ENGINEERING SOLUTION OF A BASIC CALL-CENTER MODEL: SUPPLEMENTARY MATERIAL

by

Ward Whitt

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

December 15, 2003; Revision: August 6, 2004

Abstract

This paper contains additional material supplementing the main paper to be published in *Management Science*.

Abstract from main paper

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. Simulation experiments show that the approximation is quite accurate. The $M/M/r/s+M(n)$ model is tractable because the number of customers in the system over time is a birth-and-death process. After making an additional approximation, steady-state waiting-time and response-time distributions are characterized via their Laplace transforms. Then the approximate distributions are computed by numerically inverting the transforms. 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.

1. Introduction

In the main paper, Whitt (2005a), we develop an algorithm to rapidly compute approximations for all the standard steady-state performance measures in 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 (FCFS) service discipline. We especially want the approximations to be effective 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., $1 - 10\%$). We want to allow non-exponential service-time and abandon-time distributions. For example, as observed in Brown et al. (2002), the service-time distribution might be lognormal with a squared coefficent 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. 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.

Here is how the rest of this supplement is organized: In Section 2 we show additional insight can be gained by plotting the approximate abandonment rates; this is tantamount to simply

plotting the hazard-rate function itself over the interval of relevant arguments.

The next six sections present additional experimental results showing how the approximation performs, by making comparisons to computer simulations. In Section 3 we investigate how the approximation performs for a smaller number of servers. We first note that the classical $M/GI/1/\infty$ model clearly shows limitations of the approximation in general. In particular, it shows that the approximation can perform poorly when the number of servers is small and there is negligible abandonment. However, we also show that the approximation can perform reasonably well with much fewer servers. In the main paper we only considered the case of $s = 100$ servers; here we consider $s = 20$ and $s = 5$.

In Sections 4 and 5 we evaluate the approximations when the load is very light and very heavy, respectively. In Section 4 we present some results for the case $s = 100$ and $\lambda = 90$, which we regard as a light load. Under lighter loads, the approximation tends to perform better, but it does not produce small relative errors when the congestion is low.

In Section 5 we consider heavy loads. We present several results for the case $s = 100$ and $\lambda = 120$. To show what happens with fewer servers, we also consider the case $s = 20$ and $\lambda = 24$. Both these cases have the relatively high traffic intensity $\rho = 1.2$. We consider two quite different distributions: Erlang (E_2) and lognormal ($LN(1, 4)$), which have SCV $1/2$ and 4, respectively. We consider each combination of these two distributions as service-time distribution and time-to-abandon distribution. We show that the approximation performs remarkably well in those cases. However, because of the approximation assumption that abandonments are relatively rare compared to arrivals, we would like to see if the approximation will eventually break down if we make the loads heavy enough. To investigate the possible performance degradation, we consider the case of $s = 100$ and $\lambda = 200$, which yields about 50% abandonment. however, even in that case, the approximation performs well.

In Section 6 we show in more detail that the approximation is effective for making staffing decisions. We also show that the approximation does a much better job than the simple Erlang A model, obtained by choosing an exponential abandon-time distribution with the same mean as the given mean.

In Section 7 we investigate how the approximation performs for two uniform distributions, one on the interval $(0, 2)$ and the other on the interval $(0.5, 1.5)$. We consider each of these two uniform distributions in the role of the service-time distribution and the abandon-time distribution, considering all four combinations. The experiment strongly supports the main conclusion that the hazard rate for smaller arguments matters most for the abandon-time

distribution, while only the mean matters for the service-time distribution.

In Section 8 we present a few other examples that were deleted from the main paper in order to meet space limitations. Finally, in Section 9 we present some additional details on calculating performance measures in the approximating $M/M/s/r + M(n)$ model.

We conclude this introduction by mentioning that important exact results for the $M/M/s/r+GI$ model appear in Brandt and Brandt (2002) and Mandelbaum and Zeltyn (2004). Those papers, and earlier references cited there, should be consulted for additional insights.

2. Plotting Approximate Abandonment Rates

As discussed in Section 3 of the main paper, our main idea is to develop appropriate approximate state-dependent Markovian abandonment rates. Those approximate rates provide the $M(n)$ in the approximating model $M/M/s/r + M(n)$. The approximation is remarkably simple: We let the abandonment rate for the j^{th} customer from the end of the queue, when the queue length is k, for $k \geq j$, be

$$\alpha_{k,j} \equiv \alpha_j \equiv h(j/\lambda),$$

where h is the hazard-rate (or failure-rate) function, i.e.,

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

with $f(t)$ being the pdf and $F^c(t) \equiv 1 - F(t)$ being the ccdf of the abandon time, and λ being the arrival rate. We note that exact state-dependent non-Markovian abandonment rates for the $M/M/s/r + GI$ model were determined by Brandt and Brandt (2002).

In order to see how different abandon-time distributions affect performance, it is thus instructive to look at the hazard rate functions h and the approximate state-dependent rates α_j. Moreover, it is useful to look at the approximate total abandonment rate

$$\delta(k) \equiv \sum_{j=1}^{k} \alpha_j, \quad k \geq 0 \ .$$

It is interesting to look at the functions α_j and δ_k for values of j and k that occur reasonably often. When $s \approx 100$, we also have $\lambda \approx 100$. In that case the queue length is typically less than 40. The most common values for j and k may even be less than 10. That means that the hazard-rate function is mostly only relevant over the domain $(0, 0.4)$, say. Certainly, the tail of the abandon-time distribution is relatively unimportant.

In this section we complement the main paper by plotting the estimated state-dependent abandonment rates, α_j and δ_k for values of j and k that typically occur. We consider the case $\lambda = 100$ and thus let j and k range from 0 to 40.

In Figures 1 and 2 we display the functions α_j and δ_k for the Erlang E_2 abandon-time distribution with mean 1. In Figures 3 and 4 we display the functions α_j and δ_k for the lognormal $LN(1,1)$ abandon-time distribution with mean 1 and $SCV = 1$. In Figures 5 and 6 we display the functions α_j and δ_k for the lognormal $LN(1,4)$ abandon-time distribution with mean 1 and $SCV = 4$ and thus variance 4. We compare the cumulative abandonments for these three distributions with the total exponential abandonment rate when the individual customers have mean abandon time 1 in Figure 7.

As can be seen from Figures 1, 3 and 5, Erlang (E_k) and lognormal ($LN(a,b)$) distributions have zero hazard at the origin, i.e., $h(0) = 0$. For the E_k and $LN(1,1)$ distributions, that implies few abandonments for small queue sizes. However, for more variable lognormal distributions, the behavior of h at the origin can be misleading. To illustrate, note that the $LN(1,4)$ plots in Figures 5 and 6 show that the values $h(0)$ and $h'(0)$ need not be too descriptive of the hazard function away from the origin, even for time arguments that are not too large. Higher SCV values leads to even more extreme behavior.

In Table 2 in the main paper, we saw that the E_2 abandon-time distribution with mean 1 behaved much the same as the lognormal $LN(1,1)$ abandon-time distribution. Figure 7 shows that it is to be expected given the approximate abandonment rates determined from the respective hazard-rate functions.

The limiting behavior of the abandonment rates δ_k as $k \to \infty$ are determined by Brandt and Brandt (2002), but it is not clear that these limits are relevant for system performance in common cases in which k is not too large.

3. Smaller Numbers of Servers

In this section we consider what happens with a smaller number of servers. All the examples in the main paper had 100 servers. We now want to show what happens when the number of servers is reduced. We emphasize that our approximation is *not* intended for this case.

A very important reference case is the elementary $M/GI/1/\infty$ model; it is a special case occurring when $s = 1$, r is large, the load is not too large ($\rho < 1$) and the mean time to abandon, $E[T]$ is large. In that case, our approximation essentially reduces to the $M/M/1/\infty$ model. The steady-state performance of the $M/GI/1/\infty$ model is characterized by the well-

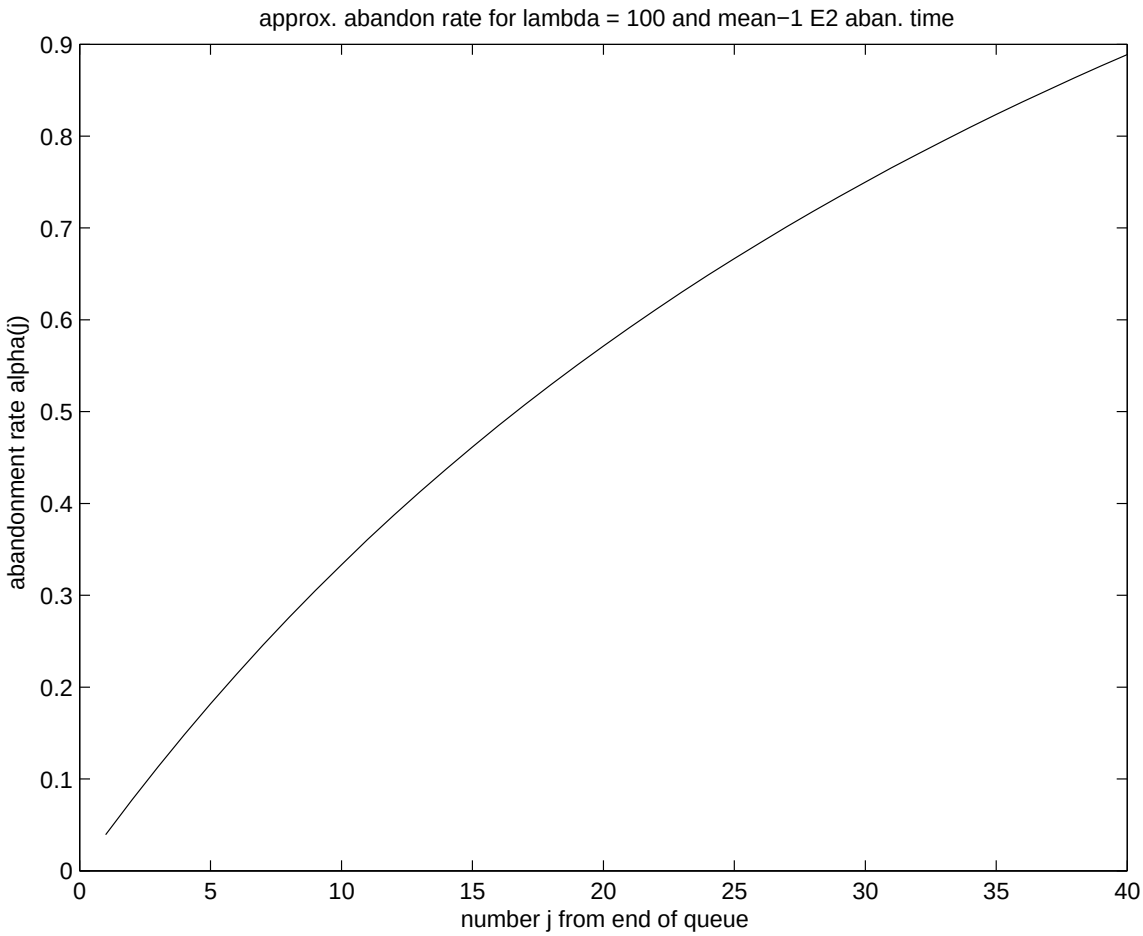

Figure 1: The approximate abandonment rate $\alpha_j = h(j/\lambda)$ for the customer j positions from the end of the queue in the $M/GI/s/r + E_2$ model with $\lambda = 100$ and Erlang E_2 abandon time having mean 1.

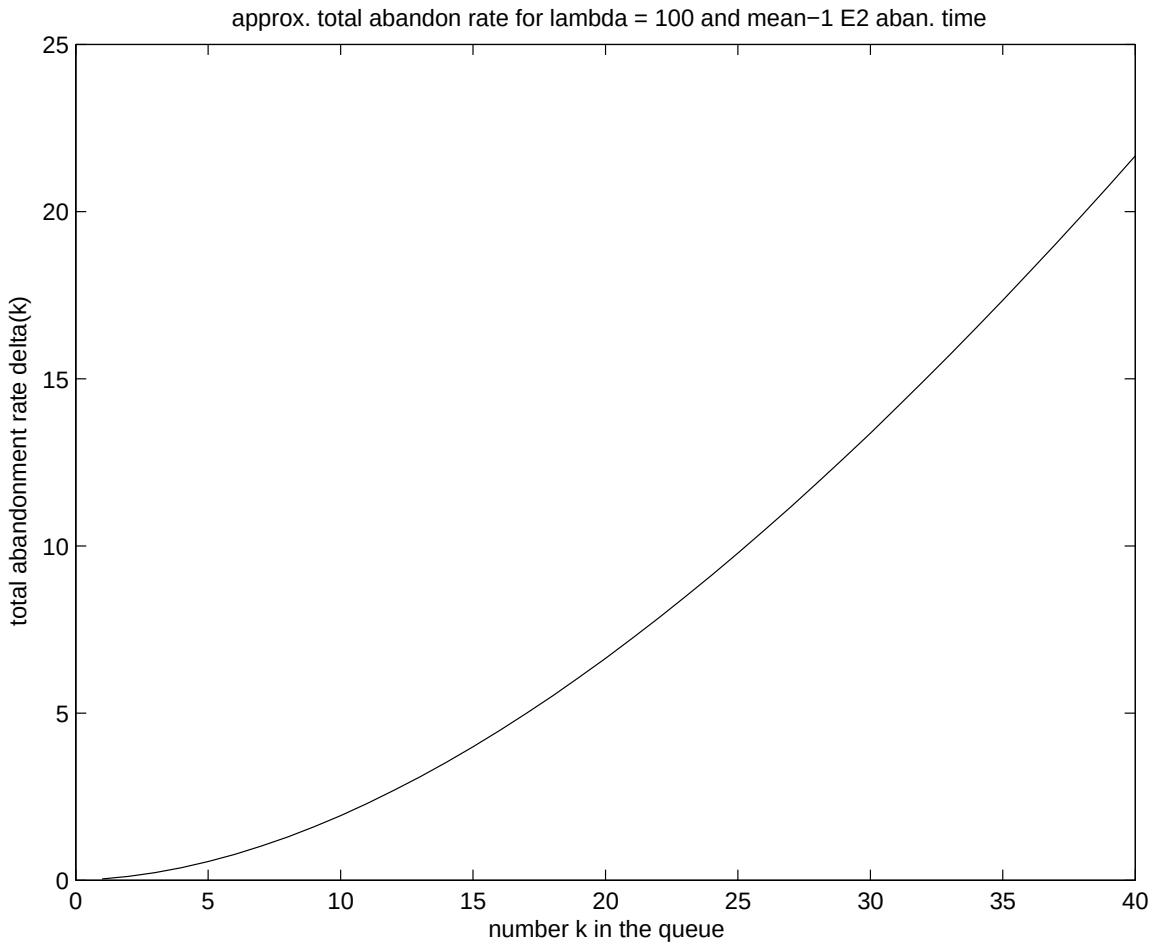

Figure 2: The approximate total abandonment rate δ_k when there are k customers in the queue in the $M/GI/s/r + E_2$ model with $\lambda = 100$ and Erlang E_2 abandon time having mean 1.

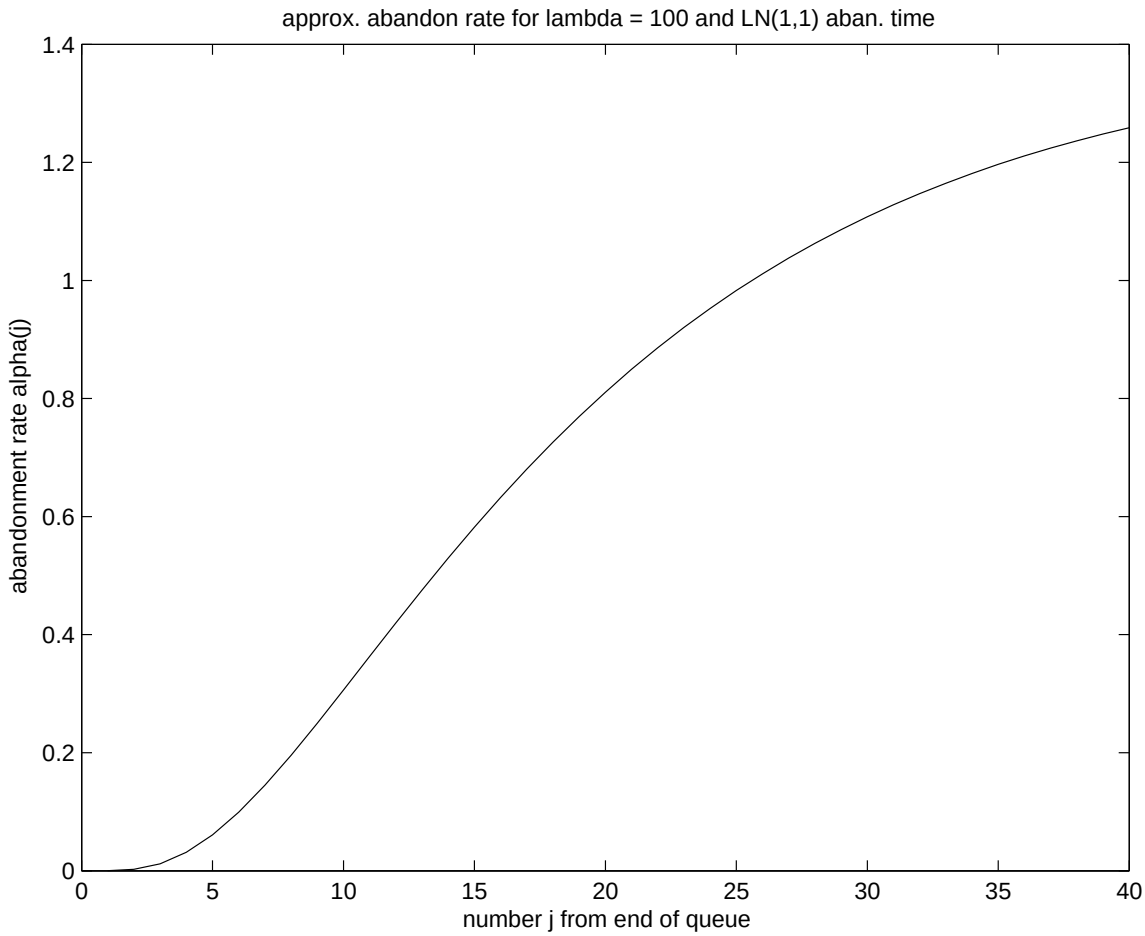

Figure 3: The approximate abandonment rate $\alpha_j = h(j/\lambda)$ for the customer j positions from the end of the queue in the $M/GI/s/r + LN(1,1)$ model with $\lambda = 100$ and lognormal $LN(1,1)$ abandon time having mean 1 and $SCV = 1$.

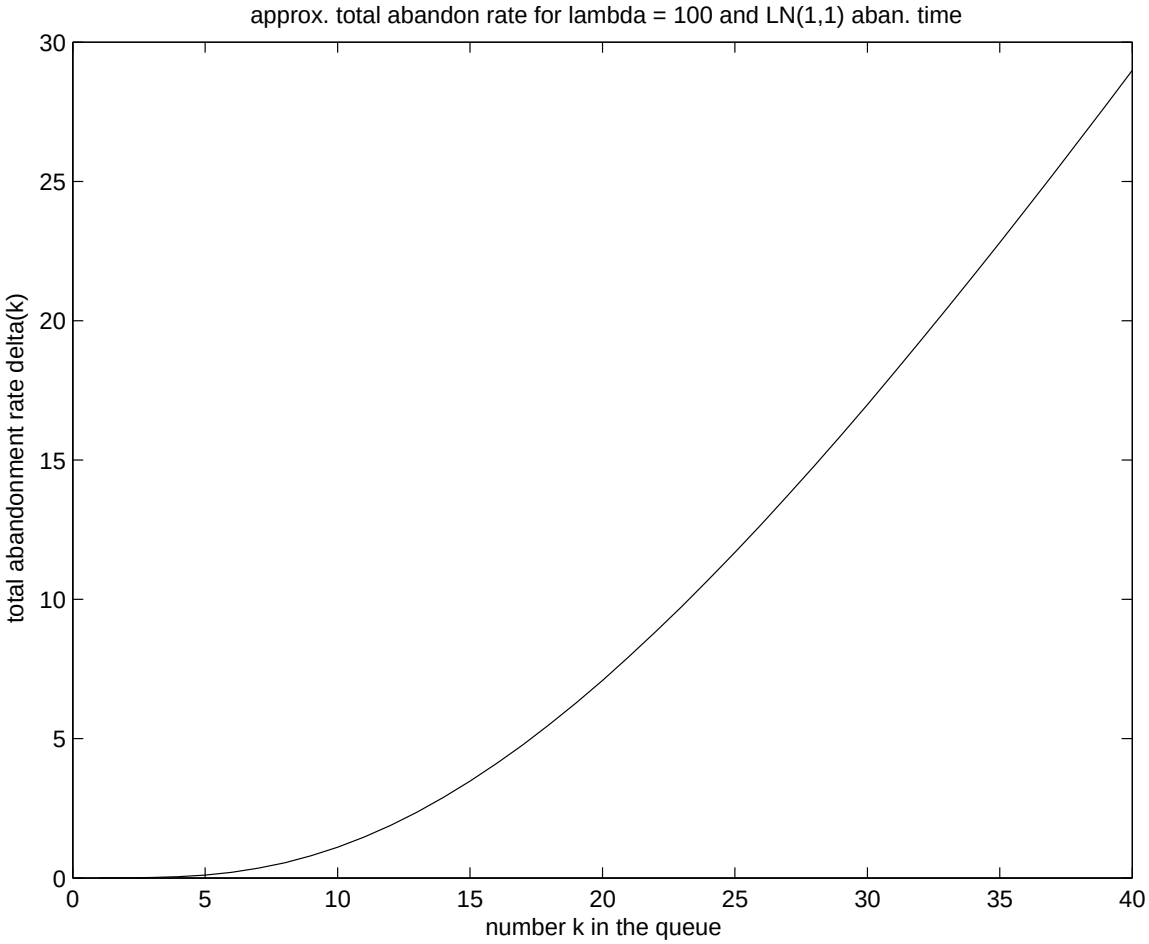

Figure 4: The approximate total abandonment rate δ_k when there are k customers in the queue in the $M/GI/s/r + LN(1,1)$ model with $\lambda = 100$ and lognormal $LN(1,1)$ abandon time having mean 1 and $SCV = 1$.

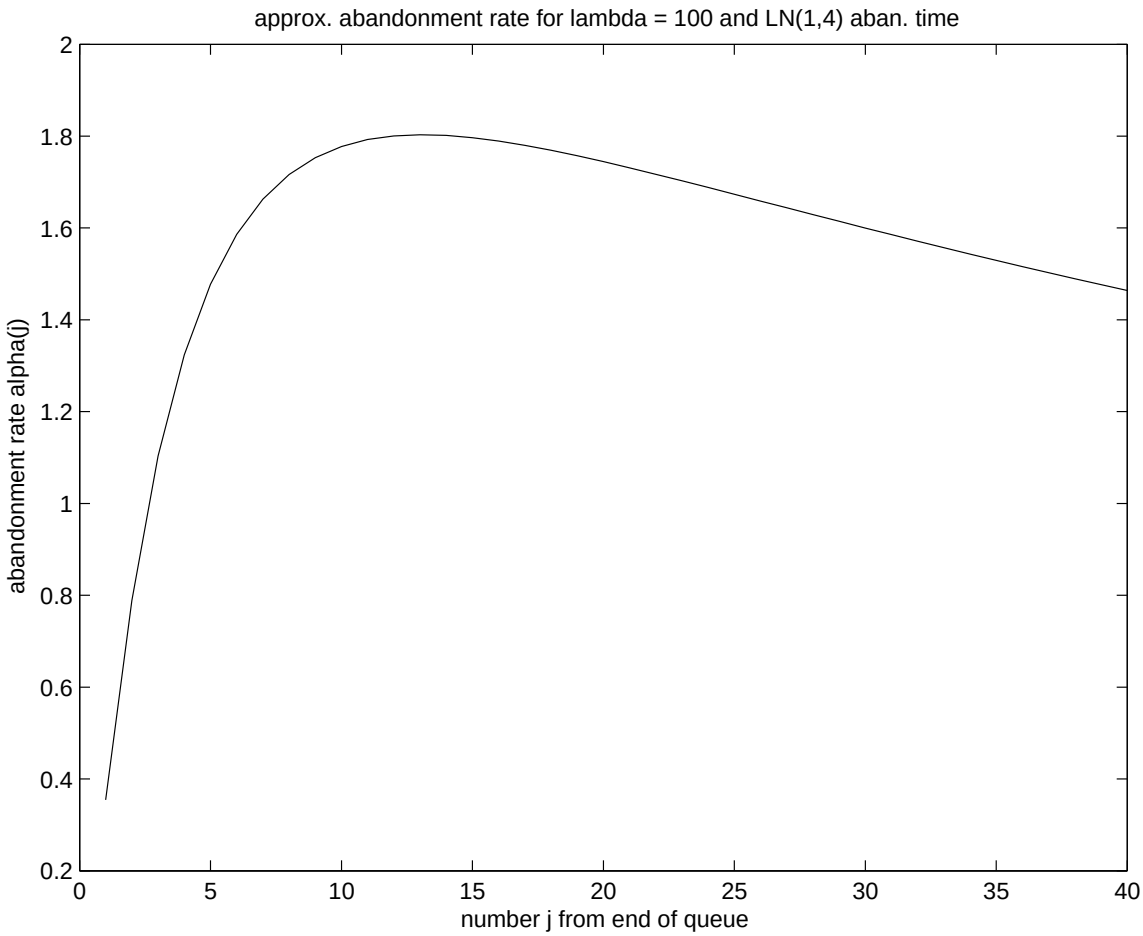

Figure 5: The approximate abandonment rate $\alpha_j = h(j/\lambda)$ for the customer j positions from the end of the queue in the $M/GI/s/r + LN(1,4)$ model with $\lambda = 100$ and lognormal $LN(1,4)$ abandon time having mean 1 and $SCV = 1$.

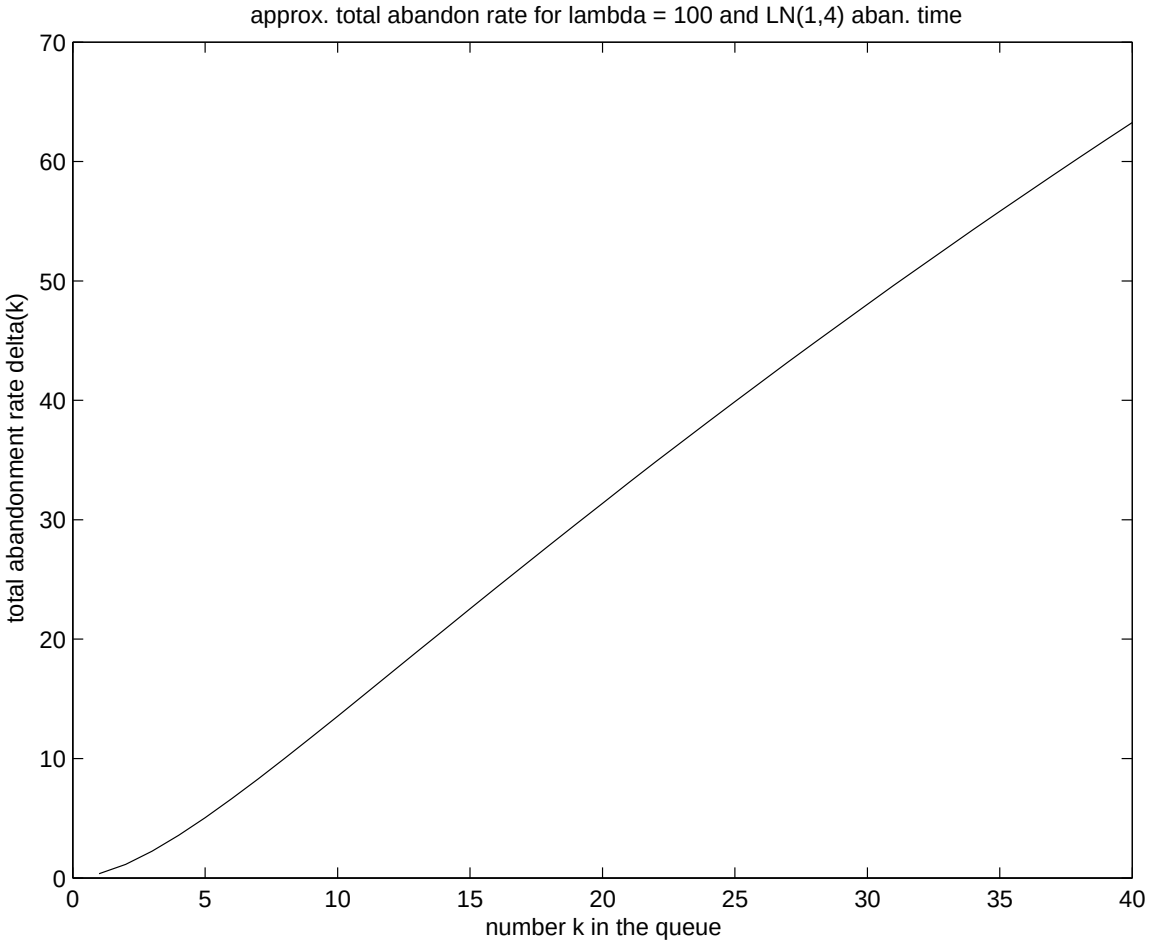

Figure 6: The approximate total abandonment rate δ_k when there are k customers in the queue in the $M/GI/s/r + LN(1,4)$ model with $\lambda = 100$ and lognormal $LN(1,4)$ abandon time having mean 1 and $SCV = 1$.

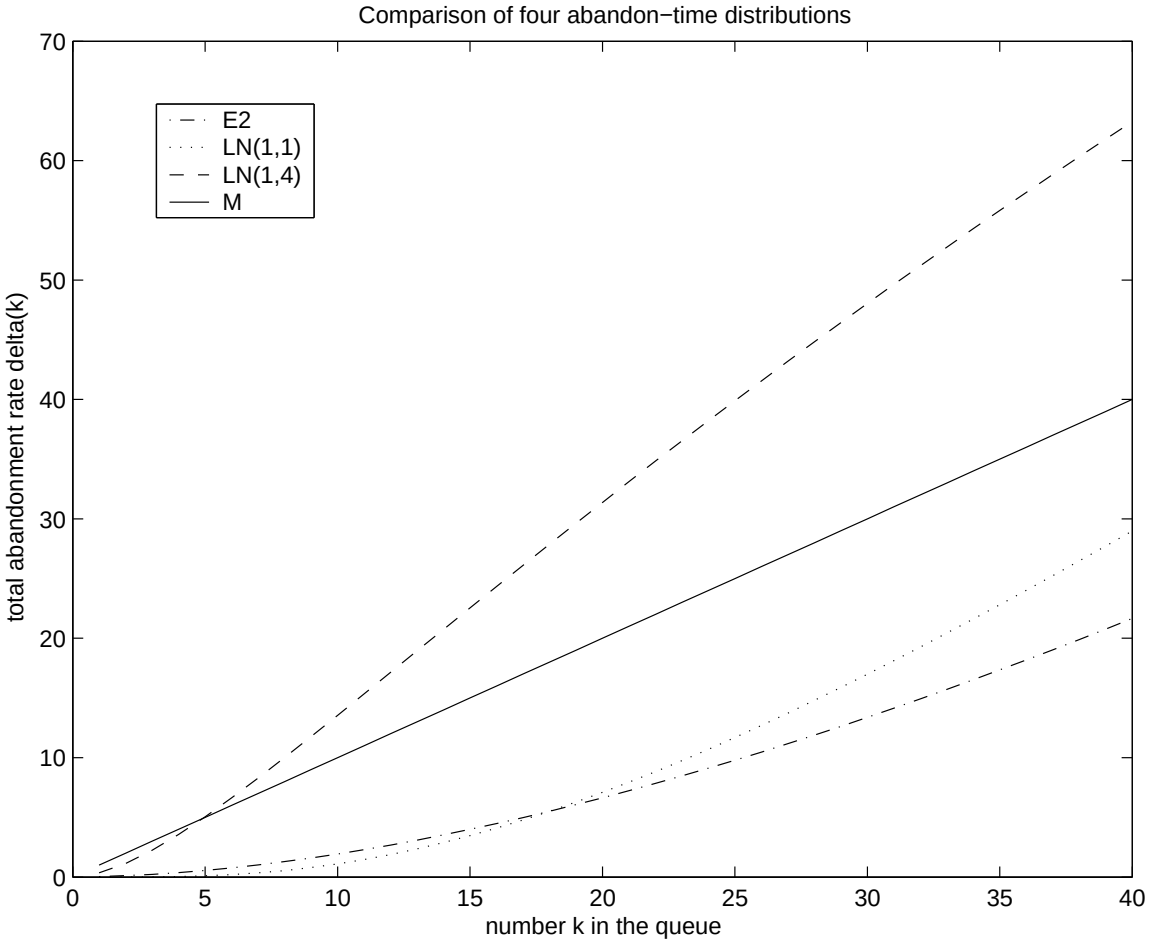

Figure 7: A comparison of four abandon-time distributions: the approximate total abandonment rate δ_k when there are k customers in the queue in the $M/GI/s/r + GI$ model with $\lambda = 100$ and four different abandon-time distributions having mean 1: Erlang E_2, lognormal $LN(1, 1)$ and $LN(1, 4)$, and exponential M.

known Pollaczek-Khintchine formulas. For example, when the mean service time is $\mu^{-1} = 1$, the mean steady-state waiting time is

$$E[W] = \frac{(1 + c_s^2)}{2} \frac{\rho}{1 - \rho} \ , \tag{3.1}$$

where $\rho \equiv \lambda/\mu$ is the traffic intensity and c_s^2 is the squared coefficient of variation (SCV, variance divided by the square of the mean) of the service-time distribution. Note that the mean waiting time is directly proportional to $1 + c_s^2$. Since the SCV of the service-time distribution plays a prominent role in the exact formulas, we see that our approximation can perform badly when the SCV c_s^2 is not close to 1. A serious limitation of our approximation is thus quite clear: The service-time approximation will produce poor results when the number of servers is too small, the SCV of the service time is not near 1 and there is negligible abandonment (e.g., the load is not too large and the mean time to abandon is not small).

Heavy-traffic limits for the $GI/GI/s/\infty$ model (without abandonments) shows that the same phenomenon occurs in heavy traffic for multiple servers, when there is no abandonment, so the $M/GI/1/\infty$ problem does not go away completely by just increasing the number of servers. But the behavior of multi-server $GI/GI/s/\infty$ queues, with moderately many servers and no abandonments, is by now quite well understood; e.g., see Whitt (1993). Thus limitations of our proposed approximations are well known.

But the situation is less clear when there are substantial abandonments. Then the simple $M/GI/1/\infty$ description no longer describes the steady-state behavior well. To provide additional insight into what happens, we now consider smaller numbers of servers. Instead of $s = 100$ servers, we now consider $s = 20$ and $s = 5$. In Table 1 here we repeat the experiment in Table 1 in the main paper with $s = 20$. In particular, we simulate the $M/E_2/20/100 + E_2$ model and the $M/M/20/100 + E_2$ model with 20 servers and arrival rate $\lambda = 18$. (As always, the mean service time is $\mu^{-1} = 1$. The mean time to abandon is $E[T] = 1$. The results in Table 1 show that the approximation still performs quite well for a smaller number of servers, but we do not regard 20 as an extremely small number. We think of the approximations as being aimed for $s = 100$, but applying reasonably well for $s \geq 20$. As before, the results in Table 1 are better than an $M/M/s/r + M$ approximation obtained by directly matching the mean abandon time, but they are not so much better.

We draw attention to two particular approximations that do not perform so well in Table 1. First, there is significant error in the abandonment probability. The exact value is 0.030, while the approximation is 0.037, an error of 23%. The error was only 10% in Table 1 in the

main paper. The 23% error is not too large for most applications, we think, but it should be noted.

Second, we see significant error in the approximation for the conditional waiting time distribution, given that the customer abandons. No such errors were present in the tables of the main paper, but we did observe in other experiments that this is a weak point of the approximation: The approximation is not consistently good in predicting the conditional waiting time distribution for the customers that abandon. The approximation is much more accurate for the conditional waiting time distribution for the customers that are served. We regard the conditional distribution for served customers as more important for applications. in the usual case of few abandonments, the overall waiting time distribution is well predicted by the approximation, because a small probability is attached to the less accurate conditional probability.

In Table 2 we consider an even smaller number of servers, in particular $s = 5$. In particular, we simulate the $M/E_2/5/100 + E_2$ model and the $M/M/5/100 + E_2$ model with 5 servers and arrival rate $\lambda = 4.0$. The results in Table 2 show that the approximation begins to degrade when the number of servers is so small. Nevertheless, the approximation performs reasonably well for this small number of servers. However, in this case, the approximation is not too different from a direct $M/M/s/r + M$ approximation obtained by directly matching the mean abandon time.

In Tables 1 and 2, the quality of the direct Erlang-A model approximation is better than in Table 1 of the main paper. That can be explained, at least in part, by the fact that the traffic intensities are smaller in these examples with fewer servers. Before, the traffic intensity was $102/100 = 1.02$. Here the traffic intensities are $18/20 = 0.90$ and $4/5 = 0.80$. Under light loads, the $M/GI/s/r + GI$ model behaves much like the $M/GI/\infty$ model, which is insensitive to the service-time distribution beyond its mean.

4. Lighter Loads

In this section we consider what happens with lighter loads. In fact, we performed an extensive set of experiments with $s = 100$ and *three* values of λ: $\lambda = 90$, $\lambda = 98$ and $\lambda = 102$. We found that the most challenging case was $\lambda = 102$, so we only displayed results for that case in the main paper. Here we supplement those results by displaying a few of the results for $s = 100$ and $\lambda = 90$.

Just as in the previous section, there is an important theoretical reference case: Here it is

	$M/GI/20/100 + GI$ *model with* $\lambda = 18$ *and* $E[T]$ *= 1.0*			
		$M/GI/20/100 + E_2$		$M/M/20/100 + M$
Performance Measure	*simulation*		*approx.*	exact
	M service	E_2 service		
$P(W = 0)$	0.587 ±0.00107	0.577 ±0.00096	0.599 –	0.651 –
$P(A)$	0.0342 ±0.00017	0.0295 ±0.00014	0.0369 –	0.0499 –
$E[Q]$	1.45 ±0.0059	1.36 ±0.0053	1.31 –	0.90 –
$Var(Q)$	7.19 ±0.033	6.13 ±0.032	6.07 –	3.68 –
$E[N]$	18.84 ±0.012	18.83 ±0.0095	18.65 –	18.0 –
$E[W\|S]$	0.075 ±0.00028	0.0711 ±0.00026	0.0682 –	0.0460 –
$Var(W\|S)$	0.0178 ±0.000075	0.0148 ±0.000071	0.0154 –	0.0094 –
$E[W\|A]$	0.240 ±0.00041	0.216 ±0.00042	0.195 –	0.124 –
$Var(W\|A)$	0.0213 ±0.00011	0.0167 ±0.000064	0.0192 –	0.0122 –
$P(W \leq 0.1\|S)$	0.742 ±0.00084	0.740 ±0.00081	0.758 –	0.826 –
$P(W \leq 0.1\|A)$	0.167 ±0.00075	0.194 ±0.0011	0.289 –	0.528 –
$P(W \leq 0.2\|S)$	0.844 ±0.00070	0.854 ±0.00060	0.860 –	0.915 –
$P(W \leq 0.2\|A)$	0.457 ±0.00089	0.517 ±0.0012	0.585 –	0.798 –

Table 1: A comparison of approximate steady-state performance measures in the $M/GI/20/100 + E_2$ model with simulations in the case of exponential (M) and Erlang (E_2) service-time distributions. The models have arrival rate $\lambda = 18$, mean service time $\mu^{-1} = 1$, and mean time to abandon $E[T] = 1.0$. These results are also compared to the exact numerical results for the purely Markovian $M/M/20/100 + M$ model, having exponential service-time and time-to-abandon distributions with the same mean. The half-width of the 95% confidence interval is given for each simulation estimate.

Performance Measure	$M/GI/5/100+GI$ *model, with* $\lambda=4$ *and* $E[T]$ *= 1.0*			
	$M/GI/5/100+E_2$ *simulation*		$M/GI/5/100+E_2$ *approx.*	$M/M/5/100+M$ exact
	M service	E_2 service		
$P(W=0)$	0.594 ±0.00038	0.584 ±0.00022	0.618 –	0.629 –
$P(A)$	0.083 ±0.00011	0.074 ±0.000091	0.096 –	0.103 –
$E[Q]$	0.550 ±0.00084	0.521 ±0.00051	0.438 –	0.410 –
$Var(Q)$	1.32 ±0.0027	1.18 ±0.0018	0.94 –	0.91 –
$E[N]$	4.22 ±0.0019	4.22 ±0.0011	4.05 –	4.00 –
$E[W\|S]$	0.113 ±0.00015	0.112 ±0.000093	0.092 –	0.086 –
$Var(W\|S)$	0.048 ±0.000080	0.042 ±0.000062	0.037 –	0.044 –
$E[W\|A]$	0.404 ±0.00023	0.363 ±0.00025	0.279 –	0.250 –
$Var(W\|A)$	0.068 ±0.00016	0.052 ±0.000090	0.056 –	0.054 –
$P(W\leq 0.1\|S)$	0.724 ±0.00030	0.710 ±0.00016	0.762 –	0.779 –
$P(W\leq 0.1\|A)$	0.076 ±0.00022	0.087 ±0.00025	0.252 –	0.310 –
$P(W\leq 0.2\|S)$	0.790 ±0.00024	0.783 ±0.00019	0.827 –	0.841 –
$P(W\leq 0.2\|A)$	0.235 ±0.00037	0.268 ±0.00057	0.466 –	0.532 –

Table 2: A comparison of approximate steady-state performance measures in the $M/GI/5/100+E_2$ model with simulations in the case of exponential (M) and Erlang (E_2) service-time distributions. The models have arrival rate $\lambda=4$, mean service time $\mu^{-1}=1$, and mean time to abandon $E[T]=1.0$. These results are also compared to the exact numerical results for the purely Markovian $M/M/5/100+M$ model, having exponential service-time and time-to-abandon distributions with the same mean. The half-width of the 95% confidence interval is given for each simulation estimate.

the $M/GI/\infty$ queue. When s and r are not too small and the loads are light, the $M/GI/s/r+GI$ model behaves much like the associated $M/GI/\infty$ model. And the $M/GI/\infty$ model is known to have the insensitivity property: Its steady-state performance in the $M/GI/\infty$ model depends upon the service-time distribution only through its mean, so that the service-time M approximation is strongly supported theoretically, and thus is clearly reasonable in this domain.

However, a caveat is in order: Under light loads, the congestion is low, so the performance measures will take small values. Experience shows that we cannot expect good relative accuracy in this domain. But that is not a serious problem, because engineering applications rarely require such relative accuracy.

We now give some examples to show what actually happens. In these examples we let $s = 100$, $\mu = 1$ and $\lambda = 90$, implying that $\rho \equiv \lambda/s\mu = 0.90$, which is a *light load* for this many servers; for discussion, see Whitt (1992, 1993). In Table 3 we display results for the $M/E_2/100/200 + E_2$ model with three different values for the mean time to abandon $E[T]$: 0.25, 1.00 and 4.00.

We regard the quality of the approximations in these cases as excellent, in fact perhaps even phenomenal. However, the relative accuracy of the probability of abandonment, $P(A)$, and the conditional mean waiting times, $E[W|S]$ and $E[W|A]$, is not always good.

In Table 4 we perform a similar experiment for the $M/M/100/200 + LN(E[T], 4)$ model. Now the time to abandon is given the relatively highly variable lognormal distribution with SCV 4. Again we let $s = 100$, $\mu = 1$ and $\lambda = 90$, so that $\rho = 0.90$. Again we consider three values for the mean time to abandon $E[T]$: 0.25, 1.00 and 4.00.

For the most part, the approximation is amazingly accurate. However, as we have observed in on several occasions, the approximation for the conditional waiting-time distribution, given that the customer abandons, is not consistently good. That is a weak point of the approximation. Since abandonment is relatively rare, that difficulty has negligible effect upon the overall waiting-time distribution. Its accuracy is about the same as for the conditional waiting-time distribution, given that the customer is served.

5. Heavy Loads

In this section we present results from experiments showing the performance of the approximation under heavy loads. We consider the same $M/GI/100/200 + GI$ model as before, but now we increase the arrival rate. We are motivated to consider this case, because the

lightly loaded $M/E_2/100/200 + E_2$ model, $\lambda = 90$						
	$E[T] = 0.25$		$E[T] = 1.00$		$E[T] = 4.00$	
perf. meas.	*sim.*	*approx.*	*sim.*	*approx.*	*sim.*	*approx.*
$P(W = 0)$	0.852	0.859	0.806	0.807	0.792	0.786
	± 0.0007	–	± 0.0015	–	± 0.0012	–
$P(A)$	0.0090	0.0107	0.0024	0.0034	0.00028	0.00047
	± 0.00009	–	± 0.00003	–	± 0.000006	–
$E[Q]$	0.544	0.521	1.15	1.27	1.48	1.82
	± 0.04	–	± 0.013	–	± 0.015	–
$E[N]$	89.74	89.56	90.94	90.96	91.47	91.78
	± 0.04	–	± 0.04	–	± 0.04	–
$E[W\|S]$	0.0056	0.0054	0.0126	0.0138	0.0164	0.0202
	± 0.00005	–	± 0.00014	–	± 0.0002	–
$E[W\|A]$	0.051	0.045	0.095	0.098	0.137	0.159
	± 0.00014	–	± 0.0007	–	± 0.003	–
$P(W \leq 0.1\|S)$	0.9922	0.9923	0.956	0.949	0.940	0.924
	± 0.00012	–	± 0.0006	–	± 0.0007	–
$P(W \leq 0.1\|A)$	0.926	0.938	0.597	0.579	0.416	0.358
	± 0.0010	–	± 0.0042	–	± 0.009	–

Table 3: A comparison of exact numerical results with simulations for the steady-state performance measures in the $M/E_2/100/200 + E_2$ model with $\lambda = 90$ and $\mu = 1$ for three different mean times to abandon: $E[T] = 0.25$, 1.00 and 4.00. The half-width of the 95% confidence interval is given for each simulation estimate.

lightly loaded $M/M/100/200 + LN(E[T], 4)$ model, $\lambda = 90$

	$E[T] = 0.25$		$E[T] = 1.00$		$E[T] = 4.00$	
perf. meas.	*sim.*	*approx.*	*sim.*	*approx.*	*sim.*	*approx.*
$P(W = 0)$	0.894 ±0.00062	0.900 –	0.842 ±0.0009	0.846 –	0.801 ±0.0023	0.801 –
$P(A)$	0.0156 ±0.00010	0.0165 –	0.0083 ±0.00005	0.0089 –	0.0023 ±0.00005	0.0026 –
$E[Q]$	0.259 ±0.0018	0.23 –	0.714 ±0.0035	0.668 –	1.43 ±0.027	1.40 –
$E[N]$	88.84 ±0.030	88.74 –	89.94 ±0.03	89.87 –	91.17 ±0.071	91.17 –
$E[W\|S]$	0.0025 ±0.000017	0.0022 –	0.0075 ±0.00004	0.0070 –	0.0156 ±0.00029	0.0152 –
$E[W\|A]$	0.027 ±0.00011	0.022 –	0.059 ±0.0002	0.051 –	0.121 ±0.00088	0.109 –
$P(W \leq 0.05\|S)$	0.987 ±0.00013	0.989 –	0.939 ±0.0003	0.943 –	0.889 ±0.0019	0.891 –
$P(W \leq 0.05\|A)$	0.879 ±0.0016	0.906 –	0.480 ±0.002	0.575 –	0.138 ±0.0019	0.208 –
$P(W \leq 0.10\|S)$	0.9991 ±0.000026	0.9993 –	0.981 ±0.00011	0.984 –	0.941 ±0.0013	0.943 –
$P(W \leq 0.10\|A)$	0.995 ±0.00025	0.996 –	0.862 ±0.0016	0.894 –	0.455 ±0.0050	0.523 –

Table 4: A comparison of exact numerical results with simulations for the steady-state performance measures in the $M/M/100/200 + LN(1, 4)$ model with $\lambda = 90$ and $\mu = 1$ for three different mean times to abandon: $E[T] = 0.25$, 1.00 and 4.00. The half-width of the 95% confidence interval is given for each simulation estimate.

abandonment-rate approximation is partly based on the assumption that abandonments are relatively rare compared to arrivals and service completions. Under heavier loads, that assumption will not hold.

We first consider the case $\lambda = 120$. We let the mean service time and mean time to abandon be 1. We consider two different distributions for the service times and abandon times: Erlang, E_2, and lognormal, $LN(1,4)$. The E_2 distribution has SCV $c_s^2 = 1/2$, while the $LN(1,4)$ distribution has SCV $c_s^2 = 4$. The hazard rate functions for these two distributions have been plotted in Figures 1 and 5. The associated total abandonment rates have been plotted in Figures 2 and 6. So we know the distributions we are considering.

We consider four cases, with each of these two distributions serving as the service-time distribution and time-to-abandon distribution. The results are displayed in Tables 5 and 6. The Table 5 shows the results for the E_2 time-to-abandon distribution, while Table 6 shows the results for the $LN(1,4)$ time-to-abandon distribution. These results show that the approximation performs spectacularly well under heavy loads. In these cases, we see that performance primarily depends on the service-time distribution through its mean, but depends significantly upon the time-to-abandon distribution beyond its mean. In each table, the results are nearly the same for the two service-time distributions. However, changing from table to table, with the different time-to-abandon distributions, we see significant differences. In Table 5, the mean queue length is about 40, while in Table 6 the mean queue length is about 14. In Table 5, the conditional mean waiting time for served customers is about 0.35, while in Table 6 the conditional mean waiting time for served customers is about 0.125.

In Tables 5 and 6 the probability of no delay, $P(W = 0)$, is very small, as should be expected because of the heavy load. The approximation correctly predicts that, but the relative accuracy is not good. That is a common phenomenon for the approximation; it does not produce small relative error for small values. However, we do not consider that a problem for most engineering applications.

From Tables 5 and 6, we see that some performance measures are the same for all four cases. In particular, the abandonment probability is very consistent throughout, being approximately $1/6 = 0.16667$. That phenomenon and other regularity in the heavily loaded cases can be explained by many-server heavy-traffic fluid limits in the overloaded or efficiency-driven (ED) limiting regime; see Whitt (2004a, b, 2005b). Indeed, the deterministic fluid approximation performs spectacularly well in this overloaded regime. It is significant that the fluid approximation strongly supports our approximation approach under heavy loads: The fluid ap-

$M/GI/100/200 + E_2$ *model with* $\lambda = 120$ *and* $E[T] = 1.0$

Perf. Meas.	*service-time distribution* E_2	$LN(1,4)$	*approx.*
$P(W = 0)$	0.00046 ±0.00006	0.0068 ±0.00035	0.0014 –
$P(A)$	0.16653 ±0.00035	0.16683 ±0.00060	0.1667 –
$E[Q]$	40.25 ±0.057	39.56 ±0.097	39.95 –
$Var(Q)$	139.6 ±0.69	221.6 ±1.09	153.9 –
$E[N]$	140.3 ±0.057	139.5 ±1.22	139.9 –
$E[W\|S]$	0.353 ±0.00051	0.343 ±0.00094	0.351 –
$Var(W\|S)$	0.0097 ±0.000058	0.0176 ±0.000087	0.0118 –
$E[W\|A]$	0.247 ±0.00025	0.261 ±0.00041	0.245 –
$Var(W\|A)$	0.0130 ±0.000032	0.0163 ±0.000065	0.0141 –
$P(W \leq 0.2\|S)$	0.063 ±0.0009	0.140 ±0.0018	0.084 –
$P(W \leq 0.2\|A)$	0.363 ±0.0008	0.348 ±0.00086	0.379 –
$P(W \leq 0.4\|S)$	0.680 ±0.0020	0.661 ±0.0024	0.672 –
$P(W \leq 0.4\|A)$	0.901 ±0.0006	0.851 ±0.0011	0.709 –

Table 5: A comparison of approximations with simulation estimates of steady-state performance measures in $M/GI/100/200 + E_2$ models under heavy load, specifically for $\lambda = 120$. The mean time to abandon is $E[T] = 1$. The two service times are Erlang (E_2) with $c_s^2 = 1/2$ and lognormal (LN(1,4) with $c_s^2 = 4$. The half-width of the 95% confidence interval is given for each simulation estimate. The common approximation based on the $M/M/100/200 + M(n)$ model are also displayed.

$M/GI/100/200 + LN(1,4)$ *model with* $\lambda = 120$ *and* $E[T] = 1.0$ *service-time distribution*			
Perf. Meas.	E_2	$LN(1,4)$	*approx.*
$P(W = 0)$	0.032 ±0.00037	0.065 ±0.00077	0.049 –
$P(A)$	0.1678 ±0.00023	0.1696 ±0.00054	0.1685 –
$E[Q]$	14.51 ±0.018	14.52 ±0.043	14.02 –
$Var(Q)$	61.1 ±0.18	81.5 ±0.30	66.6 –
$E[N]$	114.4 ±0.019	114.2 ±0.47	113.8 –
$E[W\|S]$	0.126 ±0.00017	0.125 ±0.00040	0.122 –
$Var(W\|S)$	0.0046 ±0.000014	0.0066 ±0.000027	0.0053 –
$E[W\|A]$	0.095 ±0.00008	0.103 ±0.00014	0.093 –
$Var(W\|A)$	0.0030 ±0.000006	0.0039 ±0.000014	0.0034 –
$P(W \leq 0.1\|S)$	0.359 ±0.0011	0.404 ±0.0020	0.398 –
$P(W \leq 0.1\|A)$	0.593 ±0.00054	0.554 ±0.00090	0.605 –
$P(W \leq 0.2\|S)$	0.859 ±0.00065	0.821 ±0.0015	0.856 –
$P(W \leq 0.2\|A)$	0.954 ±0.00025	0.920 ±0.00051	0.846 –

Table 6: A comparison of approximations with simulation estimates of steady-state performance measures in $M/GI/100/200 + LN(1,4)$ models under heavy load, specifically for $\lambda = 120$. The mean time to abandon is $E[T] = 1$. The two service times are Erlang (E_2) with $c_s^2 = 1/2$ and lognormal (LN(1,4) with $c_s^2 = 4$. The half-width of the 95% confidence interval is given for each simulation estimate. The common approximation based on the $M/M/100/200 + M(n)$ model are also displayed.

proximation depends upon the service-time distribution only through its mean, but it depends upon the time-to-abandon distribution beyond its mean.

We repeat the experiments above with a smaller number of servers in Tables 7 and 8. Specifically, we let $s = 20$ and $\lambda = 24$, so that the traffic intensity is again $\rho = 1.2$.

From Tables 7 and 8, we see that the $M/M/s/r + M(n)$ approximation still performs well for this smaller number of servers. For $s = 20$ and $\lambda = 24$, we again see that the performance does indeed primarily depend on the service-time distribution only through its mean, while it strongly depends on the time-to-abandon distribution beyond its mean. That is demonstrated by the similarity between the two simulation results within each table and the differences between the simulation results in the two different tables.

The traffic intensity $\rho = 1.20$ in the four tables so far is greater than we expect to see in call-center applications. However, the results show that the approximation does not break down under this heavy load. The abandonment rate is still less than 20% in all these examples. The assumption used in developing the approximations that abandonments are relatively rare compared to arrivals obviously operates with great force when the abandonment rate is lower than 5%. Nevertheless, the assumption remains roughly reasonable when the abandonment rate increases to 20%.

It is of course interesting to see if the quality of the approximation will indeed eventually degrade if the abandonment rate is high enough. To investigate this feature, we consider the unrealistic case of $s = 100$ and $\lambda = 200$. Then the abandonment rate is about 50%. Results for this case are displayed in Table 9. Surprisingly, even in this case, the approximations perform well. It thus remains to identify heavy-load cases causing the performance of the approximation to degrade.

6. Staffing Application

In Section 2 of the main paper we asserted that the approximation if effective for staffing decisions. In this section we provide extra details supporting that claim. In particular, we give results for the $M/E_2/s/200 + E_2$ model with arrival rate $\lambda = 100$ and different numbers of servers. The models have common arrival rate $\lambda = 100$, mean service time $\mu^{-1} = 1$, number of extra waiting spaces $r = 200$ and mean time to abandon 1.0. The goal is to choose the number s of servers so that both $P(A) \leq 0.05$ and $P(W \leq 0.1|S) \geq 0.8$. In Table 10 we display results for $s = 103$, $s = 104$ and $s = 105$. The middle case, $s = 104$, is the least number of servers meeting the performance constraints.

$M/GI/20/100 + E_2$ *model with* $\lambda = 24$ *and* $E[T] = 1.0$			
	service-time distribution		
Perf. Meas.	E_2	$LN(1,4)$	*approx.*
$P(W = 0)$	0.068 ±0.0004	0.126 ±0.0007	0.096 –
$P(A)$	0.1748 ±0.00031	0.1838 ±0.00034	0.1783 –
$E[Q]$	7.7 ±0.013	7.6 ±0.013	7.2 –
$Var(Q)$	25.2 ±0.04	33.3 ±0.07	25.3 –
$E[N]$	27.5 ±0.015	27.2 ±0.015	26.9 –
$E[W\|S]$	0.322 ±0.0005	0.307 ±0.0005	0.303 –
$Var(W\|S)$	0.042 ±0.00006	0.061 ±0.00013	0.046 –
$E[W\|A]$	0.309 ±0.0003	0.351 ±0.0004	0.288 –
$Var(W\|A)$	0.028 ±0.00005	0.040 ±0.00009	0.031 –
$P(W \leq 0.2\|S)$	0.297 ±0.0009	0.384 ±0.0008	0.350 –
$P(W \leq 0.2\|A)$	0.292 ±0.0004	0.253 ±0.0005	0.356 –
$P(W \leq 0.4\|S)$	0.647 ±0.0010	0.659 ±0.0009	0.676 –
$P(W \leq 0.4\|A)$	0.725 ±0.0007	0.641 ±0.0008	0.753 –

Table 7: A comparison of approximations with simulation estimates of steady-state performance measures in $M/GI/20/100 + E_2$ models under heavy load, but with a smaller number of servers, specifically for $s = 20$ and $\lambda = 24$. The mean time to abandon is $E[T] = 1$. The two service times are Erlang (E_2) with $c_s^2 = 1/2$ and lognormal (LN(1,4) with $c_s^2 = 4$. The half-width of the 95% confidence interval is given for each simulation estimate. The common approximation based on the $M/M/20/100 + M(n)$ model are also displayed.

$M/GI/20/100 + LN(1,4)$ *model with* $\lambda = 24$ *and* $E[T] = 1.0$ *service-time distribution*			
Perf. Meas.	E_2	$LN(1,4)$	*approx.*
$P(W = 0)$	0.210 ±0.00047	0.254 ±0.00064	0.252 –
$P(A)$	0.1914 ±0.00022	0.1993 ±0.00044	0.1974 –
$E[Q]$	3.15 ±0.0036	3.26 ±0.0081	2.90 –
$Var(Q)$	9.6 ±0.10	12.0 ±0.037	9.7 –
$E[N]$	22.56 ±0.0052	22.48 ±0.0092	22.17 –
$E[W\|S]$	0.129 ±0.00016	0.130 ±0.00035	0.119 –
$Var(W\|S)$	0.0166 ±0.000025	0.0227 ±0.00010	0.0177 –
$E[W\|A]$	0.139 ±0.00008	0.159 ±0.00022	0.130 –
$Var(W\|A)$	0.0091 ±0.00002	0.0140 ±0.00006	0.0115 –
$P(W \leq 0.1\|S)$	0.494 ±0.00065	0.537 ±0.0009	0.549 –
$P(W \leq 0.1\|A)$	0.420 ±0.00035	0.380 ±0.0007	0.486 –
$P(W \leq 0.2\|S)$	0.731 ±0.00042	0.731 ±0.0009	0.758 –
$P(W \leq 0.2\|A)$	0.778 ±0.00018	0.717 ±0.0006	0.784 –

Table 8: A comparison of approximations with simulation estimates of steady-state performance measures in $M/GI/20/100 + LN(1,4)$ models under heavy load, but with a smaller number of servers, specifically for $s = 20$ and $\lambda = 24$. The mean time to abandon is $E[T] = 1$. The two service times are Erlang (E_2) with $c_s^2 = 1/2$ and lognormal (LN(1,4) with $c_s^2 = 4$. The half-width of the 95% confidence interval is given for each simulation estimate. The common approximation based on the $M/M/20/100 + M(n)$ model are also displayed.

$M/GI/100/400 + GI$ *model with* $\lambda = 200$ *and* $E[T] = 1.0$				
	time-to-abandon distribution			
	E_2		$LN(1,4)$	
Perf. Meas.	*sim.*	*approx.*	*sim.*	*approx.*
$P(W = 0)$	0.00027	≈ 0	0.00028	≈ 0
	± 0.000006	–	± 0.000015	–
$P(A)$	0.50046	0.5000	0.49913	0.5000
	± 0.00092	–	± 0.00073	–
$E[Q]$	131.2	135.0	65.2	63.2
	± 0.30	–	± 0.20	–
$Var(Q)$	219.4	174.0	137.7	127.4
	± 3.6	–	± 1.7	–
$E[N]$	231.2	235.0	165.2	163.3
	± 0.30	–	± 0.20	–
$E[W\|S]$	0.8326	0.864	0.443	0.427
	± 0.0021	–	± 0.0016	–
$Var(W\|S)$	0.0084	0.0077	0.0072	0.0064
	± 0.00015	–	± 0.00011	–
$E[W\|A]$	0.479	0.243	0.209	0.205
	± 0.00069	–	± 0.00060	–
$Var(W\|A)$	0.049	0.52	0.0163	0.0157
	± 0.00023	–	± 0.00011	–

Table 9: A comparison of approximations with simulation estimates of steady-state performance measures in $M/M/100/400 + GI$ models under extremely heavy load, specifically for $s = 100$ and $\lambda = 200$. The mean time to abandon is $E[T] = 1$. The two time-to-abandon distributions are Erlang (E_2) with $c_s^2 = 1/2$ and lognormal (LN(1,4) with $c_s^2 = 4$. The half-width of the 95% confidence interval is given for each simulation estimate. The run lengths were divided by 10 for this example.

From the simulation results and the associated displayed approximations, we see that the approximation yields the correct decision in this case, as well as good approximations for the performance measures themselves. In contrast, if we used the $M/M/s/r + M$ (Erlang A) model, obtaining the exponential abandon-time distribution by simply picking an exponential distribution with the same mean as the given Erlang E_2 abandon-time distribution, we would select $s = 99$ servers, an error of 5%, in the decision variable. We display the exact numerical results for the $M/M/s/r + M$ model with $\lambda = 100$ and four values of s in Table 11.

7. Two Uniform Distributions

In this section we present additional evidence to show that the approximation is effective. In particular, in this section we consider two uniform distributions with mean 1. The first is $U(0, 2)$, which is uniform on the interval $(0, 2)$, while the second is $U(0.5, 1.5)$, which is uniform on the interval $(0.5, 1.5)$. From some perspectives, the two uniform distributions are not too different. The uniform on $(0, 2)$ is somewhat more variable than the uniform on $(0.5, 1.5)$. However, from the perspective of abandonments, the $U(0.5, 1.5)$ distribution is very different, because it does not permit any very small values. It does not permit any values less than 0.5. As a consequence, the hazard rates for small values are strikingly different for these two uniform distributions.

We now analyze the multiserver queue with each of the four combinations of these two uniform distributions playing the role of the service-time distribution and the abandonment-time distribution. In Table 12 we analyze the $M/GI/100/200 + U(0, 2)$ model in which the service-time distribution is either $U(0, 2)$ or $U(0.5, 1.5)$. In Table 13 we analyze the $M/GI/100/200 + U(0.5, 1.5)$ model in which the service-time distribution is either $U(0, 2)$ or $U(0.5, 1.5)$. We see that the specific uniform distribution makes a great difference as an abandon-time distribution, but very little difference as a service-time distribution. In all these cases, just as in most previous tables, the arrival rate is $\lambda = 102$, the individual mean service time is $\mu^{-1} = 1$, there are $s = 100$ servers, there are $r = 200$ extra waiting spaces and the mean abandon time is 1.

In order to better understand these uniform-distribution examples, we also present plots of the approximate abandonment rates, α_j and δ_k, for relevant values of j and k in Figures 8 – 11. From these plots, we see that the two uniform distributions should indeed behave very differently as abandonment distributions.

As mentioned before, exact non-Markovian abandonment rates for the $M/M/s/r + GI$

$M/E_2/s/200 + E_2$ *model with common* $\lambda = 100$ *and different* s

	$s = 103$		$s = 104$		$s = 105$	
Perf. Meas.	*sim.*	*approx.*	*sim.*	*approx.*	*sim.*	*approx.*
$P(W = 0)$	0.472 ±0.0023	0.491 –	0.527 ±0.0016	0.541 –	0.576 ±0.0020	0.589 –
$P(A)$	0.0135 ±0.00011	0.0161 –	0.0108 ±0.00009	0.0132 –	0.0087 ±0.00008	0.0108 –
$E[Q]$	5.34 ±0.032	5.53 –	4.48 ±0.027	4.68 –	3.75 ±0.030	3.94 –
$Var(Q)$	59.8 ±0.43	67.2 –	50.6 ±0.44	57.5 –	42.4 ±0.41	48.6 –
$E[N]$	104.0 ±0.060	103.9 –	103.4 ±0.037	103.4 –	102.9 ±0.057	102.9 –
$E[W\|S]$	0.052 ±0.00031	0.054 –	0.044 ±0.00026	0.045 –	0.037 ±0.00029	0.038 –
$Var(W\|S)$	0.0054 ±0.000039	0.0063 –	0.0046 ±0.000037	0.0053 –	0.0039 ±0.000036	0.0045 –
$E[W\|A]$	0.124 ±0.00037	0.127 –	0.119 ±0.00064	0.122 –	0.114 ±0.00052	0.117 –
$Var(W\|A)$	0.0051 ±0.000049	0.0060 –	0.0047 ±0.000062	0.0057 –	0.0044 ±0.000056	0.0054 –
$P(W \le 0.1\|S)$	0.776 ±0.0013	0.769 –	0.815 ±0.0013	0.806 –	0.848 ±0.0015	0.840 –
$P(W \le 0.1\|A)$	0.426 ±0.0014	0.423 –	0.452 ±0.0032	0.446 –	0.479 ±0.0027	0.470 –
$P(W \le 0.2\|S)$	0.941 ±0.00073	0.928 –	0.956 ±0.00058	0.944 –	0.967 ±0.00063	0.958 –
$P(W \le 0.2\|A)$	0.851 ±0.0017	0.828 –	0.869 ±0.0027	0.845 –	0.886 ±0.0026	0.862 –

Table 10: A comparison of approximations with simulations in $M/E_2/s/200 + E_2$ models with common arrival rate $\lambda = 100$ but different numbers s of servers, to show the impact on staffing decisions. The models have common arrival rate $\lambda = 100$, mean service time $\mu^{-1} = 1$, number of extra waiting spaces $r = 200$ and mean time to abandon 1.0. The goal is to choose the number s of servers so that both $P(A) \le 0.05$ and $P(W \le 0.1|S) \ge 0.8$. The three displayed cases are $s = 103$, $s = 104$ and $s = 105$. Here the waiting-time constraint is binding, so that the appropriate number of servers is $s = 104$.

$M/M/s/200+M$ *model with common* $\lambda = 100$ *and different* s

	exact numerical values			
Perf. Meas.	$s = 98$	$s = 99$	$s = 100$	$s = 104$
$P(W = 0)$	0.407	0.447	0.487	0.642
$P(A)$	0.0505	0.0450	0.0399	0.0233
$E[Q]$	5.05	4.50	3.99	2.33
$Var(Q)$	43.8	39.6	35.4	21.1
$E[N]$	100.0	100.0	100.0	100.0
$E[W\|S]$	0.049	0.044	0.039	0.023
$Var(W\|S)$	0.0043	0.0038	0.0034	0.0020
$E[W\|A]$	0.067	0.065	0.063	0.054
$Var(W\|A)$	0.0032	0.0030	0.0028	0.0023
$P(W \le 0.1\|S)$	0.795	0.823	0.847	0.922
$P(W \le 0.1\|A)$	0.763	0.778	0.792	0.843
$P(W \le 0.2\|S)$	0.963	0.970	0.977	0.992
$P(W \le 0.2\|A)$	0.969	0.973	0.977	0.987

Table 11: Approximations by the $M/M/s/200+M$ model for staffing in the $M/E_2/s/200+E_2$ models with common arrival rate $\lambda = 100$ but different numbers s of servers, to show the impact on staffing decisions. The approximation here matches the mean abandon time. The models have common arrival rate $\lambda = 100$, mean service time $\mu^{-1} = 1$, number of extra waiting spaces $r = 200$ and mean time to abandon 1.0. The goal is to choose the number s of servers so that both $P(A) \le 0.05$ and $P(W \le 0.1|S) \ge 0.8$. The four displayed cases are $s = 98$, $s = 99$, $s = 100$ and $s = 104$. Here the waiting-time constraint is binding, so that the appropriate number of servers is $s = 99$. The correct value for the $M/E_2/s/200 + E_2$ model is $s = 104$, as can be seen from Table 10.

$M/GI/100/200 + U(0,2)$ *model with mean time to abandon = 1.0*

	service-time distribution		
Perf. Meas.	$U(0,2)$	$U(0.5,1.5)$	*approx.*
$P(W=0)$	0.295 ±0.0013	0.276 ±0.0017	0.320 –
$P(A)$	0.0401 ±0.00020	0.0382 ±0.00014	0.0433 –
$E[Q]$	7.84 ±0.040	7.49 ±0.034	8.39 –
$Var(Q)$	72.8 ±0.59	62.3 ±0.47	91.0 –
$E[N]$	105.7 ±0.049	105.6 ±0.040	106.0 –
$E[W\|S]$	0.077 ±0.00038	0.073 ±0.00034	0.082 –
$Var(W\|S)$	0.0068 ±0.000053	0.0057 ±0.000041	0.0087 –
$E[W\|A]$	0.084 ±0.00035	0.077 ±0.00025	0.096 –
$Var(W\|A)$	0.0046 ±0.000045	0.0038 ±0.000031	0.0062 –
$P(W \le 0.1\|S)$	0.662 ±0.0017	0.674 ±0.0018	0.651 –
$P(W \le 0.1\|A)$	0.663 ±0.0013	0.703 ±0.0015	0.607 –
$P(W \le 0.2\|S)$	0.906 ±0.0013	0.928 ±0.00098	0.874 –
$P(W \le 0.2\|A)$	0.930 ±0.0013	0.952 ±0.00086	0.889 –

Table 12: A comparison of approximations with simulation estimates of steady-state performance measures in models with uniform service-time and abandon-time distributions. The specific model is in $M/GI/100/200 + U(0,2)$, where the abandon-time distribution is uniform on the interval $(0,2)$. models with two different service-time distributions having common mean 1: $U(0,2)$ and $U(0.5,1.5)$, both having mean 1.0. 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 $U(0,2)$ 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.

$M/GI/100/200 + U(0.5, 1.5)$ *model with mean time to abandon = 1.0*			
	service-time distribution		
Perf. Meas.	$U(0, 2)$	$U(0.5, 1.5)$	*approx.*
$P(W = 0)$	0.0438	0.0296	0.0776
	$\pm$0.0014	$\pm$0.0010	–
$P(A)$	0.0225	0.0215	0.0254
	$\pm$0.00022	$\pm$0.00026	–
$E[Q]$	38.2	39.7	35.1
	$\pm$0.20	$\pm$0.17	–
$Var(Q)$	381.7	339.9	416.7
	$\pm$3.4	$\pm$3.3	–
$E[N]$	137.9	139.5	134.5
	$\pm$0.21	$\pm$0.17	–
$E[W\|S]$	0.370	0.386	0.340
	$\pm$0.0019	$\pm$0.0015	–
$Var(W\|S)$	0.0331	0.0290	0.0379
	$\pm$0.00033	$\pm$0.00030	–
$E[W\|A]$	0.558	0.553	0.504
	$\pm$0.00023	$\pm$0.00017	–
$Var(W\|A)$	0.0023	0.0019	0.0060
	$\pm$0.000024	$\pm$0.000010	–
$P(W \leq 0.2\|S)$	0.202	0.165	0.267
	$\pm$0.0043	$\pm$0.0033	–
$P(W \leq 0.2\|A)$	0.000	0.000	0.000
	–	–	–
$P(W \leq 0.4\|S)$	0.499	0.460	0.549
	$\pm$0.0023	$\pm$0.0034	–
$P(W \leq 0.4\|A)$	0.000	0.000	0.000
	–	–	–

Table 13: A comparison of approximations with simulation estimates of steady-state performance measures in models with uniform service-time and abandon-time distributions. The specific model is in $M/GI/100/200 + U(0.5, 1.5)$, where the abandon-time distribution is uniform on the interval $(0.5, 1.5)$. models with two different service-time distributions having common mean 1: $U(0, 2)$ and $U(0.5, 1.5)$, both having mean 1.0. 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 $U(0.5, 1.5)$ 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.

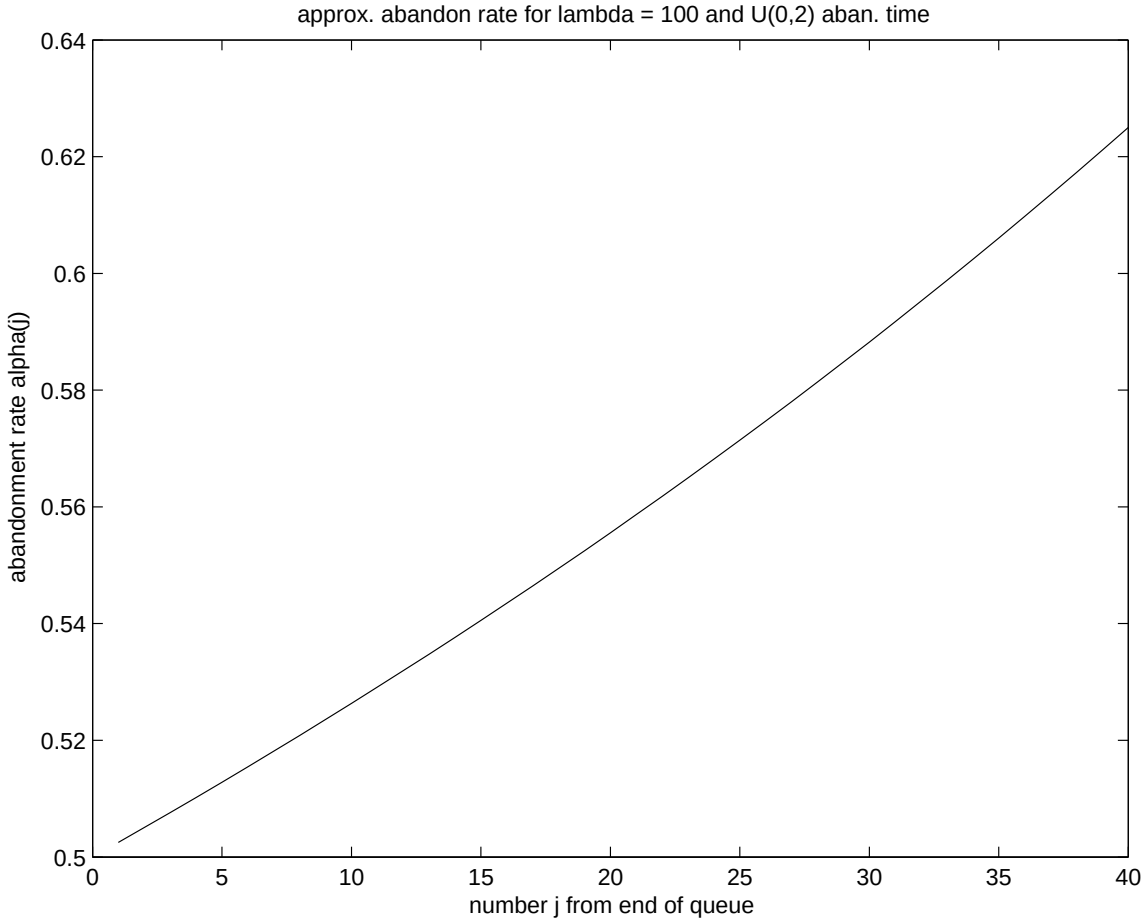

Figure 8: The approximate abandonment rate $\alpha_j = h(j/\lambda)$ for the customer j positions from the end of the queue in the $M/GI/s/r + U(0,2)$ model with $\lambda = 100$ and uniform $U(0,2)$ abandon time having mean 1.

model are determined by Brandt and Brandt (2002). See Figure 11 in Mandelbaum and Zeltyn (2004) for exact abandonment rates for different distributions.

8. Other Examples

In this section we present some additional tables. These tables originally were in the main paper, but were deleted to meet space limitations.

The first table, Table 14, repeats Table 3 in the main paper. It presents additional evidence for the performance of the algorithm when the service times are exponential, complementing Tables 2-4 in the main paper. Table 14 here describes results for the $M/M/100/200+LN(4,4)$ model, having arrival rate $\lambda = 102$, IID exponential service times with mean $\mu^{-1} = 1$, $s = 100$ servers, $r = 200$ extra waiting spaces (designed to make blocking negligible) and a lognormal $LN(4,4)$ abandon-time distribution with mean 4 and $SCV = 4$, and thus variance 64. Table

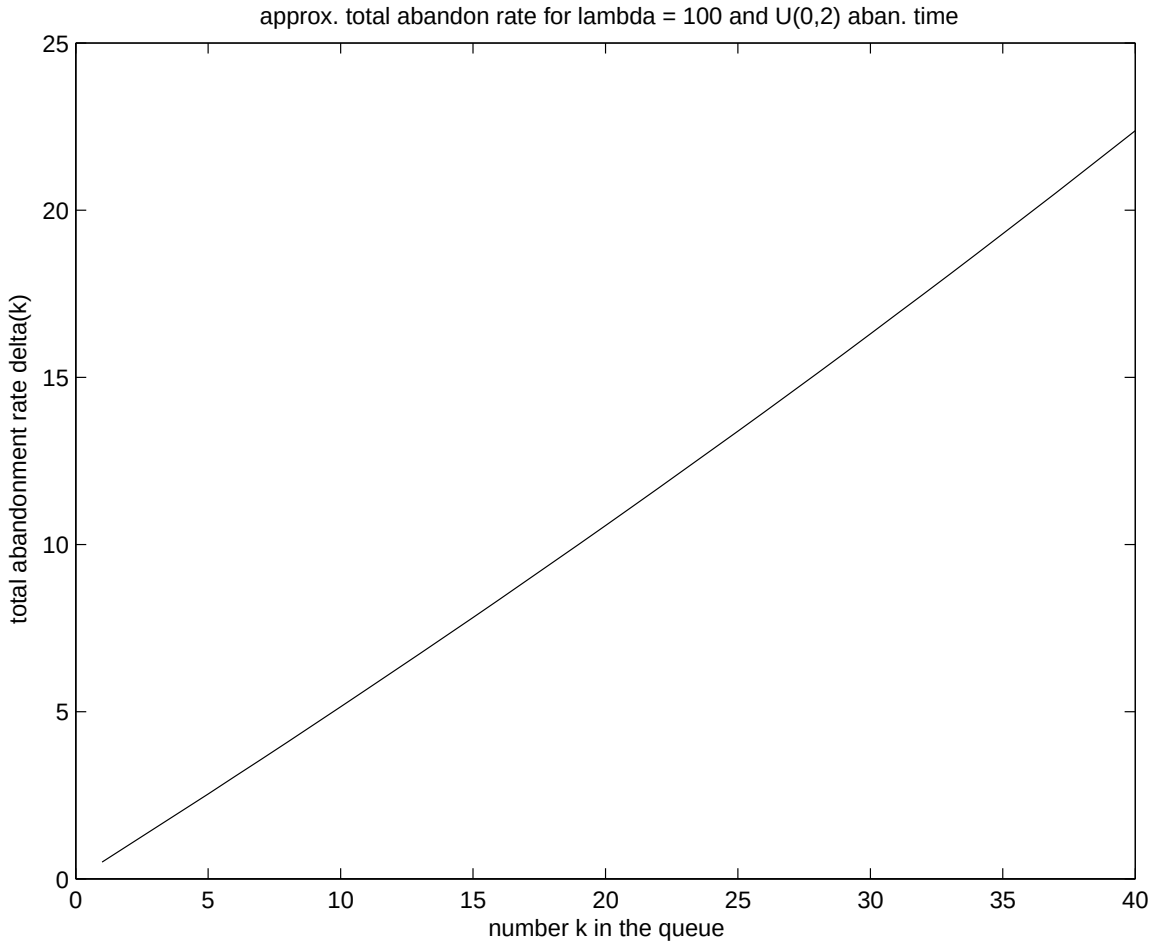

Figure 9: The approximate total abandonment rate δ_k when there are k customers in the queue in the $M/GI/s/r+U(0,2)$ model with $\lambda = 100$ and uniform $U(0,2)$ abandon time having mean 1.

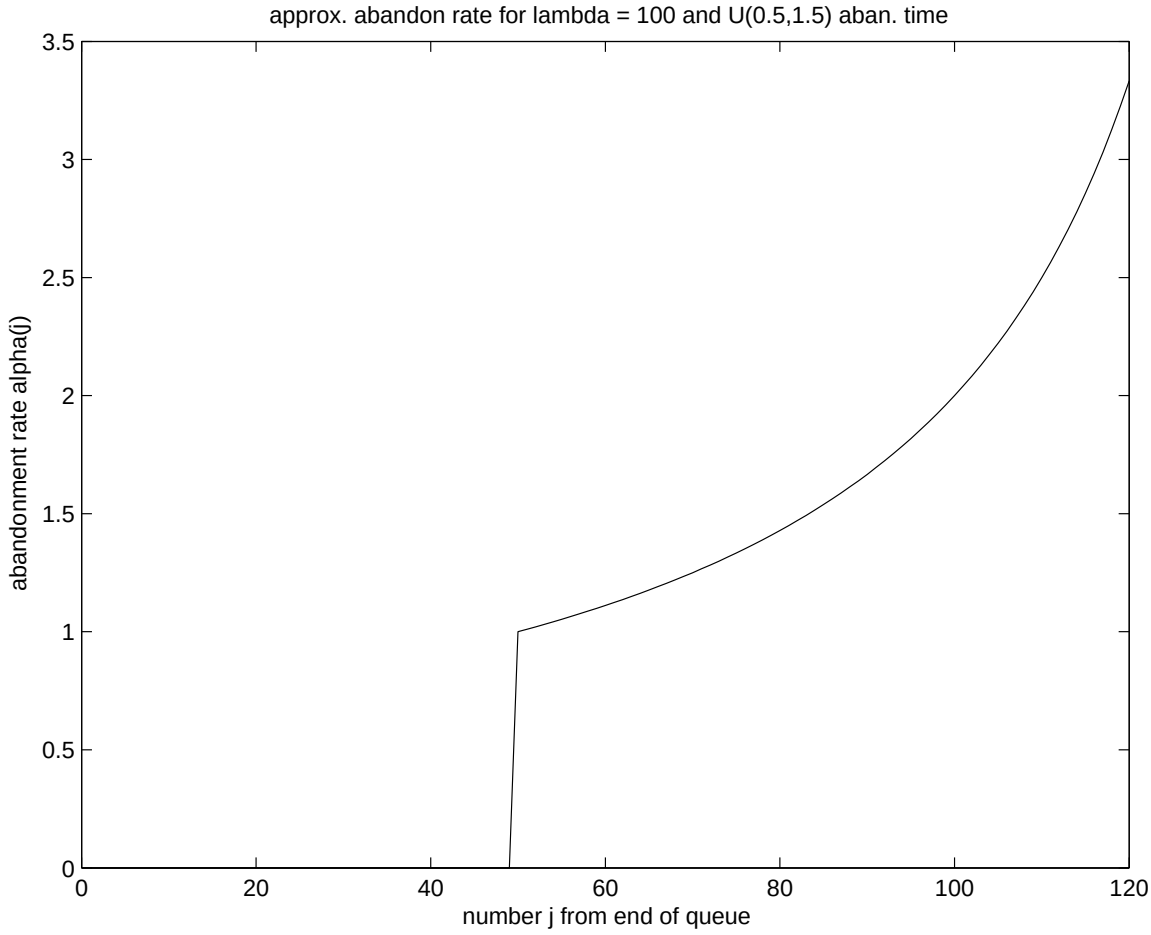

Figure 10: The approximate abandonment rate $\alpha_j = h(j/\lambda)$ for the customer j positions from the end of the queue in the $M/GI/s/r+U(0.5, 1.5)$ model with $\lambda = 100$ and uniform $U(0.5, 1.5)$ abandon time having mean 1.

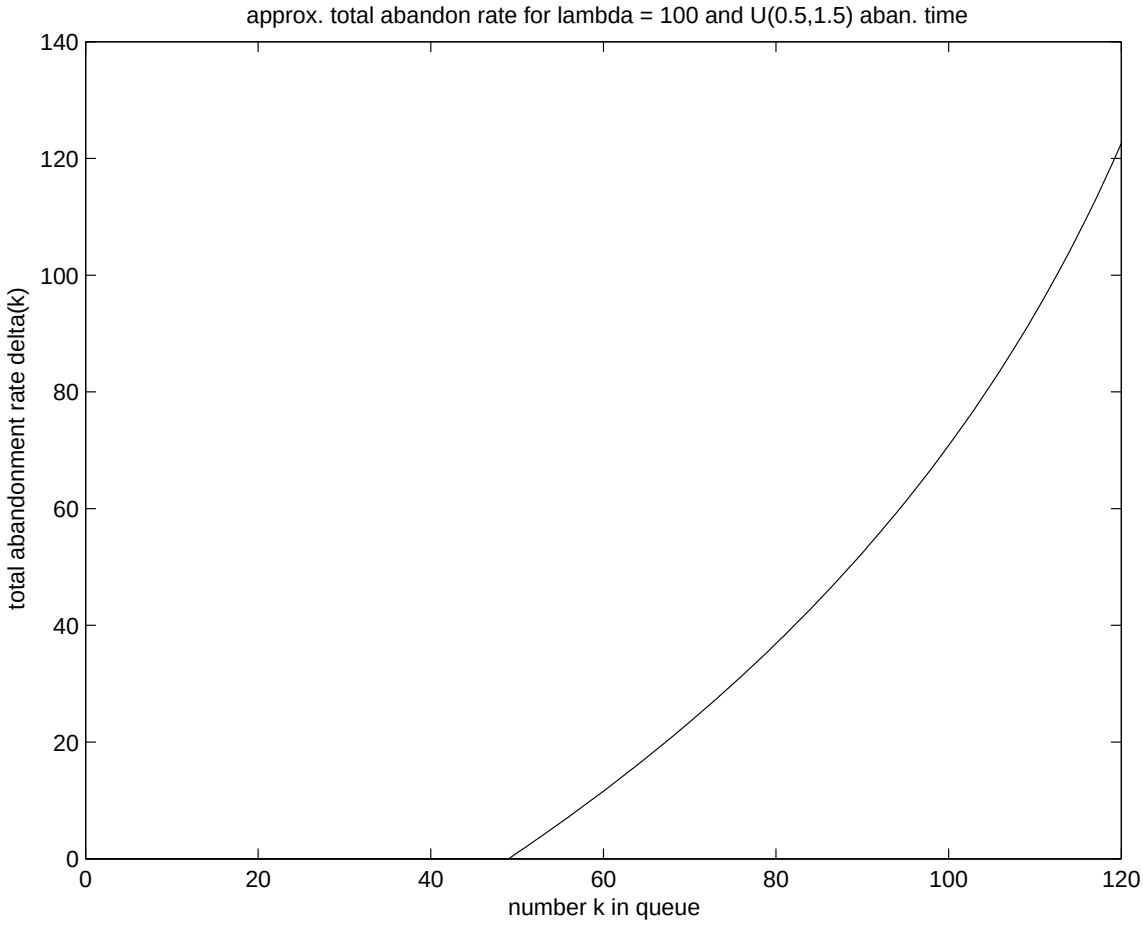

Figure 11: The approximate total abandonment rate δ_k when there are k customers in the queue in the $M/GI/s/r + U(0.5, 1.5)$ model with $\lambda = 100$ and uniform $U(0.5, 1.5)$ abandon time having mean 1.

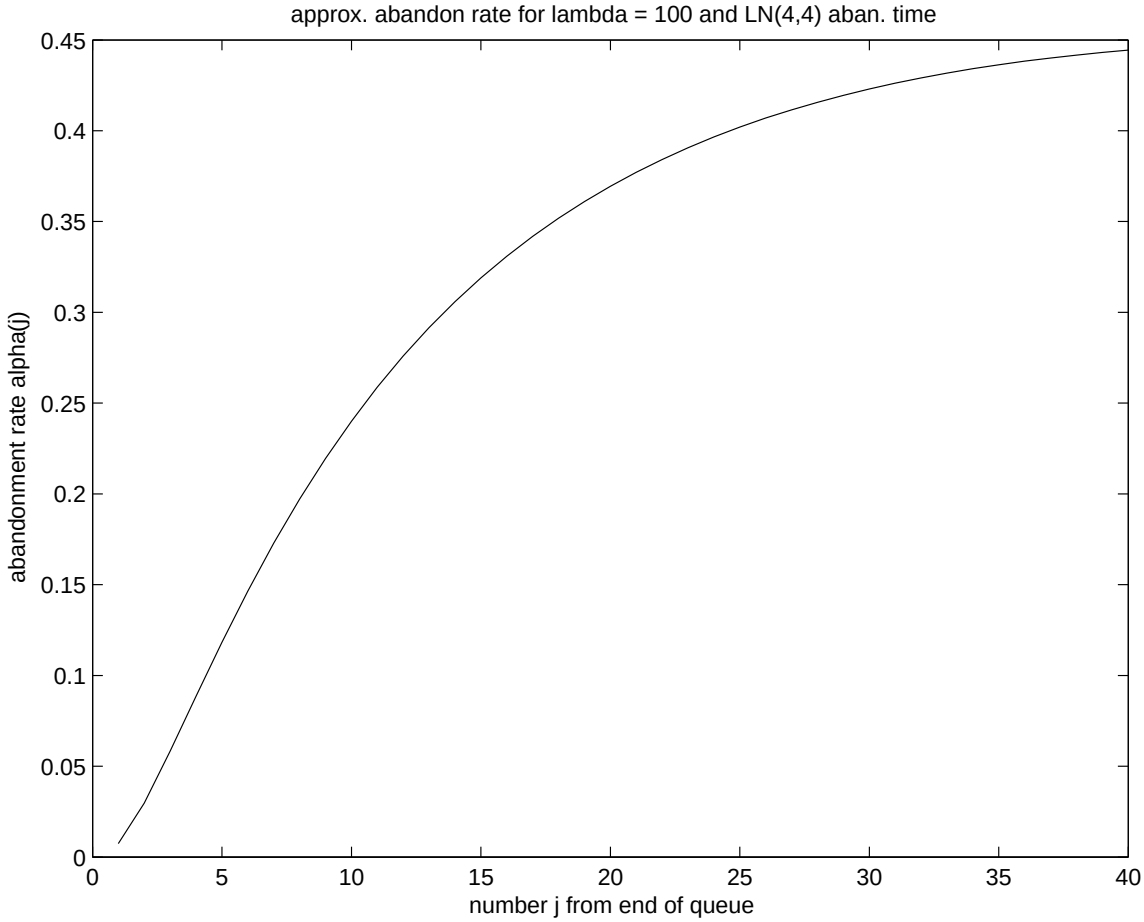

Figure 12: The approximate abandonment rate $\alpha_j = h(j/\lambda)$ for the customer j positions from the end of the queue in the $M/GI/s/r + LN(4,4)$ model with $\lambda = 100$ and lognormal $LN(4,4)$ abandon time having mean 4 and $SCV = 4$ and thus variance 64.

14 is similar to Table 3 in the main paper, but it has a more variable abandon-time distribution. Again, we see that the approximation performs well. This more-variable lognormal abandon-time distribution turns out to produce steady-state performance very close to that of an exponential abandon-time distribution with the same mean, as can be seen from Table 14, even though the lognormal hazard function is zero at the origin. We gain a better understanding when we plot the approximate abandonment rates α_j and δ_k in Figures 12 and 13. We make an explicit comparison with the exponential distribution in Figure 14. We see that δ_k is indeed quite close for these two distributions, even though the tail behavior is of course wildly different.

The next two tables, Tables 15 and 16, describe the performance of the overall algorithm, complementing the results displayed in Section 6 of the main paper. Paralleling Tables 1 and 4 in the main paper, Table 15 below describes the performance of the $M/E_2/100/200 + E_2$

	model, mean time to abandon = 4.0		
	$M/M/100/200 + LN(4,4)$		$M/M/100/200 + M$
Performance Measure	*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	–	–
$E[W\|S]$	0.1446	0.1419	0.1455
	± 0.00091	–	–
$Var(W\|S)$	0.0175	0.0169	0.0207
	± 0.00013	–	–
$E[W\|A]$	0.1878	0.1786	0.1429
	± 0.00048	–	–
$Var(W\|A)$	0.0105	0.0105	0.0137
	± 0.000048	–	–
$P(W \leq 0.1\|S)$	0.444	0.449	0.469
	± 0.0025	–	–
$P(W \leq 0.1\|A)$	0.212	0.248	0.449
	± 0.0010	–	–
$P(W \leq 0.2\|S)$	0.680	0.687	0.687
	± 0.0028	–	–
$P(W \leq 0.2\|A)$	0.602	0.632	0.737
	± 0.0023	–	–

Table 14: A comparison of approximations for steady-state performance measures with simulations in the $M/M/100/200 + LN(4,4)$ model, which has exponential service times and mean abandon time 4. The lognormal abandon-time distribution has squared coefficient of variation 4.0 and thus variance 64.0. 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 4.0. The half-width of the 95% confidence interval is given for each simulation estimate. The exact results are shown for the corresponding $M/M/100/200 + M$ model.

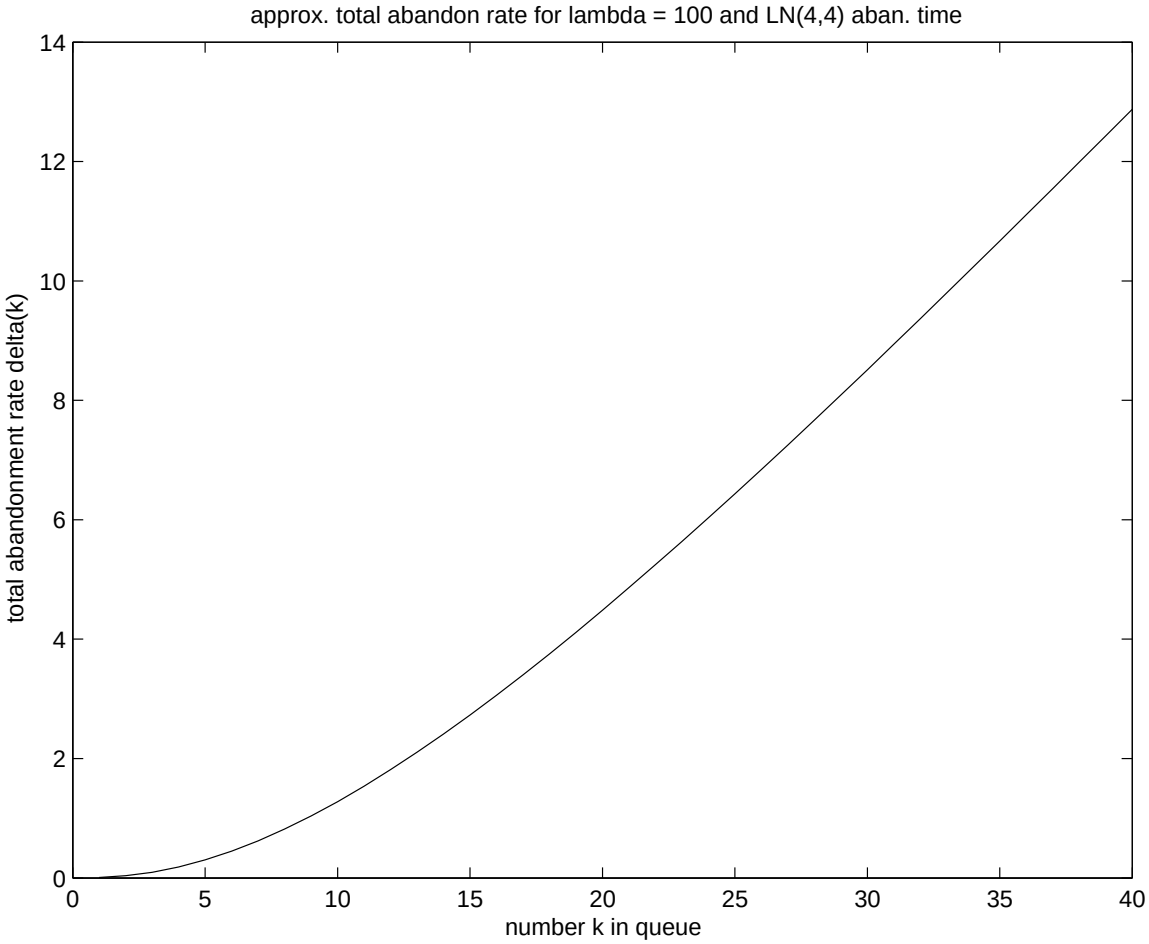

Figure 13: The approximate total abandonment rate δ_k when there are k customers in the queue in the $M/GI/s/r + LN(4,4)$ model with $\lambda = 100$ and lognormal $LN(4,4)$ abandon time having mean 4 and $SCV = 4$ and thus variance 64.

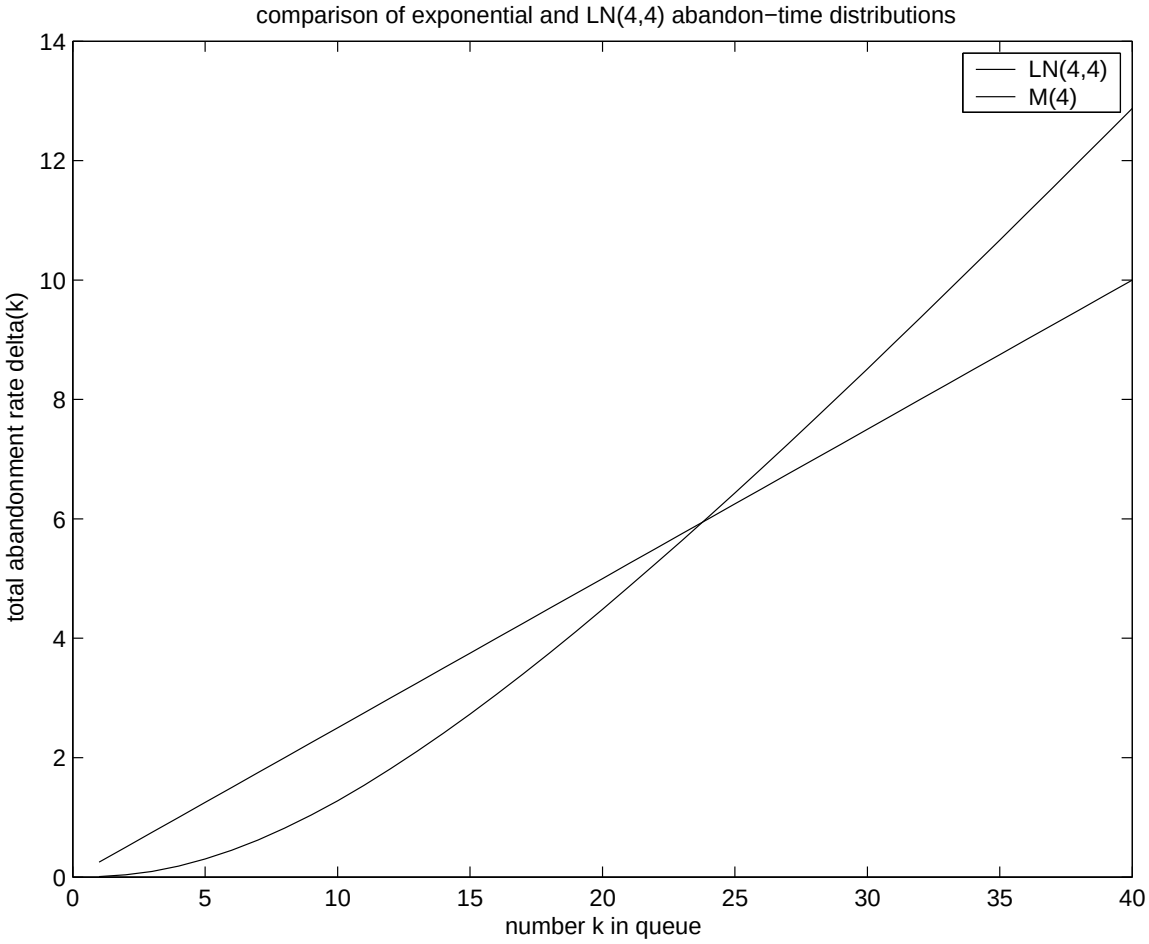

Figure 14: A comparison of two abandon-time distributions: the approximate total abandonment rate δ_k when there are k customers in the queue in the $M/GI/s/r + GI$ model with $\lambda = 100$ and two different abandon-time distributions having mean 4: lognormal $LN(4, 4)$ and exponential M(4). Both have mean 4, but $LN(4, 4)$ has $SCV = 4$ and thus variance 64.

	model, mean time to abandon = 0.25		
	$M/E_2/100/200+E_2$		$M/M/100/200+M$
Performance Measure	*sim.*	*approx. numerical*	exact numerical
$P(W=0)$	0.461	0.491	0.594
	$\pm$0.0015	–	–
$P(A)$	0.053	0.056	0.064
	$\pm$0.00029	–	–
$E[Q]$	3.07	2.84	1.62
	$\pm$0.014	–	–
$E[N]$	99.7	99.1	97.1
	$\pm$0.027	–	–
$E[W\|S]$	0.0284	0.0263	0.0148
	$\pm$0.00012	–	–
$E[W\|A]$	0.0603	0.0547	0.0321
	$\pm$0.00009	–	–
$P(W \le 0.05\|S)$	0.747	0.768	0.883
	$\pm$0.0012	–	–
$P(W \le 0.05\|A)$	0.440	0.510	0.783
	$\pm$0.0011	–	–
$P(W \le 0.1\|S)$	0.931	0.936	0.980
	$\pm$0.00048	–	–
$P(W \le 0.1\|A)$	0.865	0.886	0.972
	$\pm$0.00066	–	–

Table 15: 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 = 0.25. 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 0.25. 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.

model. Before, in Tables 1 and 4 of the main paper, the mean abandon time was 1 and 4, respectively, which is equal to the mean service time and greater than the mean service time. Now we consider the case in which the mean abandon time is less than a mean service time. Specifically, the model has arrival rate $\lambda = 102$, mean service time $\mu^{-1} = 1$, $s = 100$ servers, $r = 200$ extra waiting spaces and mean time to abandon 0.25.

Next, Table 16 describes steady-state performance in the $M/GI/100/200+M$ model with mean time to abandon = 1.0 and different service-time distributions. Paralleling Tables 5 and 6 in the main paper, the service-time distributions considered here are E_2, M and $LN(1,1)$, all with mean 1. As in Tables 5 and 6 in the main paper, we see that the performance does not depend greatly upon the service-time distribution beyond its mean.

	$M/GI/100/200+M$ *model with mean time to abandon = 1.0*			
	service-time distribution			
Perf. Meas.	E_2	M	$LN(1,1)$	*approx., exact for M service*
$P(W=0)$	0.390 ±0.0017	0.409 ±0.0013	0.397 ±0.0020	0.408 –
$P(A)$	0.0477 ±0.00027	0.0497 ±0.00020	0.0489 ±0.00025	0.0499 –
$E[Q]$	4.87 ±0.025	5.07 ±0.024	4.99 ±0.031	5.09 –
$Var(Q)$	38.3 ±0.28	44.4 ±0.30	41.4 ±0.33	44.6 –
$E[N]$	102.0 ±0.039	102.0 ±0.036	102.0 ±0.051	102.0 –
$E[W\|S]$	0.0471 ±0.00023	0.0489 ±0.00023	0.0481 ±0.00030	0.0490 –
$Var(W\|S)$	0.00354 ±0.000023	0.00418 ±0.000027	0.0039 ±0.000030	0.0042 –
$E[W\|A]$	0.0605 ±0.00023	0.0665 ±0.00021	0.0635 ±0.00021	0.0666 –
$Var(W\|A)$	0.00252 ±0.000023	0.00312 ±0.000018	0.00283 ±0.000019	0.0031 –
$P(W \le 0.1\|S)$	0.813 ±0.0012	0.799 ±0.0012	0.806 ±0.0015	0.799 –
$P(W \le 0.1\|A)$	0.805 ±0.0015	0.768 ±0.0013	0.786 ±0.0013	0.767 –
$P(W \le 0.2\|S)$	0.977 ±0.00040	0.965 ±0.00057	0.970 ±0.00053	0.964 –
$P(W \le 0.2\|A)$	0.983 ±0.00044	0.971 ±0.00054	0.977 ±0.00040	0.970 –

Table 16: A comparison of simulation estimates of steady-state performance measures in $M/GI/100/200+M$ models with three different service-time distributions having common mean 1: E_2 with $SCV=0.5$, M with $SCV=1$ and $LN(1,1)$ with $SCV=1$. 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 M 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.

9. More on Calculating Performance Measures

In this final section we supplement Section 8 of the main paper by including some additional material on calculating steady-state performance measures. First, in Subsection 9.1 we indicate a few basic performance measures we can calculate from the steady-state distribution of the number of customers in the system. Then in Subsection 9.2 we indicate how to calculate the steady-state response-time distribution and its moments. (The response time is the waiting time plus the service time.) Finally, in Subsection **??** we indicate how to calculate the waiting-time distribution for an arbitrary entering customer and its moments by combining corresponding results for those entering customers that are served and those entering customers that abandon.

9.1. Associated Rates and Steady-State Probabilities

Given the steady-state distribution of the number of customers in the system, we can calculate associated rates and steady-state probabilities (which correspond to long-run proportions): First, the steady-state probability that all servers are busy at an arbitrary time is

$$P(AllBusy) = \sum_{j=s}^{j=s+r} p_j \tag{9.1}$$

Next, the probability that the waiting room is full at an arbitrary time is

$$P(Full) = p_{s+r} \ . \tag{9.2}$$

Letting $x \wedge y$ denote $\min\{x, y\}$, the expected number of busy servers at an arbitrary time is

$$E[Busy] \equiv E[N \wedge s] = \sum_{j=1}^{j=s-1} j p_j + P(AllBusy)s \ . \tag{9.3}$$

The throughput, say θ, is the rate of service completions; it is

$$\theta = E[Busy]\mu \ . \tag{9.4}$$

9.2. The Response Time

In the paper we considered W, the steady-state waiting time until beginning service for a customer that enters the system, for the approximating $M/M/s/r + M(n)$ model. We now consider the steady-state response time for a customer that enters the system, which is the waiting time plus the service time. Let T be a response time for a customer that enters the

system. Equivalently, T is the time spent in the system. As with the waiting time, we must be careful to differentiate between customers that eventually are served and customers that eventually abandon. For customers who abandon, $T = W$, so we do not consider that case. Here we only consider customers who are served.

We first compute the expectation of T for served customers; i.e., we compute $E[T; S] = E[T1_{\{S\}}]$, where S is the event that the customer is eventually served, 1_B is the indicator function of the event B ($1_B(\omega) = 1$ if $\omega \in B$, and $1_B(\omega) = 0$ otherwise). The calculation if much like the calculation for waiting times. We exploit the fact that the service time is independent of the waiting time. We also exploit the approximations developed in Section 7.2 of the main paper. In particular, just as in the main paper, we draw heavily on the approximations in (7.10) and (7.11).

Using properties of the exponential distribution, we obtain

$$E[T; S] = \left(\sum_{k=0}^{s-1} p_k^a\right) \frac{1}{\mu} + \sum_{k=0}^{r-1} p_{s+k}^a \Gamma_{k+1} \left(\frac{1}{\mu} + \sum_{j=1}^{k+1} m_{k+1,j}\right) \tag{9.5}$$

and

$$E[T^2; S] = \left(\sum_{k=0}^{s-1} p_k^a\right) \frac{2}{\mu^2} + \sum_{k=0}^{r-1} p_{s+k}^a \Gamma_{k+1} (V_{k+1} + M_{k+1}^2) \tag{9.6}$$

where

$$V_{k+1} \equiv \frac{1}{\mu^2} \; + \; \sum_{j=1}^{k+1} m_{k+1,j}^2 \tag{9.7}$$

and

$$M_{k+1} \equiv \frac{1}{\mu} \; + \; \sum_{j=1}^{k+1} m_{k+1,j} \; . \tag{9.8}$$

Then the first and second moments of the conditional time to complete service given that service is completed are

$$E(T|S) = \frac{E[T; S]}{P(S)} \;\; \text{and} \;\; E(T^2|S) = \frac{E[T^2; S]}{P(S)} \; . \tag{9.9}$$

The conditional variance and standard deviation are then

$$Var(T|S) \equiv E(T^2|S) - (E(T|S))^2 \tag{9.10}$$

and

$$SD(T|S) \equiv \sqrt{Var(T|S)} \; . \tag{9.11}$$

We now characterize the response-time distribution via its Laplace transform. Then we can apply numerical transform inversion to calculate the distribution itself. For that purpose,

let $\hat{t}(z) \equiv E[e^{-zT}1_{\{S\}}]$ be the Laplace transform of T for served customers (Laplace-Stieltjes Transform of its cdf).

Paralleling (9.5), we have

$$\hat{t}_s(z) \equiv E[e^{-zT}1_{\{S\}}] = \left(\sum_{k=0}^{s-1} p_k^a\right)\left(\frac{\mu}{\mu+z}\right) + \sum_{k=0}^{r-1} p_{s+k}^a \Gamma_{k+1}\hat{d}_{k+1}(z) \ , \tag{9.12}$$

where

$$\hat{d}_{k+1}(z) \equiv \left(\frac{\mu}{\mu+z}\right)\Pi_{j=1}^{k+1}\left(\frac{m_{k+1,j}^{-1}}{m_{k+1,j}^{-1}+z}\right) \ . \tag{9.13}$$

We can now easily calculate the cumulative distribution function (cdf) by numerical transform inversion. We obtain the cdf $P(T \le t; S)$ for any desired t by numerically inverting its Laplace transform $\hat{t}(z)/z$, e.g., by using the Fourier-series method described in Abate and Whitt (1995). The associated conditional response-time cdf is

$$P(T \le t|S) = \frac{P(T \le t; S)}{P(S)} \ . \tag{9.14}$$

10. 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 the papers. 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. Computing* 7, 36–43.

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.
Available at http://iew3.technion.ac.il/serveng/References.references.html.

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 Spectrum* 26, 377–411.

Whitt, W. 1992. Understanding the Efficiency of Multi-Server Service Systems. *Management Science* 38, 708–723.

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

Whitt, W. 2004a. Fluid models for multi-server queues with abandonments. Department of Industrial Engineering and Operations Research, Columbia University. Available at http://columbia.edu/∼ww2040.

Whitt, W. 2004b. Two fluid approximations for multi-server queues with abandonments Department of Industrial Engineering and Operations Research, Columbia University. Available at http://columbia.edu/∼ww2040.

Whitt, W. 2005a. Engineering solution of a basic call-center model. *Management Science*, to appear. (main paper) Available at http://columbia.edu/∼ww2040.

Whitt, W. 2005b. Efficiency-driven heavy-traffic approximations for multi-server queues with abandonments. *Management Science*, to appear.
Available at http://columbia.edu/∼ww2040.