ENGINEERING SOLUTION OF A BASIC CALL-CENTER MODEL

by

Ward Whitt

Department of Industrial Engineering and Operations Research
Columbia University, New York, NY 10027

Abstract

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, IID service times with a general distribution, s servers, r extra waiting spaces and IID customer abandonment times with a general distribution. Empirical studies 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.

Keywords: 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.

December 9, 2003

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 email 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, Koole and Mandelbaum (2002) 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 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) and Brown et al. (2002).

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 multi-site, multi-class 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 and IID customer abandonment times with a general distribution (the final GI). 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 fifteen-sixty 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) and Garnett, Mandelbaum and Reiman (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 Bacceli and Hebuterne (1981), Brandt and Brandt (1999, 2002), Mandelbaum and Zeltyn (2003) and references therein, but 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, Mandelbaum and Shimkin (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 $s = 1000$) and there is non-negligible customer abandonment (e.g., $1 - 5\%$). We want to allow non-exponential service-time and abandon-time distributions, but realistic distributions are not radically different from exponential. For example, as observed in Brown et al. (2002), the service-time distribution might be lognormal with a squared coefficent of variation (CSQ, 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. 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 Section 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 Section 2 we start by presenting simulation results to show that it is not sufficient to just use the corresponding Erlang A model, obtained by using exponential service-time and abandon-time distributions with the given means. In Section 3 we introduce the state-dependent Markovian approximation for the abandonments. In Section 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 Section 5 we discuss the simple exponential approximation for the more general GI service times. In Section 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 Section 7 we specify a general Markovian call-center model that we will analyze; it is the $M(n)/M/s/r+M(n)$ model, which as state-dependent arrival rates as well as state-dependent abandonment rates to allow for treating balking and retrying. In Section 8 we derive the steady-state performance measures in the $M(n)/M/s/r + M(n)$ model, some of which require approximations. In Section 9 we discuss fitting the model parameters to call-center data. Finally, in Section 10 we draw conclusions.

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.

Indeed, simulations show that the Erlang A model does not provide a consistently good approximation for the $M/GI/s/r+GI$ model. For example, 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. An E_k distribution has $CSQ = 1/k$. Thus the E_2 abandon-time distribution here has mean 1, $CSQ = 1/2$ and variance 1/2.

In Table 1 we compare simulations of the $M/E_2/100/200 + E_2$ model with simulations of the $M/M/100/200 + M$ model with the same arrival rate, mean service time and mean time to abandon. All simulation experiments reported in the paper are based on ten 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.

To define the performance measures we examine, let S be the event that a typical customer that enters the system (is not blocked) eventually will be served; let A be the event that a typical customer that enters the system abandons before starting service; let W be the steady-state waiting time (before beginning service) for a typical entering customer; 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 $Var(Q)$, 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; $E[W|S]$ and $Var(W|S)$, the conditional mean and variance of the waiting time of an entering customer, given that the entering customer eventually will be served; $E[W|A]$ and $Var(W|A)$, the conditional mean and variance of the waiting time of an entering customer, given that the entering customer eventually will abandon; $P(W \leq t|S)$, the conditional probability that an entering customer waits less than time t, given that the customer eventually will be served; and $P(W \leq t|A)$, 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. We have chosen the waiting room size r sufficiently large so that blocking is negligible and so not a factor.

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.

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,

	model, with mean time to abandon = 1.0			
	$M/E_2/100/200+E_2$		$M/M/100/200+M$	
Performance Measure	*sim.*	*approx.*	*sim.*	exact
$P(W=0)$	0.217	0.250	0.4092	0.4083
	$\pm$0.0021	–	$\pm$0.0013	–
$P(A)$	0.0351	0.0381	0.0498	0.0499
	$\pm$0.00029	–	$\pm$0.00020	–
$E[Q]$	11.52	11.41	5.073	5.092
	$\pm$0.075	–	$\pm$0.024	–
$Var(Q)$	112.0	121.9	44.4	44.6
	$\pm$0.71	–	$\pm$0.30	–
$E[N]$	109.9	109.5	102.0	102.0
	$\pm$0.092	–	$\pm$0.036	–
$E[W\|S]$	0.1115	0.1102	0.0489	0.0490
	$\pm$0.00071	–	$\pm$0.00023	–
$Var(W\|S)$	0.0101	0.0119	0.00418	0.0042
	$\pm$0.000061	–	$\pm$0.000027	–
$E[W\|A]$	0.1508	0.1521	0.0665	0.0666
	$\pm$0.00042	–	$\pm$0.00021	–
$Var(W\|A)$	0.0067	0.0079	0.0031	0.0031
	$\pm$0.000044	–	$\pm$0.000018	–
$P(W\leq 0.1\|S)$	0.510	0.528	0.7994	0.7986
	$\pm$0.0030	–	$\pm$0.0012	–
$P(W\leq 0.1\|A)$	0.305	0.316	0.7678	0.7671
	$\pm$0.0014	–	$\pm$0.0013	–
$P(W\leq 0.2\|S)$	0.795	0.786	0.9648	0.9644
	$\pm$0.0023	–	$\pm$0.00057	–
$P(W\leq 0.2\|A)$	0.740	0.726	0.9705	0.9702
	$\pm$0.0019	–	$\pm$0.00054	–

Table 1: 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 = 1.0. The two models have common arrival rate $\lambda = 102$, mean service time $\mu^{-1} = 1$, number of servers $s = 100$, number of extra waiting spaces $r = 200$ and mean time to abandon 1.0. Both simulation estimates and numerical results are shown for the two models. The half-width of the 95% confidence interval is given for each simulation estimate.

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. (That corresponds to the classic 80/20 rule used in many call centers, meaning that 80% of all calls should be answered within 20 seconds, 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 3 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 j^{th} from the end of a queue of length k will abandon at rate $\alpha_{k,j}$,

$1 \le j \le k$, independent of the rest of the history up to that point. We will first develop a way to define suitable infinitesimal rates $\alpha_{k,j}$ 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, which states that 80% of all calls should be answered within 20 seconds. Since the mean length of the calls themselves tends to be 200 seconds 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 $1 - 4\%$ 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

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

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, Mandelbaum and Reiman (2002), Chapter 10 of Whitt (2002a) Whitt (2002c) and Jelenkovic, Mandelbaum and Momcilovic (2002). 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 hazard (or failure-rate) function

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

where $F^c(t) \equiv 1 - F(t)$ is the complementary cdf (ccdf). To understand abandonment behavior, the key quantity is the hazard-rate function h in (3.2) for relatively small time arguments.

In fact, in some circumstances it may even be possible to use only the single value $h(0) = f(0)$, the value of the hazard-rate function at the origin, since only small times should matter. Indeed, that was our initial idea. In the process of doing this research, we discovered that this idea also has been advanced by Mandelbaum and Zeltyn (2003). However, we found that the single value $h(0)$ did not work well. For example, distributions such as lognormal have $h(0) = 0$ and yet often produce many relatively small values.

Our goal is to produce 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 j^{th} 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 j^{th} from the end of the queue arrived.

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

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

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

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

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

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

Then the approximate total abandonment rate would be

$$\delta_k \equiv \lambda \int_{0}^{k/\lambda} h(t)\, dt = -\lambda \log_e F^c(k/\lambda) \ . \tag{3.6}$$

Our definition in (3.3) makes the rate $\alpha_{k,j}$ independent of k, so we have an unnecessary subscript k, but we believe that the extra subscript can be useful. When working with call-center data, an attractive alternative for generating the desired state-dependent abandonment rates is to directly estimate the rates $\alpha_{k,j}$, without making direct reference to the time-to-abandon cdf F or its hazard-rate function h. Clearly, we can estimate $\alpha_{k,j}$ by the ratio of the observed number of abandonments in state (k, j) divided by the total observed time spent in state (k, j).

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 $\alpha_{k,j}$ for the j^{th} customer from the end of a queue of length k, they develop an approximate abandonment rate β_j for the j^{th} 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 Section 8. 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 approximations for abandonments proposed in Section 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.

Given the $M/M/s/r+GI$ model, when we apply Section 3, we obtain the $M/M/s/r+M(n)$ model as an approximation. We then exploit further approximations to solve that model, as developed in Section 8.

In Table 2 we show results for the $M/M/100/200+E_2$ and $M/M/100/200+LN(1,1)$ models with common arrival rate $\lambda = 102$, mean service time $\mu^{-1} = 1$, $s = 100$ servers, $r = 200$ extra waiting spaces and mean abandon time 1. The Erlang E_2 abandon time has $CSQ = 1/2$ and thus variance $= 1/2$; the lognormal $LN(1,1)$ abandon time has $CSQ = 1$ and thus 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 both the $M/M/100/200+E_2$ and $M/M/100/200+LN(1,1)$ models. Moreover, the steady-state performance measures are quite different from the associated Erlang A model.

To show some other cases, we present two additional tables. In Table 3 we show results for a less variable lognormal abandon-time distribution with a greater mean, mean 4. Specifically, we let the lognormal distribution have $CSQ = 0.25$, and thus variance 4. Again the approximation agrees closely with the simulation results, and the performance is very different from that of the corresponding Erlang A model with the same mean service time and mean abandon time. Since the congestion is much greater in this case, we make the number of waiting spaces larger to avoid significant blocking; in particular, we let $r = 300$.

As illustrated by Tables 2 -3, 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 non-exponential service times.

It should be noted that many exact results can be computed for the $M/M/s/\infty+GI$ model, as shown by Brandt (1999), but the exact algorithm is complicated. When we go further to treat non-exponential service times commonly occurring in practice, it is not evident that greater accuracy for the $M/M/s/r + M(n)$ model provides an important benefit.

5. Treating the Service Times

In Section 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. Actually, the new model is somewhat more complicated, because we not only define a total state-dependent abandonment rate δ_k when there are k customers waiting in the queue, but we also define separate individual state-dependent abandonment rates $\alpha_{k,j}$ for each of the k customers in the queue, depending

	model, mean time to abandon = 1.0				
	$M/M/100/200 + E_2$		$M/M/100/200 + LN(1,1)$		$M/M/100/200 + M$
Perf. Meas.	*sim.*	*approx.*	*sim.*	*approx.*	exact
$P(W = 0)$	0.246 ±0.0020	0.250 –	0.242 ±0.0026	0.247 –	0.408 –
$P(A)$	0.0378 ±0.00032	0.0381 –	0.0376 ±0.00032	0.0379 –	0.0499 –
$E[Q]$	11.75 ±0.075	11.41 –	11.42 ±0.071	11.02 –	5.09 –
$Var(Q)$	129.2 ±0.94	121.9 –	115.6 ±0.46	107.2 –	44.6 –
$E[N]$	109.9 ±0.091	109.5 –	109.6 ±0.092	109.1 –	102.0 –
$E[W\|S]$	0.1133 ±0.00072	0.1102 –	0.1094 ±0.00067	0.1058 –	0.0490 –
$Var(W\|S)$	0.0119 ±0.000083	0.113 –	0.0104 ±0.000042	0.0097 –	0.0042 –
$E[W\|A]$	0.1628 ±0.00063	0.1521 –	0.1788 ±0.00026	0.1642 –	0.0666 –
$Var(W\|A)$	0.0079 ±0.000061	0.0076 –	0.0054 ±0.000024	0.0054 –	0.0031 –
$P(W \leq 0.1\|S)$	0.520 ±0.0026	0.528 –	0.518 ±0.0028	0.527 –	0.799 –
$P(W \leq 0.1\|A)$	0.273 ±0.0019	0.316 –	0.140 ±0.00064	0.204 –	0.767 –
$P(W \leq 0.2\|S)$	0.775 ±0.0023	0.786 –	0.792 ±0.0018	0.807 –	0.964 –
$P(W \leq 0.2\|A)$	0.688 ±0.0027	0.726 –	0.644 ±0.00066	0.706 –	0.970 –

Table 2: A comparison of approximations for steady-state performance measures with simulations in two models with exponential service times and mean abandon time 1. The two models have Erlang E_2 and lognormal $LN(1,1)$ abandon-time distributions. The lognormal abandon-time distribution has $CSQ = 1.0$ and thus variance = 1.0. The two models have common arrival rate $\lambda = 102$, mean service time $\mu^{-1} = 1$, number of servers $s = 100$, number of extra waiting spaces $r = 200$ and mean time to abandon 1.0. The half-width of the 95% confidence interval is given for each simulation estimate. The exact numerical results are also displayed for the corresponding model with an exponential abandon-time distribution.

	model, mean time to abandon = 4.0		
	$M/M/100/300 + LN(4, 0.25)$		$M/M/100/300 + M$
Performance Measure	*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	–	–
$E[W\|S]$	1.154	1.144	0.1455
	± 0.0073	–	–
$E[W\|A]$	1.327	1.288	0.1429
	± 0.0015	–	–
$P(W \leq 0.4\|S)$	0.0702	0.0710	0.469
	± 0.0032	–	–
$P(W \leq 0.4\|A)$	0.000093	0.0000	0.449
	± 0.0032	–	–

Table 3: A comparison of steady-state performance measures in the $M/M/100/300 + LN(4, 0.25)$ and $M/M/100/300 + M$ model with time-to-abandon distribution having mean 4.0. The lognormal time-to-abandon distribution has squared coefficient of variation 0.25 and thus variance 4.0. The two models have common arrival rate $\lambda = 102$, mean service time $\mu^{-1} = 1$, number of servers $s = 100$, number of extra waiting spaces $r = 300$ and mean time to abandon 4.0. Both simulation estimates and numerical results are shown for the lognormal model. The half-width of the 95% confidence interval is given for each simulation estimate.

on their position in the queue. 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 now go further and develop a Markovian approximation for the service times. 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 model in Section 8.

We primarily make this second model approximation because it produces a Markovian model that we can analyze. However, unlike the direct approximation by the Erlang A model, it also turns out to be surprisingly accurate. 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 known that the insensitivity 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 $M/GI/s/\infty$) or the associated intermediate finite-waiting room models $M/GI/s/r$. Under heavier loads, the insensitivity we are using as an approximation becomes much more reasonable because of the abandonments. 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 (2002b), based on the heavy-traffic limit for the $G/H_2^*/s/r$ model

established in Whitt (2002c). 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 $CSQ = 1$, then the entire diffusion approximation is independent of the service-time distribution beyond its mean. Consistent with that theoretically-based approximation, our approximations tend to perform better when the service-time CSQ is close to 1.

In summary, 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-rate 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.

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 Section 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

	model, mean time to abandon = 4.0		
	$M/E_2/100/200+E_2$		$M/M/100/200+M$
Performance Measure	*sim.*	*approx. numerical*	exact numerical
$P(W=0)$	0.056	0.0764	0.226
	$\pm$0.0016	–	–
$P(A)$	0.0236	0.0253	0.0364
	$\pm$0.00036	–	–
$E[Q]$	41.6	41.8	14.84
	$\pm$0.44	–	–
$E[N]$	141.2	141.2	113.1
	$\pm$0.39	–	–
$E[W\|S]$	0.407	0.409	0.1455
	$\pm$0.0042	–	–
$E[W\|A]$	0.413	0.430	0.1429
	$\pm$0.0023	–	–
$P(W \le 0.1\|S)$	0.133	0.161	0.4688
	$\pm$0.0032	–	–
$P(W \le 0.1\|A)$	0.0462	0.050	0.4493
	$\pm$0.00078	–	–
$P(W \le 0.2\|S)$	0.234	0.261	0.6865
	$\pm$0.0047	–	–
$P(W \le 0.2\|A)$	0.166	0.164	0.7366
	$\pm$0.0025	–	–

Table 4: 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. The two models have common arrival rate $\lambda = 102$, mean service time $\mu^{-1} = 1$, number of servers $s = 100$, number of extra waiting spaces $r = 200$ and mean time to abandon 4. Both simulation estimates and numerical results are shown for the Erlang model. The half-width of the 95% confidence interval is given for each simulation estimate.

abandon times, such as = 0.25, the results are quite close to the Erlang A model, but they are very different for larger mean abandon times. To illustrate, we show the case of mean abandon time 4.0 in Talbe 3.

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 5 we consider $M/GI/100/200+LN(1,1)$ models with common lognormal abandon-time distribution having mean = 1.0 and $CSQ = 1.0$; and in Table 6 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 non-exponential 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 (2002b), the performance degrades as the service-time CSQ deviates more from 1, the CSQ of an exponential distribution. In particular, we see degradation of performance for the $LN(1,4)$ service time in Table 5 and the D service time in Table 6, but even in these cases the errors are not too great.

7. A General Markovian Call-Center Model

In this section we define a general Markovian call-center model that includes the state-dependent abandonment rates developed in Section 3. Moreover, it includes the $M/M/s/r + M(n)$ model proposed as the approximation for the $M/GI/s/r + GI$ model. It goes beyond that to treat balking and retrying as well. In the next section we show how to calculate all desired steady-state performance measures for the model, exploiting a further approximation to treat the state-dependent abandonment rates.

The general model primarily is the $M(n)/M/s/r + M(n)$ queueing model, which has s servers, r extra waiting spaces and state-dependent arrivals (the first $M(n)$) and state-dependent abandonment (the final $M(n)$). We go beyond state-dependent abandonments to also treat balking and retrials, but all within the framework of the $M(n)/M/s/r+M(n)$ model. We could have state-dependent service as well ($M(n)$ instead of M), which could occur because of additional abandonment after service has begun, but we do not discuss that extension here.

We also go beyond the $M(n)/M/s/r+M(n)$ model: As indicated in Section 3, we not only define a total state-dependent abandonment rate when there are k customers waiting in the queue, but we also define a state-dependent abandonment rate for each of the k customers in the queue, depending on their position in the queue.

We assume that the queue operates in a standard manner: Arriving customers enter service immediately upon arrival if there is a free server. Arrivals finding all s servers busy and all r available waiting spaces full are blocked and lost. From an applied point of view, we think of blocked customers and abandoning customers as possibly retrying again later, but we model the retrials as totally independent events. We assume that retrials by blocked customers and abandoning customers are captured in the model indirectly by increased birth rates beyond the exogenous arrival rate. We work within the framework of the one-dimensional $M(n)/M/s/r+M(n)$ model. It thus remains to determine appropriate retrial rates when that

$M/GI/100/200 + LN(1,1)$ *model with mean time to abandon = 1.0*

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

Table 5: 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: E_2 with $CSQ = 0.5$, M with $CSQ = 1.0$, $LN(1,1)$ with $CSQ = 1.0$ and $LN(1,4)$ with $CSQ = 4.0$. The models have common arrival rate $\lambda = 102$, mean service time $\mu^{-1} = 1.0$, number of servers $s = 100$, number of extra waiting spaces $r = 200$ and $LN(1,1)$ abandon-time distribution with mean 1.0. The half-width of the 95% confidence interval is given for each simulation estimate. The common approximations based on the $M/M/100/200 + M(n)$ model are also displayed.

$M/GI/100/200 + E_2$ *model with mean time to abandon = 1.0*

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

Table 6: 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: deterministic (D) with $CSQ = 0$, E_2 with $CSQ = 0.5$, M with $CSQ = 1$ and $LN(1,1)$ with $CSQ = 1$. The models have common arrival rate $\lambda = 102$, mean service time $\mu^{-1} = 1$, number of servers $s = 100$, number of extra waiting spaces $r = 200$ and E_2 abandon-time distribution with mean 1.0. The half-width of the 95% confidence interval is given for each simulation estimate. The common approximations based on the $M/M/100/200 + M(n)$ model are also displayed.

feature is present. See Section 9 for further discussion.

If an arriving customer cannot enter service immediately and is not blocked, the customer may balk (leave immediately) because the customer does not want to wait at all before starting service. If an arriving customer cannot begin service immediately and does not leave immediately (because of blocking or balking), the customer joins the end of the queue. Waiting customers are served in a first-come, first-served (FCFS) manner. However, customers waiting in queue may elect to abandon. We assume that customers do not abandon after they have started service, but the model could easily be modified to allow for abandonment while in service. We assume that customers who abandon, balk or are blocked all may retry, but that these retrials are reflected by the state-dependent arrival rate.

Let $N(t)$ be the number of customers in the system at time t. In the $M(n)/M/s/r+M(n)$ queueing model, the stochastic process $\{N(t) : t \geq 0\}$ is a birth-and-death process. The birth rate λ_k is the state-dependent arrival rate; it is state-dependent to account for balking and retrials. 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.

We now specify the birth rates and death rates in detail. Key primitives are the exogenous arrival rate λ and the individual mean service time $1/\mu$. The birth and death rates are modified to account for blocking, balking, reneging (abandonment) and retrying. We will discuss how to select the model parameters in Section 9.

First, the birth rate in state k (the state-dependent arrival rate) is

$$\lambda_k = \begin{cases} \lambda + \zeta_k \ , & 0 \leq k \leq s-1 \ , \\ \\ \lambda(1-\beta) + \zeta_k \ , & s \leq k \leq s+r-1 \ , \end{cases} \tag{7.1}$$

where λ is the exogenous arrival rate, β is the probability that an exogenously arriving customer who finds all servers busy will balk, and ζ_k is the retrial arrival rate when there are k customers in the system, $0 \leq k \leq s+r-1$. We also introduce ζ_r to describe the retrial rate when the system is full. Thus, even though none of these arrivals enter the system, we assume that there is a total arrival rate of $\lambda_{s+r} = \lambda + \zeta_r$ when the system is full, all of whom are immediately blocked.

Second, the death rate in state k is

$$\mu_k = \begin{cases} k\mu \ , & 1 \leq k \leq s \ , \\ \\ s\mu + \delta_{k-s} \ , & s+1 \leq k \leq s+r \ , \end{cases} \tag{7.2}$$

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.

As indicated above, we go beyond this basic birth-and-death process model to define abandonment rates for each of the separate customers in queue. We let α_k be the abandonment rate for the k^{th} customer from the end of the queue, whenever there are at least k customers in queue. Thus, α_1 is the abandonment rate for the last customer in queue, while α_2 is the abandonment rate for the second-to-last customer in queue, and so forth. Thus, the total abandonment rate when there are k customers waiting in queue is

$$\delta_k = \sum_{j=1}^{j=k} \alpha_j \ . \tag{7.3}$$

Equivalently, we can start with the total abandonment rate, letting the total abandonment rate be δ_k when there are k customers waiting in queue. Then, assuming that the total abandonment rate δ_k is nondecreasing in k, we let

$$\alpha_k = \delta_k - \delta_{k-1}, \quad 1 \le k \le r \ , \tag{7.4}$$

where $\delta_0 \equiv 0$.

Above we have assumed that the abandonment rate for the k^{th} customer from the end of the queue is α_k, independent of the total number of customers waiting in queue (provided that there are at least k waiting). If it is deemed appropriate, it is easy to generalize that by introducing parameters $\alpha_{k,j}$, $1 \le j \le k$, letting $\alpha_{k,j}$ be the abandonment rate for the j^{th} customer from the end of the queue when the total number of customers is k. Our assumption above is the special case in which $\alpha_{k,j} = \alpha_j$ for all $k \ge j \ge 1$. We believe that special case is natural, but it can be generalized.

8. Steady-State Distribution of the Markovian Model

We now show how to calculate all the standard performance measures for the Markovian call-center model introduced in Section 7. We start by calculating the steady-state distribution of the basic birth-and-death process. Then we describe the experience of entering customers. 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, Choudhury and Whitt (1999) for an overview of the inversion algorithms.

8.1. Steady-State Distribution of the Birth-And-Death Process

The key to successfully analyzing the model is the fact that the stochastic process $\{N(t) : t \geq 0\}$ is a birth-and-death process. Since the state space is finite, there is a unique limiting steady-state distribution. Let N be a random variable with the limiting steady-state distribution of $N(t)$. The steady-state distribution is

$$p_k \equiv P(N = k) \equiv \lim_{t\to\infty} P(N(t) = k|N(0) = i) \ . \tag{8.1}$$

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

$$p_k\lambda_k = p_{k+1}\mu_{k+1}, \quad 0 \leq k \leq s + r - 1 \ . \tag{8.2}$$

Clearly, the birth-and-death parameters λ_k and μ_k can be totally general for this part, but we use the additional structure of the $M(n)/M/s/r + M(n)$ model later.

To calculate the steady-state distribution numerically, it is convenient to solve for it recursively. We use the structure of the $M(n)/M/s/r + M(n)$ model somewhat here: Since the probability p_s is likely to be of order $O(1)$ (assuming that the number s of servers has been chosen in a reasonable manner), it is natural to start at s and separately go up and down. For that purpose, let $x_s = 1$,

$$x_{s+k+1} = \frac{\lambda_{s+k}x_{s+k}}{\mu_{s+k+1}} \ , \quad 0 \leq k \leq r - 1 \ , \tag{8.3}$$

and

$$x_{k-1} = \frac{\mu_k x_k}{\lambda_{k-1}} \ , \quad 1 \leq k \leq s \ . \tag{8.4}$$

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

$$y = \sum_{k=0}^{s+r} x_k \ . \tag{8.5}$$

Then the steady-state probabilities are

$$p_k = x_k/y \ , \quad 0 \leq k \leq s + r \ . \tag{8.6}$$

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.

8.2. The Probability of Abandoning or Being Served

We now start to describe the experience of individual customers that join the system (including retrials, if any). Our approach is to condition on the state seen by arrivals that enter the system and then average over all the possibilities. First, we need the steady-state distribution of the number of customers in the system seem by arrivals. Let N^a be a random variable with the steady state distribution of the number of customers in the system seen by an arrival that enters the system, and let $p_k^a \equiv P(N^a = k)$; it is the limiting probability that the number of customers in the system is k just before an entering arrival. Then

$$p_k^a = \frac{\lambda_k p_k}{\sum_{j=0}^{j=s+r-1} \lambda_j p_j} \ ; \tag{8.7}$$

see Section 3.2 of Cooper (1981).

Note that the state probabilities seen by arrivals are the same as the state probabilities at an arbitrary time if there is no balking or retrials, because then the arrival process is a simple Poisson process. More generally, when 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 (PASTA) property; see Section 5.16 of Wolff (1989). However, with balking and retrials, the probabilities $\{p_k^a\}$ are in general different from the probabilities $\{p_k\}$.

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 that enters the system. First, the probability that an arriving customer who enters the system does not wait at all before starting service is exactly

$$P(NoWait) \equiv P(W = 0) = \sum_{k=0}^{s-1} p_k^a \ . \tag{8.8}$$

The situation is more complicated when the arrival must join the queue. To analyze this, we will make an approximation, which is exact only when $\delta_k = k\alpha$ or, equivalently, when $\alpha_k = \alpha$ for all k. Conditional on the arrival seeing $s + k - 1$ customers in the system upon arrival (the s customers in service and $k - 1$ others already in the queue waiting), $k \geq 1$, 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. For that, we exploit properties of independent exponential random variables.

Suppose that we consider the fate of the last customer in queue beginning at his arrival epoch. Suppose that there are $s+k$ customers in the system at this time. When there are $s+k$ customers in the system, we can think of there being $s+k$ independent exponential times to candidate next events. There are s independent exponential service times, each with rate μ; there are k independent exponential abandon times with rates α_j, $1 \leq j \leq k$. The time until the next event is the minimum of these $s+k$ exponential random variables, which has a rate equal to the sum of the rates. The probability that exponential variable j yields the minimum is the rate for that j^{th} variable divided by the sum of the rates. Moreover, the minimum and the variable yielding that minimum are independent random variables.

Thus, assuming that initially there are k customers in queue, with the new arrival being the last one, the probability that abandonment by the customer j places from the end of the queue is the first departure event is

$$\gamma_{k,1} = \frac{\alpha_j}{s\mu + \delta_k} \tag{8.9}$$

and the expected time until that first departure event is

$$m_{k,1} = \frac{1}{s\mu + \delta_k} \ . \tag{8.10}$$

The difficulty is going on to subsequent departure events. First, we need to account for the fact that, after the initial arrival time, the customers remaining in the queue are staying longer, so that the hazard rate should apply to a larger time argument. After two departures, a customer who was j^{th} from the end of the queue, if he is still present, should have an abandonment rate that changes from $\alpha_j = h(j/\lambda)$ to $\alpha_{j+2} = h((j+2)/\lambda)$.

Moreover, there is a difficulty in proceeding forward to subsequent departure events. We can proceed forward exactly, obtaining a new collection of independent exponential random variables for each subsequent departure event, but to do so properly, we need to keep track of which customers have generated previous departure events, because the relevant rates depend on which customers remain in the system.

To simplify matters, we make an approximation: We assume that the remaining total rate before the j^{th} departure event is

$$RemainingRate(j) \approx (\delta_k - \delta_{j-1}) \ , \tag{8.11}$$

where $\delta_0 = 0$.

Assuming that there are initially k waiting customers, let $\gamma_{k,j}$ be the probability that the customer initially k^{th} in line abandons in the j^{th} subsequent departure event, given that the

customer has not abandoned previously. Let $m_{k,j}$ be the mean time between the $(j-1)^{\text{st}}$ and j^{th} departure events (where the 0^{th} departure event occurs at time 0). By the reasoning above, we let

$$\gamma_{k,j} \approx \frac{\alpha_j}{s\mu + (\delta_k - \delta_{j-1})} \tag{8.12}$$

and

$$m_{k,j} \approx \frac{1}{s\mu + (\delta_k - \delta_{j-1})} \ , \tag{8.13}$$

for $1 \leq j \leq k$, where $\delta_0 \equiv 0$. Then the probability that customer $s+k$ eventually receives service is

$$\Gamma_k = (1-\gamma_{k,1})(1-\gamma_{k,2})\dots(1-\gamma_{k,k}) \tag{8.14}$$

for $\gamma_{k,j}$ in (8.12).

Note that in the special case $\alpha_k = \alpha$ for all k, the definitions above are exact, reducing to

$$\gamma_{k,j} = \frac{\alpha}{s\mu + (k-j+1)\alpha} \ \text{ and } \ m_{k,j} = \frac{1}{s\mu + (k-j+1)\alpha} \ , \tag{8.15}$$

where $\delta_0 \equiv 0$. Thus, our approximate algorithm produces the exact performance measures for the $M/M/s/r+M$ model, as we have confirmed with simulation.

Proceeding forward, we emphasize that we are now using the approximations in (8.12) and (8.13) as well as the approximate $M(n)/M/s/r+M(n)$ model. Hence subsequent performance measures are not exact for the $M(n)/M/s/r+M(n)$ model. (However, they are exact for the $M(n)/M/s/r+M$ special case.)

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

$$P(S) = \left(\sum_{k=0}^{s-1} p_k^a \right) + \sum_{k=0}^{r-1} p_{s+k}^a \Gamma_{k+1} \ , \tag{8.16}$$

for Γ defined in (8.14), drawing on the approximations in (8.12) and (8.13).

Since 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

$$P(A) = 1 - P(S) \ . \tag{8.17}$$

8.3. The Waiting Time for Customers Who Are Served

Let W be the waiting time (until beginning service) for a customer that enters the system. However, we must be careful, because we must differentiate between customers that eventually are served and customers that eventually abandon. In this subsection we consider only customers that are served.

We now compute the expectation of W for served customers; i.e., we compute $E[W; Served] = E[W1_{\{Served\}}]$, 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). Using properties of the exponential distribution, we obtain

$$E[W; S] = \sum_{k=0}^{r-1} p^a_{s+k} \Gamma_{k+1} \sum_{j=1}^{k+1} m_{k+1,j} \tag{8.18}$$

and

$$E[W^2; S] = \sum_{k=0}^{r-1} p^a_{s+k} \Gamma_{k+1} (V_{k+1} + M^2_{k+1}) \tag{8.19}$$

where

$$V_{k+1} \equiv \sum_{j=1}^{k+1} m^2_{k+1,j} \tag{8.20}$$

and

$$M_{k+1} \equiv \sum_{j=1}^{k+1} m_{k+1,j} \ . \tag{8.21}$$

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

$$E(W|S) = \frac{E[W; S]}{P(S)} \ \text{ and } \ E(W^2|S) = \frac{E[W^2; S]}{P(S)} \ . \tag{8.22}$$

The conditional variance and standard deviation are then

$$Var(W|S) \equiv E(W^2|S) - (E(W|S))^2 \tag{8.23}$$

and

$$SD(W|S) \equiv \sqrt{Var(W|S)} \ . \tag{8.24}$$

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 $\hat{w}_s(z) \equiv E[e^{-z(W)} 1_{\{S\}}]$ be the Laplace transform of W for served customers (Laplace-Stieltjes Transform of its cdf). Paralleling (8.18), we have

$$\hat{w}_s(z) = \sum_{k=0}^{r-1} p^a_{s+k} \Gamma_{k+1} \hat{e}_{k+1}(z) \ , \tag{8.25}$$

where

$$\hat{e}_{k+1}(z) \equiv \Pi_{j=1}^{k+1} \left(\frac{m^{-1}_{k+1,j}}{m^{-1}_{k+1,j} + z} \right) \ . \tag{8.26}$$

We can now calculate the cdf by numerical transform inversion. Specifically, we obtain the cdf $P(W \le t; S)$ for any desired t by numerically inverting its Laplace transform $\hat{w}(z)/z$,

e.g., by using the Fourier-series method described in Abate and Whitt (1995). The associated conditional waiting-time cdf is

$$P(W \le t|S) = \frac{P(W \le t; S)}{P(S)} \ . \tag{8.27}$$

8.4. The Time to Abandon

As in (8.17), 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,

$$P(A) = \sum_{k=0}^{r-1} p^a_{s+k}(1 - \Gamma_{k+1}) \ , \tag{8.28}$$

$$E[W1_{\{A\}}] = \sum_{k=0}^{r-1} p^a_{s+k} E[W_{k+1}1_{\{A\}}] \tag{8.29}$$

and

$$E[W^2 1_{\{A\}}] = \sum_{k=0}^{r-1} p^a_{s+k} E[W^2_{k+1}1_{\{A\}}] \ , \tag{8.30}$$

where

$$\begin{aligned} E[W_k 1_{\{A\}}] = \ & \gamma_{k,1}m_{k,1} + (1-\gamma_{k,1})\gamma_{k,2}(m_{k,1}+m_{k,2}) \\ & +(1-\gamma_{k,1})(1-\gamma_{k,2})\gamma_{k,3}(m_{k,1}+m_{k,2}+m_{k,3}) \\ & +\ldots+(1-\gamma_{k,1})\ldots(1-\gamma_{k,k-1})\gamma_{k,k}(m_{k,1}+\ldots+m_{k,k}) \end{aligned} \tag{8.31}$$

and

$$\begin{aligned} E[W^2_k 1_{\{A\}}] = \ & \gamma_{k,1}2m^2_{k,1} + (1-\gamma_{k,1})\gamma_{k,2}(m^2_{k,1}+m^2_{k,2}+(m_{k,1}+m_{k,2})^2)+\ldots+ \\ & (1-\gamma_{k,1})(1-\gamma_{k,2})\ldots(1-\gamma_{k,k-1})\gamma_{k,k}(m^2_{k,1}\ldots+m^2_{k,k}+(m_{k,1}+\ldots+m_{k,k})^2) \end{aligned} \tag{8.32}$$

The associated conditional moments are

$$E(W|A) = \frac{E[W1_{\{A\}}]}{P(A)} \ \text{ and } \ E(W^2|A) = \frac{E[W^2 1_{\{A\}}]}{P(A)} \ , \tag{8.33}$$

for $P(A)$ in (8.28). Finally, the conditional variance and standard deviation are

$$Var(W|A) = E(W^2|A) - (E(W|A))^2 \tag{8.34}$$

and

$$SD(W|A) = \sqrt{Var(W|A)} \ . \tag{8.35}$$

Now let $\hat{a}(z) \equiv E[e^{-zW}1_{\{A\}}]$ be the Laplace transform of W for entering customers who abandon. Paralleling (8.25), we have

$$\hat{a}(z) = \sum_{k=0}^{r-1} p_{s+k}^{a} \hat{a}_{k+1}(z) \ , \tag{8.36}$$

where

$$\begin{aligned} \hat{a}_k(z) &= \gamma(k,1)\left(\frac{m_{k,1}^{-1}}{m_{k,1}^{-1}+z}\right) \\ &\quad + \sum_{j=2}^{k} \gamma_{k,j}\left(\frac{m_{k,j}^{-1}}{m_{k,j}^{-1}+z}\right)\Pi_{\ell=1}^{j-1}\left[(1-\gamma_{k,\ell})\left(\frac{m_{k,\ell}^{-1}}{m_{k,\ell}^{-1}+z}\right)\right] \ . \end{aligned} \tag{8.37}$$

Paralleling $P(W \le t; S)$ above, we can compute $P(W \le t; A)$ 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

$$P(W \le t|A) = \frac{P(W \le t; A)}{P(A)} \ . \tag{8.38}$$

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 whether they abandon or are served.

9. 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. (2002) 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.

Since the abandonment rate $\alpha_{k,j}$ applies to the customer j^{th} from the end of a queue of length k, it is natural to use the estimator $\hat{\alpha}_{k,j}$, defined as the number of abandonments by customers j^{th} from the end of 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. The goal is to obtain

such estimates for a relatively long time interval over which the operating conditions do not change significantly. Of course, that condition is easily achieved in a computer simulation.

Similarly, we can indirectly account for retrials and balking by directly estimating state-dependent arrival rates. Paralleling the estimator $\hat{\alpha}_{k,j}$ defined above, we would estimate the state-dependent birth rate λ_k by $\hat{\lambda}_k$, defined as the number of arrivals during a time interval $[0,t]$ occurring when the number in system is k divided by the length of time that the number of customers in the system is k. Furthermore, we can indirectly account for abandonment after service has begun by directly estimating state-dependent service rates in all states. If the estimators reveal significant state-dependence, we can deduce that effects such as abandonments, retrials and balking must be happening.

We intend to test these statistically-fit $M(n)/M(n)/s/r + M(n)$ approximations in future work.

10. 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 have shown that the Markovian model $M/M/s/r + M(n)$ with state-dependent abandonment rates often can serve as an excellent approximation for the relatively intractable $M/GI/s/r + GI$ model. Moreover, in Sections 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, and in Section 8 we have developed approximate solutions for all the standard steady-state performance measures in the model. The algorithm exploits numerical transform inversion plus an approximation to describe customer experience with the state-dependent abandonment rates. The key analysis approximation appears in (8.11)–(8.13).

As indicated in Section 9, 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. 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. 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.

11. 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 was supported by National Science Foundation Grant DMS-02-2340.

References

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

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

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

Baccelli, F., G. Hebuterne. 1981. On queues with impatient customers. in *Performance '81*, ed. E. Gelenbe, North-Holland, Amsterdam, pp. 159–179.

Bolotin, V. 1994. Telephone circuit holding-time distributions. In *Proceedings of the International Teletraffic Congress, ITC* 14, J. Labetoulle and J. W. Roberts (eds.), North-Holland, Amsterdam, 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. 2002. Statistical analysis of a telephone call center: a queueing-science perspective. Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA.

Cooper, R. B. 1981. *Introduction to Queueing Theory*, second edition, North Holland, Amsterdam.

Gans, N., G. Koole, A. Mandelbaum. 2003. Telephone call centers: Tutorial, Review and Research Prospects. *Manufacturing and Service Opns. Mgmt.* 5, to appear.

Garnett, O., A. Mandelbaum, M. I. Reiman. 2002. Designing a call center with impatient customers. *Manufacturing and Service Opns. Mgmt.*, 4, 208-227.

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

Jelenkovic P., A. Mandelbaum, P. Momcilovic. 2002. Heavy Traffic Limits for Queues with Many Deterministic Servers. Columbia University.

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

Mandelbaum, A., S. Zeltyn. 2003. The impact of customers patience on delay and abandonment: some empirically-driven experiments with the $M/M/n+G$ queue. The Technion, Israel.

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

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

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

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

Whitt, W. 2002b. A diffusion approximation for the $G/GI/n/m$ queue. Department of Industrial Engineering and Operations Research, Columbia University. *Operations Res.*, to appear.

Whitt, W. 2002c. Heavy-traffic limits for the $G/H_2^*/n/m$ queue. Department of Industrial Engineering and Operations Research, Columbia University.

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 Science* 48, 566–583.