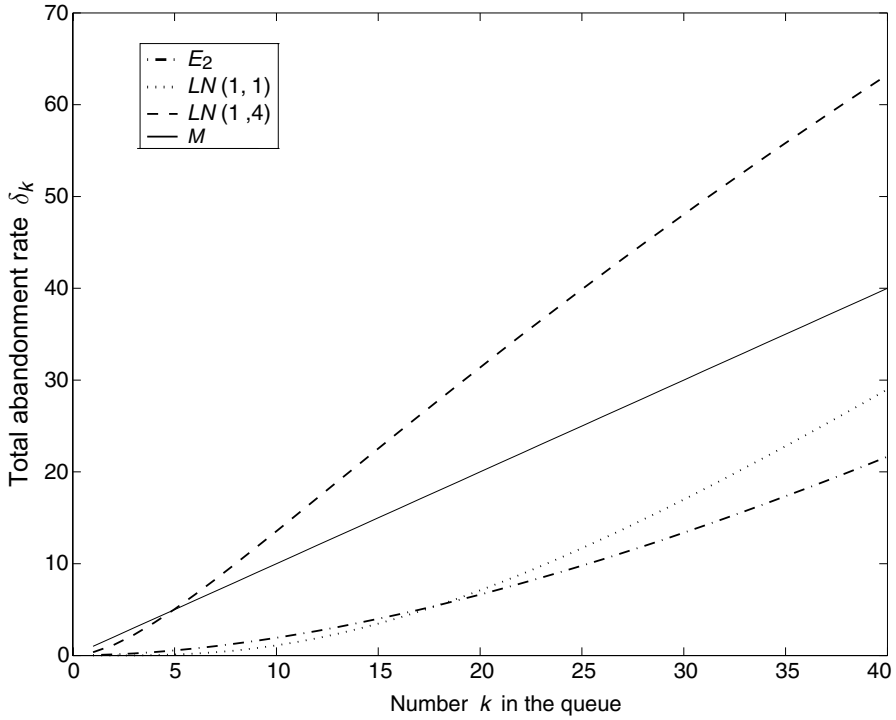
In Table 2, we show results for the $M/M/100/200 + GI$ model with Erlang and lognormal abandon times, common arrival rate $\lambda = 102$, and mean abandon time 1. By $LN(a, b)$, we mean a lognormal distribution with mean a and SCV b. Thus the lognormal $LN(1, 1)$ abandon time has $SCV = 1$ and variance 1. We also display the exact numerical results for the corresponding $M/M/100/200 + M$ model for comparison. From Table 2, we see that the approximations agree quite closely with the simulations. For example, the approximation error for the probability of no delay, $P(W = 0)$, in the $M/M/s/r + E_2$ model is only 2%, compared to 15% in the $M/E_2/100/200 + E_2$ model in Table 1. As in Table 1, the steady-state performance measures are quite different from the associated Erlang-A model.

With $s = 100$ servers each working at rate 1, the arrival rate $\lambda = 102$ is a relatively heavy load. We consider that case in most of our examples throughout this paper. For each of these examples, we also performed simulations with arrival rates $\lambda = 98$ and $\lambda = 90$. The quality of the approximation for the $M/M/s/r + GI$ model at these lighter loads is consistently better. That should be expected because abandonments are less frequent. Some results for the case $\lambda = 90$ are in the online supplement (Whitt 2004b).

Even though the E_2 and $LN(1, 1)$ distributions are quite different, Table 2 shows that the performance with these two abandon-time distributions is quite close. That is easy to understand when we look at the hazard functions and the approximate total abandonment rates δ_k produced by the approximation in (3.3)–(3.4). To make that clear, we plot the resulting function δ_k for four different abandon-time distributions in Figure 1. (Since the queue length only rarely exceeds 40, we plot δ_k for $0 \le k \le 40$. Since $\alpha_j = h(j/\lambda)$ when $k \le 40$, the hazard function is only relevant over the initial subinterval $[0, 0.4]$.)

For comparison, we include the abandonment rate δ_k for the exponential (M) and $LN(1, 4)$ distributions in Figure 1, in addition to the E_2 and $LN(1, 1)$ distributions used in Table 2. From Figure 1, we see that

Figure 1 A Comparison of Four Abandon-Time Distributions



Note. The approximate total abandonment rate δ_k when there are k customers in the queue in the $M/M/s/r + GI$ model and with four different abandon-time distributions having mean 1: Erlang E_2, lognormal $LN(1, 1)$ and $LN(1, 4)$, and exponential M.

for small-time arguments, the hazard function and the total abandonment rate approximations are quite close for the E_2 and $LN(1, 1)$ distributions, and these two are quite different from the other two. Consistent with Figure 1, both the approximations and the simulations are close for the M and $LN(1, 4)$ abandon-time distributions (not shown here).

To show some other cases, we present two additional tables. In Tables 3–4, we show results for lognormal abandon-time distributions with a greater mean, 4. The first lognormal distribution has SCV 4, and thus variance 64, while the second has SCV 0.25, and thus variance 4. Again, the approximations agree closely with the simulation results. For $LN(4, 4)$ in Table 3, the performance is similar to that of the Erlang-A model, but for $LN(4, 0.25)$ in Table 4, the performance is entirely different from that of the corresponding Erlang A (with the same mean service time and mean abandon time). Since the congestion is much greater in the $LN(4, 0.25)$ case, we make the number of waiting spaces larger to avoid significant blocking, in particular, we let $r = 300$.

As illustrated by Tables 2–4, simulation results show that the $M/M/s/r + M(n)$ approximation for the $M/M/s/r + GI$ model performs remarkably well. Overall, we find the weakest part of our approximation is the approximation for the nonexponential service times (see §6).

Table 3 A Comparison of Approximations for Steady-State Performance Measures with Simulations in the $M/M/100/200 + LN(4, 4)$ Model with Arrival Rate $\lambda = 102$

Performance measure	$M/M/100/200 + LN(4, 4)$		c$M/M/100/200 + M$
	Sim.	Approx.	Exact
P($W = 0$)	0.210	0.212	0.226
	±0.0019	—	—
P(A)	0.0349	0.0353	0.0364
	±0.00030	—	—
E[Q]	14.90	14.61	14.84
	±0.095	—	—
Var(Q)	187.0	180.1	214.5
	±1.37	—	—
E[N]	113.3	113.0	113.1
	±0.023	—	—
E[$W \mid S$]	0.1446	0.1419	0.1455
	±0.00091	—	—
Var($W \mid S$)	0.0175	0.0169	0.0207
	±0.00013	—	—
E[$W \mid A$]	0.1878	0.1786	0.1429
	±0.00048	—	—
Var($W \mid A$)	0.0105	0.0105	0.0137
	±0.000048	—	—
P($W \le 0.1 \mid S$)	0.444	0.449	0.469
	±0.0025	—	—
P($W \le 0.1 \mid A$)	0.212	0.248	0.449
	±0.0010	—	—
P($W \le 0.2 \mid S$)	0.680	0.687	0.687
	±0.0028	—	—
P($W \le 0.2 \mid A$)	0.602	0.632	0.737
	±0.0023	—	—

Table 4 **A Comparison of Approximations for Steady-State Performance Measures with Simulations in the $M/M/100/300+LN(4, 0.25)$ Model with Arrival Rate $\lambda = 102$**

Performance measure	$M/M/100/300+LN(4, 0.25)$		$M/M/100/300+M$
	Sim.	Approx. numerical	Exact numerical
$P(W=0)$	0.0096	0.0101	0.226
	±0.00082	—	—
$P(A)$	0.0206	0.0204	0.0364
	±0.00029	—	—
$E[Q]$	118.1	117.0	14.84
	±0.75	—	—
$E[N]$	218.0	216.9	113.1
	±0.75	—	—
$E[W \mid S]$	1.154	1.144	0.1455
	±0.0073	—	—
$E[W \mid A]$	1.327	1.288	0.1429
	±0.0015	—	—
$P(W \le 0.4 \mid S)$	0.0702	0.0710	0.469
	±0.0032	—	—
$P(W \le 0.4 \mid A)$	0.000093	0.0000	0.449
	±0.0032	—	—

Additional experiments are reported in the online supplement (Whitt 2004b). There we show that the approximation still performs quite well with fewer servers and light loads. There we also show that the approximation still performs well under heavy loads, e.g., for $s = 100$ and $\lambda = 120$ and for $s = 20$ and $\lambda = 24$. The approximation even performs well when $s = 100$ and $\lambda = 200$.

5. Treating the Service Times

In §3, we developed a state-dependent Markovian approximation for abandonments, which replaces the original $M/GI/s/r + GI$ model by the associated $M/GI/s/r + M(n)$ model, where $M(n)$ denotes state-dependent Markovian abandonments. Unfortunately, however, when the service-time distribution is not exponential, the new $M/GI/s/r + M(n)$ model is also very difficult to analyze exactly, so we need to make further approximations. We propose approximating the given general service-time distribution simply by an exponential service-time distribution with the same mean. We thereby obtain the totally Markovian $M/M/s/r + M(n)$ approximation for the original $M/GI/s/r + GI$ model. We show how to analyze this Markovian model in §7.

We primarily make this second model approximation because it produces a Markovian model that we can analyze. However, unlike the direct approximation by the full Erlang-A model, this step also turns out to be relatively accurate. That may be surprising, because the same approximation for the classical single-server $M/GI/1/\infty$ model would be terrible. For example, the mean steady-state waiting time in the $M/GI/1/\infty$ model is proportional to $1 + c_s^2$, where c_s^2 is the SCV of the service-time distribution. When c_s^2 is not nearly 1, the $M/M/1$ approximation would be very bad. However, the situation is very different when there is a large number of servers.

An important theoretical reference point is the well-known insensitivity of the Erlang loss model (also known as the Erlang-B model and $M/GI/s/0$). In the Erlang loss model, the steady-state distribution does not depend on a general service-time distribution beyond its mean. Thus the approximation we are making is exact for the $M/GI/s/0$ special case, which occurs in the limit as the abandonments get fast.

A second important theoretical reference point is the $M/GI/\infty$ model, which also has the service-time insensitivity property. Under light loads, the $M/GI/s/r + GI$ model will behave like the associated $M/GI/\infty$ model, where the service-time distribution beyond the mean has no impact on the steady-state distribution. Hence, as is borne out in simulations, we should anticipate that our approximations tend to perform better in light loads. For that reason, our examples focus more on heavier loads.

On the other hand, it is well known that the insensitivity to the service-time distribution beyond its mean in the Erlang loss system and the associated infinite-server system does *not* hold for the corresponding Erlang delay model (also known as the Erlang-C model or $M/M/s/\infty$) or the associated intermediate finite waiting room models $M/M/s/r$. However, the dependence on the service-time distribution is much less when there are multiple servers. For smaller numbers of servers, there is ample evidence, e.g., see Seelen et al. (1985) and Whitt (1993). For the larger numbers of servers common in call centers, the impact of the service-time distribution on the performance of the $M/G/s/\infty$ model can be seen from simulations by Mandelbaum and Schwartz (2002) (there $s = 100$). Since the $M/GI/s/r + GI$ model approaches the $M/GI/s/r$ model as the mean abandon time increases, we can use those no-abandonment models to see the limitations of our proposed procedure in general.

Under heavier loads, the insensitivity we are using as an approximation becomes much more reasonable because of the abandonments as well as the large number of servers, but we recognize that it is a relatively crude approximation. Assuming that abandonments are indeed occurring at a sufficient rate, the abandonments make the $M/GI/s/r + GI$ model more like the $M/GI/s/0$ model instead of the $M/GI/s/\infty$ model. As simulations show, when there is a reasonable level of abandonment, the $M/M/s/r+GI$ model is a reasonable approximation for the $M/GI/s/r+GI$ model, and our approximating $M/M/s/r + M(n)$ model is a reasonable approximation for both the $M/GI/s/r+M(n)$ and $M/GI/s/r+GI$ models.

A third relevant theoretical reference point is the diffusion approximation for the $G/GI/s/r$ model developed in Whitt (2004a), based on the heavy-traffic limit for the $G/H_2^*/s/r$ model established in Whitt (2005a). The special H_2^* service times are mixtures of an exponential distribution and an atom point mass at zero. The H_2^* service-time distribution is appealing because it leads to a one-dimensional Markov limit process for the number of customers in the system, but at the same time, it permits a two-parameter characterization of the service-time distribution, with one parameter characterizing the mean and the other characterizing the variability.

It turns out that in the special case of a Poisson arrival process (the $M/GI/s/r$ model), the proposed diffusion approximation does not depend greatly on the service-time distribution beyond its mean. Indeed, for the special case of a Poisson arrival process, the approximate probability of delay and the approximate conditional distribution of the number of busy servers, given that all servers are not busy, are independent of the service-time distribution beyond its mean. Moreover, if in addition the service-time distribution has $SCV = 1$, then the entire diffusion approximation is independent of the service-time distribution beyond its mean. Consistent with that theoretical-based approximation, our approximations tend to perform better when the service-time SCV is close to 1.

A fourth important theoretical reference point is the heavy-traffic fluid limit for the $M/GI/s+GI$ model in the overloaded or efficiency-driven (ED) regime, characterized by $s \to \infty$ and $\lambda \to \infty$ with $\mu = 1$ and $\rho \equiv \lambda/s\mu > 1$ held fixed (see Whitt 2004c, 2005b, c). The steady-state performance in the ED regime depends strongly upon the time-to-abandon distribution, but does not depend upon the service-time distribution beyond its mean.

6. Testing the General Approximation

We now evaluate the approximation of the general GI service-time distribution in the $M/GI/s/r + GI$ model by an exponential distribution with the same mean. We want to show that the performance in the $M/GI/s/r+GI$ model tends to depend on the service-time distribution primarily only through its mean, so that we can approximate the $M/GI/s/r+GI$ model by the corresponding $M/M/s/r + GI$ model. Combined with the Markovian approximation for abandonments developed in §3, we thus obtain the full approximation by a $M/M/s/r+M(n)$ model.

One such test was already performed in Table 1. There we compared the approximation to simulations of the $M/E_2/100/200 + E_2$ model with arrival rate $\lambda = 102$, mean service time $\mu^{-1} = 1$ for the case of mean abandon time $= 1$. We have also considered the Erlang model with different mean abandon times. Again, the approximation is effective. For smaller mean abandon times, such as $= 0.25$, the results are quite close to the Erlang-A model, but they are entirely different for larger mean abandon times. To illustrate, we show the case of mean abandon time 4.0 in Table 5.

Table 5 **A Comparison of Steady-State Performance Measures in the $M/E_2/100/200+E_2$ and $M/M/100/200+M$ Model with Mean Time to Abandon $= 4.0$. and Arrival Rate $\lambda = 102$**

Performance measure	$M/E_2/100/200+E_2$		$M/M/100/200+M$
	Sim.	Approx. numerical	Exact numerical
$P(W = 0)$	0.056 ±0.0016	0.0764 —	0.226 —
$P(A)$	0.0236 ±0.00036	0.0253 —	0.0364 —
$E[Q]$	41.6 ±0.44	41.8 —	14.84 —
$E[N]$	141.2 ±0.39	141.2 —	113.1 —
$E[W \mid S]$	0.407 ±0.0042	0.409 —	0.1455 —
$E[W \mid A]$	0.413 ±0.0023	0.430 —	0.1429 —
$P(W \le 0.1 \mid S)$	0.133 ±0.0032	0.161 —	0.4688 —
$P(W \le 0.1 \mid A)$	0.046 ±0.00078	0.050 —	0.4493 —
$P(W \le 0.2 \mid S)$	0.234 ±0.0047	0.261 —	0.6865 —
$P(W \le 0.2 \mid A)$	0.166 ±0.0025	0.164 —	0.7366 —

In the next tables we look at $M/GI/s/r + GI$ models with common time-to-abandon distributions, but different service-time distributions having a common mean. In Table 6, we consider $M/GI/100/200 + LN(1, 1)$ models with common lognormal abandon-time distribution having mean $= 1.0$ and $SCV = 1.0$; and in Table 7, we consider $M/GI/100/200 + E_2$ models with common E_2 abandon-time distribution having mean $= 1.0$. In each case, we consider several different service-time distributions from among: D (deterministic), E_2, M, $LN(1, 1)$, and $LN(1, 4)$. The results show, first, that the performance is indeed largely independent of the service-time distribution beyond its mean and, second, that the approximation performs remarkably well. However, the approximation is better with M service times than with the nonexponential service-time distributions. As the service-time distribution deviates more from the exponential distribution, the approximation performs worse. Consistent with the diffusion approximation for the $M/GI/s/r$ model in Whitt (2004a), the performance degrades as the service-time SCV deviates more from 1, the SCV of an exponential distribution. In particular, we see degradation of performance for

Table 6 **A Comparison of Simulation Estimates of Steady-State Performance Measures in $M/GI/100/200 + LN(1,1)$ Models with Four Different Service-Time Distributions Having Common Mean** 1.0

Performance measure	Service-time distribution				
	E_2	M	$LN(1,1)$	$LN(1,4)$	*Approx.*
$P(W=0)$	0.211	0.242	0.229	0.286	0.247
	±0.0013	±0.0026	±0.0015	±0.0020	—
$P(A)$	0.0348	0.0376	0.0366	0.0425	0.0379
	±0.00021	±0.00032	±0.00024	±0.00021	—
$E[Q]$	11.40	11.42	11.44	11.55	11.02
	±0.039	±0.071	±0.051	±0.048	—
$\mathrm{Var}(Q)$	102.7	115.6	110.6	137.6	107.2
	±0.39	±0.46	±0.43	±0.49	—
$E[N]$	109.9	109.6	109.7	109.2	109.1
	±0.053	±0.092	±0.062	±0.071	—
$E[W \mid S]$	0.1097	0.1094	0.1098	0.1096	0.1058
	±0.00037	±0.00067	±0.00047	±0.00045	—
$\mathrm{Var}(W \mid S)$	0.0091	0.0104	0.0099	0.0126	0.0097
	±0.000030	±0.000042	±0.000037	±0.000047	—
$E[W \mid A]$	0.1696	0.1788	0.1753	0.1940	0.1642
	±0.00025	±0.00026	±0.00025	±0.00041	—
$\mathrm{Var}(W \mid A)$	0.0047	0.0054	0.0051	0.0068	0.0054
	±0.000031	±0.000024	±0.000023	±0.000048	—
$P(W \le 0.1 \mid S)$	0.502	0.518	0.511	0.542	0.527
	±0.0016	±0.0028	±0.0021	±0.0020	—
$P(W \le 0.1 \mid A)$	0.157	0.140	0.146	0.117	0.204
	±0.00099	±0.00064	±0.00067	±0.00075	—
$P(W \le 0.2 \mid S)$	0.807	0.792	0.797	0.773	0.807
	±0.0011	±0.0018	±0.0016	±0.0011	—
$P(W \le 0.2 \mid A)$	0.693	0.644	0.661	0.571	0.706
	±0.0016	±0.00066	±0.0015	±0.0019	—

Note. E_2 with $SCV = 0.5$, M with $SCV = 1.0$, $LN(1,1)$ with $SCV = 1.0$ and $LN(1,4)$ with $SCV = 4.0$. The models have common arrival rate $\lambda = 102$ and $LN(1,1)$ abandon-time distribution.

the $LN(1,4)$ service time in Table 6 and the D service time in Table 7, but even in these cases the errors are not too great.

7. Steady-State Distribution of the Markovian Model

We now show how to calculate all the standard performance measures for the Markovian $M/M/s/r + M(n)$ call-center model. We start by calculating the steady-state distribution of the basic birth-and-death process. Then we describe the experience of entering customers, which requires further approximation. When we calculate waiting-time distributions, we will exploit numerical inversion of Laplace transforms, using the EULER algorithm in Abate and Whitt (1995), as already done in Whitt (1999). See Abate et al. (1999) for an overview of the inversion algorithms.

7.1. Steady-State Distribution of the Birth-and-Death Process

Let $N(t)$ be the number of customers in the system at time t. In the $M/M/s/r + M(n)$ queueing model, the stochastic process $\{N(t)\colon t \ge 0\}$ is a birth-and-death process. The birth rate is the arrival rate λ. The death rate μ_k is simply the total service rate when all servers are not busy, but when there is at least one customer waiting in queue, the death rate is the sum of the total service rate and the total abandonment rate. In particular, the death rate in state k is

$$\mu_k = \begin{cases} k\mu, & 1 \le k \le s, \\ s\mu + \delta_{k-s}, & s+1 \le k \le s+r, \end{cases} \tag{7.1}$$

where μ is the individual service rate and δ_k is the total state-dependent abandonment rate when there are k customers waiting in queue (obtained from (3.3)–(3.4) in our approximation of $M/GI/s/r + GI$).

Because the state space is finite, there is always a unique proper limiting steady-state distribution. Let N be a random variable with the limiting steady-state distribution of $N(t)$. The steady-state distribution is

$$p_k \equiv \mathrm{P}(N = k) \equiv \lim_{t\to\infty} \mathrm{P}(N(t) = k \mid N(0) = i). \tag{7.2}$$

The steady-state probabilities are determined by the local balance equations

$$p_k\lambda = p_{k+1}\mu_{k+1}, \quad 0 \le k \le s+r-1. \tag{7.3}$$

It is convenient to calculate the steady-state distribution recursively. Since the probability p_s is likely to be near the largest probability p_j (assuming that the number s of servers has been chosen in a reasonable manner), it is natural to start at s and separately go

Table 7 A Comparison of Simulation Estimates of Steady-State Performance Measures in $M/GI/100/200+E_2$ Models with Four Different Service-Time Distributions Having Common Mean 1.0

Performance measure	Service-time distribution				
	D	E_2	M	$LN(1,1)$	*Approx.*
$P(W=0)$	0.180	0.217	0.246	0.233	0.250
	±0.0013	±0.0021	±0.0020	±0.0021	—
$P(A)$	0.0309	0.0351	0.0378	0.0370	0.0381
	±0.00017	±0.00029	±0.00032	±0.00027	—
$E[Q]$	11.08	11.52	11.75	11.74	11.41
	±0.042	±0.075	±0.075	±0.063	—
$Var(Q)$	89.3	112.0	129.2	123.3	121.9
	±0.40	±0.71	±0.94	±0.72	—
$E[N]$	109.9	109.9	109.9	110.0	109.5
	±0.049	±0.092	±0.091	±0.72	—
$E[W \mid S]$	0.1078	0.1115	0.1133	0.1133	0.1102
	±0.00038	±0.00071	±0.00072	±0.00061	—
$Var(W \mid S)$	0.0079	0.0101	0.0119	0.0113	0.0113
	±0.000032	±0.000061	±0.000083	±0.000061	—
$E[W \mid A]$	0.1343	0.1508	0.1628	0.1589	0.1521
	±0.00028	±0.00042	±0.00063	±0.00039	—
$Var(W \mid A)$	0.0051	0.0067	0.0079	0.0075	0.0076
	±0.000028	±0.000044	±0.000061	±0.000047	—
$P(W \le 0.1 \mid S)$	0.501	0.510	0.520	0.514	0.528
	±0.0018	±0.0030	±0.0026	±0.0025	—
$P(W \le 0.1 \mid A)$	0.358	0.305	0.273	0.283	0.316
	±0.0014	±0.0014	±0.0019	±0.00088	—
$P(W \le 0.2 \mid S)$	0.833	0.795	0.775	0.780	0.786
	±0.0013	±0.0023	±0.0023	±0.0020	—
$P(W \le 0.2 \mid A)$	0.818	0.740	0.688	0.705	0.726
	±0.0013	0.0019	±0.0027	±0.0018	—

Note. E_2 with $SCV = 0.5$, M with $SCV = 1.0$, $LN(1,1)$ with $SCV = 1.0$, and $LN(1,4)$ with $SCV = 4.0$. The models have common arrival rate $\lambda = 102$ and E_2 abandon-time distribution.

up and down. For that purpose, let $x_s = 1$,

$$x_{s+k+1} = \frac{\lambda x_{s+k}}{\mu_{s+k+1}}, \quad 0 \le k \le r-1, \tag{7.4}$$

and

$$x_{k-1} = \frac{\mu_k x_k}{\lambda}, \quad 1 \le k \le s. \tag{7.5}$$

We then normalize to get the steady-state probabilities themselves. To do so, let the sum be

$$y = \sum_{k=0}^{s+r} x_k. \tag{7.6}$$

Then the steady-state probabilities are

$$p_k = x_k/y, \quad 0 \le k \le s+r. \tag{7.7}$$

Let $Q(t) \equiv \max\{0, N(t)-s\}$ be the queue length at time t and let $Q \equiv \max\{0, N-s\}$ be the steady-state queue length. We obtain the distribution of Q directly from the distribution of N above.

7.2. The Probability of Being Served or Abandoning

We now start to describe the experience of individual customers. Since the arrival process is Poisson, the state seen by arrivals is the same as at an arbitrary time, by the Poisson-Arrivals-See-Time-Average property (see Wolff 1989, §5.16). Thus the probability that an arrival is blocked and lost is simply $P(Loss) = p_{s+r}$. Henceforth we focus on the customers who enter the system. The probability that an admitted or entering customer finds k customers in the system is

$$p_k^a = \frac{p_k}{(1-P(Loss))} = \frac{p_k}{1-p_{s+r}}. \tag{7.8}$$

Our approach is to condition on the state seen by arrivals that enter the system and then average over all the possibilities. Let S be the event that a customer who enters the system eventually receives service and let A be the event that a customer who enters the system eventually abandons. Let W be the waiting time in queue for a customer who enters the system. First, the probability that an arriving customer who enters the system does not wait at all before starting service is exactly

$$P(NoWait) \equiv P(W=0) = \sum_{k=0}^{s-1} p_k^a. \tag{7.9}$$

The situation is more complicated when the arrival must join the queue. To analyze these situations, we

will make more approximations. Conditional on the arrival seeing $s+k-1$ customers in the system upon arrival (s customers in service and $k-1$ others already in the queue waiting), customers arriving after that customer play no role in that customer's experience. After that customer arrives, there will be $s+k$ customers in the system, with the new arrival at the end of the queue. Thus it suffices to consider the evolution of the system starting at level $s+k$, ignoring all future arrivals. Accordingly, to do further analysis, we consider the system starting at level $s+k$ and ignore future arrivals.

In that framework, we assume that successive departures (including abandonments) occur according to the minimum of independent exponential random variables. Thus we let the successive identities of departing customers and the successive intervals between departures be mutually independent random variables. Let $\gamma_{k,j}$ be the probability that the customer initially kth in line abandons in the jth subsequent departure event (among the original $s+k$ customers), given that the customer has not abandoned previously. Let $m_{k,j}$ be the mean time between the $(j-1)$st and jth departure events (where the 0th departure event occurs at time 0). We approximate these quantities by

$$\gamma_{k,j} \approx \frac{\alpha_j}{s\mu + (\delta_k - \delta_{j-1})} \tag{7.10}$$

and

$$m_{k,j} \approx \frac{1}{s\mu + (\delta_k - \delta_{j-1})} \tag{7.11}$$

for $1 \le j \le k$, where $\delta_0 \equiv 0$.

Approximation formulas (7.10)–(7.11) require explanation, which we will do below. First, note that for the $M/M/s/r+M$ model, in which $\alpha_j = \alpha$ for all j, these approximations are exact. Then

$$\gamma_{k,j} = \frac{\alpha}{s\mu + (k-j+1)\alpha} \quad \text{and}$$
$$m_{k,j} = \frac{1}{s\mu + (k-j+1)\alpha}, \tag{7.12}$$

where $\delta_0 \equiv 0$. Thus our approximate algorithm produces the exact performance measures for the $M/M/s/r+M$ model.

We now explain how we derived approximations (7.10)–(7.11). As indicated above, we start by ignoring future arrivals. At time 0—the arrival epoch of the arriving customer of interest (the last customer in the queue of length k)—we assume that the abandonment rates are as specified previously, i.e., the abandonment rate for the customer jth from the end of the queue is $\alpha_j \equiv h(j/\lambda)$, as in (3.3). There is no difficulty for the first departure; it is easy to see that formulas (7.10)–(7.11) are exact for $j=1$. We indeed have exactly the minimum of independent exponential random variables. However, there are problems when we consider subsequent departures.

To consider subsequent departures, we need to consider system dynamics over time: First, as time evolves, the waiting customers are spending more time in the system, so that their abandonment rates should change. To keep within the present framework, we want to work with the abandonment rates α_j defined in terms of the hazard function in (3.3). Because time is evolving, the hazard function should apply to a larger time argument. Here is what we do: As a further approximation, we act in this step of the approximation as if each successive departure epoch takes time $1/\lambda$. Thus, after m departures, $1 \le m \le k-j$, we let a customer who was jth from the end of the queue, if he is still present, have an abandonment rate that changes from $\alpha_j = h(j/\lambda)$ to $\alpha_{j+m} = h((j+m)/\lambda)$. As a consequence, the customer who was originally last in the queue has abandonment rate α_j for the jth departure epoch.

Even though we have specified the operative rates at successive departure epochs by the approximation above, we still need to do more in the approximation, because the evolution of the system depends on which customer departs at each departure epoch. We obtain approximations (7.10)–(7.11) by acting at any departure epoch as if all previous departures were service completions. That implies that the remaining total rate before the jth departure event should be approximately $\delta_k - \delta_{j-1}$, where $\delta_0 = 0$. Here is more explanation: With the assumption that all previous departures were service completions, the first $j-1$ customers waiting in queue, which had initial rates $(\alpha_{k-j+1}, \ldots, \alpha_k)$, have gone into service, while the remaining $k-j+1$ customers, which had initial rates $(\alpha_1, \ldots, \alpha_{k-j+1})$, have had their rate indices increase by $j-1$ to $(\alpha_j, \ldots, \alpha_k)$. By that reasoning, we obtain approximations (7.10)–(7.11). Ultimately, however, we use these approximations because they evidently work.

Given the approximations in (7.10)–(7.11), we can calculate associated performance measures. First, the probability that customer $s+k$ eventually receives service is

$$\Gamma_k = (1-\gamma_{k,1})(1-\gamma_{k,2})\cdots(1-\gamma_{k,k}) \tag{7.13}$$

for $\gamma_{k,j}$ in (7.10).

We now can (approximately) express the probability that a new arrival who enters the system eventually completes service; it is

$$\mathrm{P}(S) = \left(\sum_{k=0}^{s-1} p_k^a\right) + \sum_{k=0}^{r-1} p_{s+k}^a \Gamma_{k+1} \tag{7.14}$$

for Γ defined in (7.13), drawing on the approximations in (7.10).

Because all customers who enter the system and are not served must abandon, we can express the steady-state probability that an arrival who enters the system eventually abandons as

$$\mathrm{P}(A) = 1 - \mathrm{P}(S). \tag{7.15}$$

7.3. The Waiting Time for Customers Who Are Served

Let W be the waiting time (until beginning service) for a customer who enters the system. We want to differentiate between customers who eventually are served and customers who eventually abandon, so in this subsection we consider only entering customers who are served.

We now compute the first two moments of W for served customers, i.e., we compute $\mathrm{E}[W^j; S] = \mathrm{E}[W^j 1_S]$, where 1_B is the indicator function of the event B ($1_B(\omega) = 1$ if $\omega \in B$, and $1_B(\omega) = 0$ otherwise). We exploit the approximations in the last subsection, acting if the successive intervals between departures are independent exponential random variables with the means in (7.11). In using properties of the exponential distribution, we obtain

$$\mathrm{E}[W; S] = \sum_{k=0}^{r-1} p^a_{s+k} \Gamma_{k+1} \sum_{j=1}^{k+1} m_{k+1,\,j} \tag{7.16}$$

and

$$\mathrm{E}[W^2; S] = \sum_{k=0}^{r-1} p^a_{s+k} \Gamma_{k+1} (V_{k+1} + M^2_{k+1}), \tag{7.17}$$

where

$$V_{k+1} \equiv \sum_{j=1}^{k+1} m^2_{k+1,\,j} \tag{7.18}$$

and

$$M_{k+1} \equiv \sum_{j=1}^{k+1} m_{k+1,\,j}. \tag{7.19}$$

Then the first and second moments of the conditional waiting time given that the customer eventually completes service are

$$\mathrm{E}(W \mid S) = \frac{\mathrm{E}[W; S]}{\mathrm{P}(S)} \quad \text{and}$$
$$\mathrm{E}(W^2 \mid S) = \frac{\mathrm{E}[W^2; S]}{\mathrm{P}(S)}. \tag{7.20}$$

The conditional variance is then

$$\mathrm{Var}(W \mid S) \equiv \mathrm{E}(W^2 \mid S) - (\mathrm{E}(W \mid S))^2. \tag{7.21}$$

We can characterize the waiting-time distributions via their Laplace transforms. Then we can apply numerical transform inversion to calculate the distributions. For that purpose, let $\hat{w}_s(z) \equiv \mathrm{E}[e^{-zW} 1_{\{S,\, W>0\}}]$ be the Laplace transform of W for served customers who are not served immediately (Laplace-Stieltjes transform of its cdf). Paralleling (7.16), we have

$$\hat{w}_s(z) = \sum_{k=0}^{r-1} p^a_{s+k} \Gamma_{k+1} \hat{e}_{k+1}(z), \tag{7.22}$$

where

$$\hat{e}_{k+1}(z) \equiv \prod_{j=1}^{k+1} \left(\frac{m^{-1}_{k+1,\,j}}{m^{-1}_{k+1,\,j} + z} \right). \tag{7.23}$$

We can now calculate the cdf by numerical transform inversion. Specifically, we obtain the cdf $\mathrm{P}(0 < W \le t; S)$ for any desired t by numerically inverting its Laplace transform $\hat{w}_s(z)/z$, e.g., by using the Fourier-series method described in Abate and Whitt (1995). The associated conditional waiting-time cdf is

$$\mathrm{P}(W \le t \mid S) = \frac{\mathrm{P}(W = 0) + \mathrm{P}(0 < W \le t; S)}{\mathrm{P}(S)}. \tag{7.24}$$

7.4. The Time to Abandon

As in (7.15), let A be the event that an entering customer eventually abandons and let W be the time spent in queue by an entering customer. Let W_k be the time to abandon for a customer who starts in position k in queue. Then, reasoning as before,

$$\mathrm{P}(A) = \sum_{k=0}^{r-1} p^a_{s+k} (1 - \Gamma_{k+1}), \tag{7.25}$$

$$\mathrm{E}[W 1_A] = \sum_{k=0}^{r-1} p^a_{s+k} \mathrm{E}[W_{k+1} 1_A], \tag{7.26}$$

and

$$\mathrm{E}[W^2 1_A] = \sum_{k=0}^{r-1} p^a_{s+k} \mathrm{E}[W^2_{k+1} 1_A], \tag{7.27}$$

where

$$\begin{aligned} \mathrm{E}[W_k 1_A] = {} & \gamma_{k,1} m_{k,1} + (1-\gamma_{k,1}) \gamma_{k,2} (m_{k,1} + m_{k,2}) \\ & + (1-\gamma_{k,1})(1-\gamma_{k,2}) \gamma_{k,3} (m_{k,1} + m_{k,2} + m_{k,3}) \\ & + \cdots + (1-\gamma_{k,1}) \cdot \ldots \cdot (1-\gamma_{k,k-1}) \\ & \cdot \gamma_{k,k} (m_{k,1} + \cdots + m_{k,k}) \end{aligned} \tag{7.28}$$

and

$$\begin{aligned} \mathrm{E}[W_k^2 1_A] = {} & \gamma_{k,1} 2 m^2_{k,1} + (1-\gamma_{k,1}) \gamma_{k,2} \\ & \cdot (m^2_{k,1} + m^2_{k,2} + (m_{k,1} + m_{k,2})^2) \\ & + \cdots + (1-\gamma_{k,1})(1-\gamma_{k,2}) \cdot \ldots \cdot (1-\gamma_{k,k-1}) \gamma_{k,k} \\ & \cdot (m^2_{k,1} + \cdots + m^2_{k,k} + (m_{k,1} + \cdots + m_{k,k})^2). \end{aligned} \tag{7.29}$$

The associated conditional moments are

$$\mathrm{E}(W \mid A) = \frac{\mathrm{E}[W 1_A]}{\mathrm{P}(A)} \quad \text{and}$$
$$\mathrm{E}(W^2 \mid A) = \frac{\mathrm{E}[W^2 1_A]}{\mathrm{P}(A)} \tag{7.30}$$

for $\mathrm{P}(A)$ in (7.25). Finally, the conditional variance is

$$\mathrm{Var}(W \mid A) = \mathrm{E}(W^2 \mid A) - (\mathrm{E}(W \mid A))^2. \tag{7.31}$$

Now let $\hat{a}(z) \equiv \mathrm{E}[e^{-zW}1_A]$ be the Laplace transform of W for entering customers who abandon. Paralleling (7.22) and (7.28), we have

$$\hat{a}(z) = \sum_{k=0}^{r-1} p_{s+k}^{a} \hat{a}_{k+1}(z), \tag{7.32}$$

where

$$\hat{a}_k(z) = \gamma(k,1)\left(\frac{m_{k,1}^{-1}}{m_{k,1}^{-1}+z}\right) + \sum_{j=2}^{k} \gamma_{k,j}\left(\frac{m_{k,j}^{-1}}{m_{k,j}^{-1}+z}\right) \cdot \prod_{\ell=1}^{j-1}\left[(1-\gamma_{k,\ell})\left(\frac{m_{k,\ell}^{-1}}{m_{k,\ell}^{-1}+z}\right)\right]. \tag{7.33}$$

Paralleling $\mathrm{P}(0 < W \le t; S)$ above, we can compute $\mathrm{P}(W \le t; A)$ by numerically inverting its Laplace transform $\hat{a}(z)/z$. Then the conditional cdf of the time to abandon given that the customer does, in fact, abandon is

$$\mathrm{P}(W \le t \mid A) = \frac{\mathrm{P}(W \le t; A)}{\mathrm{P}(A)}. \tag{7.34}$$

We can easily combine the results in this section with the results in the last section to determine the waiting-time distribution of all customers, regardless of whether they abandon or are served:

$$\mathrm{P}(W \le t) = \mathrm{P}(W = 0) + \mathrm{P}(0 < W \le t; S) + \mathrm{P}(W \le t; A), \quad t > 0. \tag{7.35}$$

8. Fitting the Model Parameters

Given the $M/GI/s/r+GI$ model, it is natural to try to estimate the general service-time and abandon-time distributions directly, which is somewhat difficult because they involve censored data. We do not directly observe abandon times, because some customers are served before they would abandon (see Brown et al. 2005 for discussion).

We have shown how to derive the appropriate Markovian abandonment approximation from the abandon-time hazard function and the arrival rate λ, but an attractive alternative, which avoids directly estimating the abandon-time distribution or its hazard rate, is to directly fit a $M/M/s/r+M(n)$ model, or the more general $M(n)/M(n)/s/r+M(n)$ model, to available system data, be the data from a simulation or an actual operating call center.

We can directly estimate the total abandonment rate δ_k by the estimator $\hat{\delta}_k$, defined as the number of abandonments by customers from a queue of length k in the time interval $[0,t]$ divided by the length of time in the time interval $[0,t]$ that the queue was of length k. Since $\alpha_j = \delta_j - \delta_{j-1}$, we can also estimate α_j by the estimator $\hat{\alpha}_{2,j} = \hat{\delta}_j - \hat{\delta}_{j-1}$.

This alternative statistical approach is investigated in Pierson and Whitt (2005) and found to be effective. The simulation experiments show that the approximate abandonment rates produced by the method of §3 agree closely with the exact abandonment rates estimated from simulations when there is ample data.

9. Conclusions

The queueing model $M/GI/s/r+GI$ has long been regarded as appropriate for call centers, but it is difficult to analyze directly. We find that the steady-state behavior of the $M/GI/s/r+GI$ model is primarily affected by the service-time distribution through its mean. In contrast, the steady-state behavior of the $M/GI/s/r+GI$ model is primarily affected by the time-to-abandon distribution by its hazard function near the origin, and not its mean or tail behavior. That is perhaps the major insight about the $M/GI/s/r+GI$ model to be drawn from this work.

We have shown that the Markovian $M/M/s/r+M(n)$ model with state-dependent abandonment rates often can serve as an excellent approximation for the relatively intractable $M/GI/s/r+GI$ model. Moreover, in §§3 and 5 we have identified a simple way to construct the approximating $M/M/s/r+M(n)$ model, given the arrival rate and the abandon-time hazard function.

We can exploit birth-and-death processes to analyze the approximating $M/M/s/r+M(n)$ model, but it is not easy to describe the customer experience in this model. In §7, we introduced further approximations, making it possible to calculate approximate solutions for all the standard steady-state performance measures in the $M/M/s/r+M(n)$ model. The algorithm exploits numerical transform inversion in addition to the approximations.

We have performed computer simulations to evaluate the performance of the approximations. The examples we have examined, which are typical for call centers, indicate that the approximations are remarkably accurate. The weakest part of the approximation seems to be the treatment of nonexponential service-time distributions that are not close to exponential, as illustrated by the lognormal $LN(1,4)$ case in Table 6 and the deterministic (D) case in Table 7, but even in these cases the performance is not too bad. When the mean abandon time is large, the $M/GI/s/r+GI$ model will behave much like the associated $M/GI/s/r$ model, for which we already know much about the impact of the service-time distribution beyond its mean. Hence, some limitations of the approximation are known. However, we have not nearly explored all possible cases. For contemplated new scenarios, the approximation should be validated by comparing it with computer simulations.

As indicated in §8, once it is recognized that a state-dependent Markovian model might serve as a good approximation for the original $M/GI/s/r+GI$ model, it is natural to directly fit the Markovian $M/M/s/r+M(n)$ model to system data, which is investigated by Pierson and Whitt (2005). Moreover, it is natural to go beyond the first Markovian model with state-dependent abandonment rates to consider new Markovian models with state-dependent arrival rates and service rates as well. From a practical engineering perspective, our work suggests that the canonical model for (single-site, single-group) call centers should perhaps be the $M(n)/M(n)/s/r+M(n)$ model instead of the $M/GI/s/r+GI$ model. To some extent, that point of view already is expressed by Brandt and Brandt (1999, 2002).

The approximations for service times and abandon times proposed for the $M/GI/s/r+GI$ model in this paper can immediately be applied to more complicated models of the same kind, e.g., as occur with skill-based routing when there are multiple classes of calls and agents. It remains to determine how effective these approximations will be in other settings.

An online supplement (Whitt 2004b) to this paper is available at http://mansci.pubs.informs.org/ecompanion.html.

Acknowledgments

The author is grateful to Columbia University undergraduate Margaret Pierson for writing the $M/GI/s/r+GI$ simulation program and performing the simulation experiments. The author is also grateful to the referees for their critical reading of this paper. The author was supported by National Science Foundation Grant DMS-02-2340.

References

Abate, J., W. Whitt. 1995. Numerical inversion of Laplace transforms of probability distributions. *ORSA J. Comput.* **7** 36–43.

Abate, J., G. L. Choudhury, W. Whitt. 1999. An introduction to numerical transform inversion and its application to probability models. W. Grassman, ed. *Computational Probability*. Kluwer, Boston, MA, 257–323.

Ancker, C. J. Jr., A. V. Gafarian. 1963. Queuing with reneging and multiple heterogeneous servers. *Naval Res. Log. Quart.* **10** 125–145.

Baccelli, F., G. Hebuterne. 1981. On queues with impatient customers. F. J. Kylstra, ed. *Performance '81*. North-Holland, Amsterdam, The Netherlands, 159–179.

Bolotin, V. 1994. Telephone circuit holding-time distributions. J. Labetoulle, J. W. Roberts, eds. *Proc. Internat. Teletraffic Congress, ITC* 14. North-Holland, Amsterdam, The Netherlands, 125–134.

Brandt, A., M. Brandt. 1999. On the $M(n)/M(n)/s$ queue with impatient calls. *Performance Evaluation* **35** 1–18.

Brandt, A., M. Brandt. 2002. Asymptotic results and a Markovian approximation for the $M(n)/M(n)/s+GI$ system. *Queueing Systems* **41** 73–94.

Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, L. Zhao. 2005. Statistical analysis of a telephone call center: A queueing-science perspective. *J. Amer. Stat. Assoc.* Forthcoming.

Gans, N., G. Koole, A. Mandelbaum. 2003. Telephone call centers: Tutorial, review and research prospects. *Manufacturing Service Oper. Management* **5** 79–141.

Garnett, O., A. Mandelbaum, M. I. Reiman. 2002. Designing a call center with impatient customers. *Manufacturing Service Oper. Management* **4** 208–227.

Halfin, S., W. Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* **29** 567–588.

Jelenkovic, P., A. Mandelbaum, P. Momcilovic. 2004. Heavy traffic limits for queues with many deterministic servers. *Queueing Systems* **47** 53–69.

Mandelbaum, A., N. Shimkin. 2000. A model for rational abandonments from invisible queues. *Queueing Systems* **36** 141–173.

Mandelbaum, A., R. Schwartz. 2002. Simulation experiments with $M/G/100$ queues in the Halfin-Whitt (QED) regime. Technical report, The Technion, Israel.

Mandelbaum, A., S. Zeltyn. 2004. The impact of customers patience on delay and abandonment: Some empirically-driven experiments with the $M/M/N+G$ queue. *OR Spektrum* **26** 377–411.

Palm, C. 1937. Étude des délais d'attente. *Ericsson Technics* **5** 37–56.

Pierson, M. P., W. Whitt. 2005. A statistically fit Markovian approximation for the $M/GI/s/r+GI$ queueing model. Working paper, Department of Industrial Engineering and Operations Research, Columbia University, New York.

Puhalskii, A. A., M. I. Reiman. 2000. The multiclass GI/PH/N queue in the Halfin-Whitt regime. *Adv. Appl. Probab.* **32** 564–595.

Seelen, L. P., H. C. Tijms, M. H. van Hoorn. 1985. *Tables for Multi-Server Queues*. North-Holland, New York.

Whitt, W. 1993. Approximations for the $GI/G/m$ Queue. *Production Oper. Management* **2** 114–161.

Whitt, W. 1999. Improving service by informing customers about anticipated delays. *Management Sci.* **45** 192–207.

Whitt, W. 2002. *Stochastic-Process Limits*. Springer, New York.

Whitt, W. 2004a. A diffusion approximation for the $G/GI/n/m$ queue. *Oper. Res.* **52** 922–944.

Whitt, W. 2004b. Engineering solution of a basic call-center model: Supplementary material. http://mansci.pubs.informs.org/ecompanion.html.

Whitt, W. 2004c. Efficiency-driven heavy-traffic approximations for multi-server queues with abandonments. *Management Sci.* **50**(10) 1449–1461.

Whitt, W. 2005a. Heavy-traffic limits for the $G/H_2^*/n/m$ queue. *Math. Oper. Res.* **30**(1) 1–27.

Whitt, W. 2005b. Fluid models for multi-server queues with abandonments. *Oper. Res.* Forthcoming.

Whitt, W. 2005c. Two fluid approximations for multi-server queues with abandonments. *Oper. Res. Lett.* Forthcoming.

Wolff, R. W. 1989. *Stochastic Modelling and the Theory of Queues*. Prentice-Hall, Englewood Cliffs, NJ.

Zohar, E., A. Mandelbaum, N. Shimkin. 2002. Adaptive behavior of impatient customers in tele-queues: Theory and empirical support. *Management Sci.* **48** 566–583.