Service System Planning in the Presence of a Random Arrival Rate

Samuel G. Steckley and Shane G. Henderson
School of Operations Research and Industrial Engineering, Cornell University
Vijay Mehrotra
Decision Sciences Department, San Francisco State University

November 1, 2004

Abstract

A fundamental workforce management challenge for inbound call center managers is to determine the number of agents to be scheduled to answer calls during each time period. These decisions are typically based on the desire to minimize cost while achieving some pre-determined service objectives. These service objectives are typically functionals of the customer queue time distributions, which in turn are highly dependent on the distribution of customer arrivals.

The traditional call center modeling approach is to divide a given planning horizon into a series of time periods, and to assume a deterministic fixed-rate Poisson arrival process for each period. These arrival processes then determine the performance measures that drive the selection of staffing levels.

The arrival rate is very often not known with certainty, as we show in this paper through the analysis of historical data from several call centers. This type of uncertainty arises either because the arrival rate varies randomly over time or because the rate is simply unknown due to lack of information. In either case, the uncertainty in the arrival rate has major implications for the validity of traditional performance measures and consequently on the quality of staffing decisions.

In this paper, we consider two potential forms of uncertainty in the arrival rates, and in each case address the question of what performance measures to use in order to support staffing decisions. We also explore ways to compute appropriate estimates for these performance measures. We clarify when the analytical approximations can be expected to be accurate and describe when and how simulation should be used to provide better estimates.

1 Introduction

How should staffing levels be selected in a call center? The basic tradeoff is clear: higher staffing levels translate into both better customer service but also higher cost. Finding the "right" level involves com-

puting (service level) performance and cost for a range of potential staffing levels. Computing the cost is straightforward. Computing performance is more difficult, and is the subject of this paper. We explore the question of how to compute performance for a given staffing level, focusing on the situation when the arrival rate of calls to the call center cannot be determined with certainty. More specifically we examine the questions of *what* to compute, *how* to compute it, and what are the probable implications of ignoring uncertainty associated with the arrival process.

The term "uncertainty" has several possible interpretations, and as we argue in this paper, the interpretation can be important in selecting an appropriate performance measure to gauge performance.

One interpretation, and the one that we focus on, is as follows. On any given day (the choice of days as a time scale is arbitrary but seems appropriate) it is quite reasonable to model the arrival process of calls as a nonhomogeneous Poisson process (NHPP) with time-dependent rate function $\Lambda = (\Lambda(t) : t \geq 0)$. This follows from the Palm-Khintchine theorem (e.g., Whitt [2002b, p. 318]) that states that the superposition of arrivals from a large number of independent potential customers is well approximated by a Poisson process. The rate function depends on the propensity of customers to call, which in turn can depend on factors that cannot be planned for in advance, such as weather, marketing promotions, and competitor behavior. This situation can be modelled by viewing the rate function Λ as random. Once the rate function is realized (e.g., after the weather, marketing promotions, or behavior of competitors is revealed), the arrival rate function is then fixed, and call arrivals follow a NHPP with the realized rate function. We call this interpretation the *randomly varying arrival rate* (RVAR) case.

A second interpretation relates to forecast error. In this setting we believe that there is a true deterministic arrival rate function $\lambda = (\lambda(t) : t \geq 0)$ but we do not know what it is. Unlike the RVAR case, the arrival rate function does not vary from day to day. The uncertainty here arises due to our lack of perfect knowledge of λ, for example, the response to a one-time marketing campaign. If we model our uncertainty through a random function $\Lambda = (\Lambda(t) : t \geq 0)$ then we again have a random arrival rate, but the interpretation is quite different to the RVAR case. We call this case the *unknown arrival rate* (UAR) case.

A hybrid situation where the *distribution* of the arrival rate in the RVAR case is unknown is also possible, if not typical, but while it may be the "correct" abstraction it also seems unwieldy. We do not consider that possibility further.

The appropriate long-run performance measures differ in the RVAR and UAR cases in terms of how one should weight performance conditional on a given realized arrival rate function. In the RVAR case there are more customers expected on days when the arrival rate Λ is large, so more customers experience the performance associated with a large arrival rate. In the UAR case weighting by the arrival rate may

be inappropriate. These long-run performance measures can be viewed as "customer-focussed" since they indicate what a customer can expect in terms of performance.

In addition to long-run performance measures, we also discuss methods for determining short-run performance measures, i.e., "what might happen tomorrow." This kind of information is valuable because it can help to explain variability in daily performance. Short-run performance measures can be viewed as "manager-focussed" since they indicate what a manager could see on any particular day. But of course, long-run and short-run performance measures are relevant to both managers and customers.

Given that we can choose appropriate performance measures, we then look at how to compute them. A common approach is to use closed-form expressions based on steady-state results for simple queueing models. We explore, in some depth, the question of when those expressions are accurate. Our efforts in this direction complement those of Green and Kolesar [1991], Whitt [1991], Massey and Whitt [1998] and Green et al. [2001]. Those papers give empirical and theoretical evidence that the approximation is good for moderate to large event rates, because it is in that setting that the system quickly reaches steady state.

We explore the quality of the approximation in more detail by computing, for some simple models, approximations for the difference between true and expected performance. Those calculations shed further light on when one can feel confident in using closed-form approximations based on simple queueing models. For the most appropriate queueing model that explicitly models customer abandonment the approximations are very good when the arrival rate does not change too rapidly, or when the call center is lightly loaded. When the approximations are not of high quality, we might instead use simulation. We briefly discuss how to design the simulation experiments in order to efficiently compute the desired performance measures.

Grassmann [1988] modelled forecast errors using a random arrival rate. Thompson [1999] and Jongbloed and Koole [2001] gave methods for staffing when the arrival rate is random. Whitt [1999] suggested a particular form of random arrival rate for capturing forecast uncertainty. Chen and Henderson [2001] studied the potential impact in predictions of ignoring the issue. Ross [2001, Chapter 4] developed extensions to the "square-root staffing rule" to account for a random arrival rate. Avramidis et al. [2004] developed several different arrival process models and compared their fit to call center data. They also found that performance measures depend fairly strongly on the arrival rate process. Deslauriers et al. [2004] show that it is appropriate in their setting to weight performance by the arrival rate. Gans et al. [2003] discuss this issue as part of a survey of the area of call center design and management. Brown et al. [2002] developed an autoregressive model for the arrival rate that can capture correlation across different days. Harrison and Zeevi [2005] developed an economic model based on attaching costs to abandonment and agent levels. Mathematical support for their model is given in Bassamboo et al. [2004]. Whitt [2004] gives an economic analysis for a special case of the Harrison-Zeevi model, offering 2 computational approaches

for estimating performance. Both the Harrison-Zeevi and Whitt papers address the RVAR case. We do not adopt an economic model here, instead working directly with performance measures associated with the waiting time distribution of a "typical" customer. In addition to a random arrival rate, Whitt [2004] deals explicitly with absenteeism, which he models through a random number of servers being available. We do not consider a random number of servers, although it is possible to capture that phenomenon in a straightforward manner in the RVAR case.

We view the main contributions of this paper as follows:

1. We give further statistical evidence from call centers that the RVAR case is a common feature. We show that the random arrival rate is not only *statistically* significant but also *practically* significant, reinforcing previous observations to that effect.
2. We distinguish two forms of uncertainty, and argue that we should use different performance measures in the two cases.
3. We briefly describe the potential impact of ignoring a random arrival rate in performance predictions.
4. We look at the use of both closed-form approximations based on simple queueing models and simulation for computing performance measures. In particular, we give further insight into when closed-form approximations may be expected to perform well, and when one should instead consider simulation. This analysis applies very generally, and not just to the RVAR or UAR situations.

The remainder of this paper is organized as follows. In §2 we analyze data from several call centers, showing in several cases that the arrival process is not well-modelled by a NHPP. Then, in §3 we consider the RVAR case and the performance measure giving the long-run fraction of customers that wait less than a prescribed amount of time in queue before receiving service. We give an expression for this quantity, and then consider approximations given by steady-state expectations. We also show that performance will typically be overestimated if a randomly-varying arrival rate is ignored. The section concludes by discussing how one can use simulation to estimate performance measures efficiently. In §4 we turn to the UAR case and again suggest appropriate performance measures. We again consider approximations based on steady-state expectations. The section concludes with a discussion of simulation procedures to estimate the performance measures. The question of whether one needs to perform simulation or not is an important one. This decision may depend on the quality of steady-state approximations. In §5 we explore this notion in more depth. We use results for diffusion approximations and birth-death systems to shed further light on when steady-state approximations can be expected to accurately represent the time-dependent performance that one actually sees. We offer some conclusions and directions for future research in §6.

2 Data analysis

The conventional approach to call center staffing is the "Stationary, Independent Period-by-Period," (SIPP) approach [Green et al., 2001]. The SIPP approach divides the planning horizon into a series of time intervals. Within each time interval a stationary queueing model is analyzed to provide estimates of performance in that period. The arrival processes in the periods are usually modelled as independent Poisson Processes, with the arrival rate for each period assumed to be fixed throughout that period. Agent requirements for each period are then determined from steady state equations that are based on the forecasted arrival and service rates, and target service objective for that period. As often noted (for example, in Brown et al. [2002], Green et al. [2001], and Avramidis et al. [2004]), there are a number of potentially significant problems associated with the standard SIPP approach, most notably the use of a period-specific arrival rate that does not vary over the planning horizon.

We obtained data from several call centers and first "cleaned" it by removing all records for weeks containing unusual events like holidays and all records for days in which known data collection problems existed. We also ensured that there were no trends in the data, since our ensuing analysis cannot distinguish these from the effects we are trying to capture. From here, we conducted extensive data analysis to examine, among other things, how well the Poisson assumption stands up to representative industry data about call arrival patterns. We provide a few illustrative results of this data analysis to motivate the analysis that follows.

Our observations reveal two common phenomena. First, for the vast majority of queues and time periods, historical data does not support the assumption of Poisson arrivals following a deterministic arrival rate λ_i during period i over the entire planning horizon. We illustrate this at the weekly, daily, and hourly levels and also conduct a statistical test of the fixed Poisson arrival rate hypothesis at the hourly level.

Second, we note that it is common to see significant correlation in call arrivals across different time periods in the same planning horizon. We briefly discuss this empirical phenomenon to motivate more general arrival rate models such as those suggested by Avramidis et al. [2004] and the one used in §3.

Whitt [2002a] describes the variability associated with Poisson arrivals to a call center as "process variability." Even under the deterministic arrival rate assumption of the standard SIPP model, we expect to see variability in the total call arrivals per week, per day, and per 15-, 30-, and 60-minute period, with the variance in the number of arrivals approximately equal to the mean. However, we have consistently observed much higher variability than we would expect under a Poisson model with stable and known parameters and ascribe a significant amount of this additional variability to the randomness of the arrival rate parameter, referred to by Whitt as "parameter variability."

For example, consider the daily and weekly historical data for a typical queue "QT" presented in Table 1 below.

Day Of Week	Mean	StDev	Var/Mean	Observations
Monday	2072	496	119	23
Tuesday	1952	483	119	23
Wednesday	1971	521	138	23
Thursday	2185	960	422	23
Friday	1990	747	280	23
TOTAL	10,170	2400	566	23

Table 1: Summary statistics for daily and weekly volumes for QT .

There is one especially interesting point to note in Table 1. It is well known that most call centers experience highest call volumes on Mondays, with daily totals typically declining over the course of the week. In this case, we see a mean for Thursday that is slightly higher than Monday, but with a much higher level of variability.

Based on the SIPP model, required staffing is based solely on the historical mean call volume, ignoring any excess variance. The result: this historical data would lead to staffing levels which overlook the high variability of Thursday, introducing a definite risk to operational performance.

We believe that this is a serious, and common, practical problem for managers, and thus this research is intended to provide guidance about staffing decisions while taking into account the relative variability of different time periods.

Next, we examine the behavior of call arrivals at the hourly level. In particular, to test the standard SIPP hypothesis that each one hour interval (e.g., Monday 10-11am) faces a Poisson arrival stream with a deterministic parameter that is the same in all weeks, we use a statistical test presented in Brown and Zhao [2002]. In describing the test, we refer to our planning horizons as "weeks" though clearly this applies to longer or shorter horizons without loss of generality.

In examining the data, we seek to test the hypothesis that each interval i faces a fixed arrival rate λ_i against the alternative hypothesis that each interval i in each week $t = 1, \ldots, n$ has a (potentially distinct) arrival rate λ_{it} such that $\sum_{t=1}^{n}(\lambda_{it} - \lambda_i)^2 > 0$.

Letting $X_1, X_2, \ldots X_n$ correspond to call volumes for a particular time interval during historical weeks $1, 2, \ldots n$, we define $Y_1, Y_2, \ldots Y_n$, where $Y_i = \sqrt{X_i + 3/8}$. Letting $\bar{Y} = n^{-1}\sum_{i=1}^{n} Y_i$, we then define our test statistic $T \equiv 4\sum_{i=1}^{n}(Y_i - \bar{Y})^2$.

As described in greater detail in Brown and Zhao [2002], we reject H_0 if $T > \chi^2_{n-1,1-\delta}$ where δ is our level of significance (we have set $\delta = 0.01$ in Table 2).

Our statistical tests, for QT and for the vast majority of queue-interval combinations in our data set, are overwhelmingly in favor of rejecting the hypothesis that the arrival rate does not depend on t. These

Hour Of Day	Mean	StDev	T	Critical Value	n
9 - 10	14	9	151	44	26
10 - 11	84	49	844	47	28
11 - 12	199	95	1113	47	28
12 - 13	243	123	1385	47	28
13 - 14	258	132	1458	46	27
14 - 15	228	183	2469	47	28
15 - 16	207	180	2057	47	28
16 - 17	195	183	2402	47	28
17 - 18	201	162	1898	47	28
18 - 19	187	91	971	47	28
19 - 20	185	76	823	47	28
20 - 21	166	76	949	47	28
21 - 22	136	75	876	46	27
22 - 23	91	45	556	46	27
23 - 24	68	35	467	47	28

Table 2: Hourly call volumes for QT on Mondays

results are consistent with §4.2 of Brown et al. [2002].

Next, we briefly present empirical evidence of correlation between time periods within a planning horizon. In Table 3 below, we show correlations between Monday call volumes and call volumes on other days of the week for several different queues from our dataset.

Queue Number	M-T Corr	M-W Corr	M-TH Corr	M-F Corr	n
1	0.44	0.41	0.38	0.19	20
2	0.88	0.76	0.58	0.60	26
3	0.92	0.69	0.37	0.04	30
4	0.81	0.67	0.56	0.70	24
5	0.42	0.49	0.48	0.51	27
6	0.89	0.78	0.56	0.55	26
7	0.81	0.79	0.70	0.66	16

Table 3: Correlation between Monday call volume and subsequent days for QT

The presence of non-zero positive correlation from one day to the next has been noted in Brown et al. [2002], and our findings confirmed this strongly, especially in the case of Monday. However, we were somewhat surprised to find multiple cases with strong correlations between Monday call volumes and Wednesday's, Thursday's and Friday's call volumes as well.

Finally, we look at the question of intra-day correlation. It is our belief that the first few hours of a day often provide significant information about the call volumes for the remainder of the day. To illustrate this, in Table 3, the values in the cells correspond to the correlation between the first three hours of each day with the subsequent hours of the respective day (the correlation estimates are based on 27-29 data

Interval	Monday	Tuesday	Wednesday	Thursday	Friday
12 - 13	0.81	0.91	0.45	0.99	0.79
13 - 14	0.50	0.77	0.46	0.90	0.69
14 - 15	0.32	0.77	0.35	0.69	0.61
15 - 16	0.21	0.85	0.37	0.32	0.65
16 - 17	0.22	0.78	0.33	0.29	0.60
17 - 18	0.27	0.88	0.40	0.67	0.88
18 - 19	0.54	0.94	0.46	0.69	0.87
19 - 20	0.69	0.88	0.45	0.82	0.71
20 - 21	0.61	0.73	0.35	0.81	0.64
21 - 22	0.71	0.74	0.34	0.80	0.50
22 - 23	0.56	0.71	0.29	0.75	0.60
23 - 24	0.35	0.59	0.21	0.57	0.77

Table 4: Correlation between first three hours with subsequent hours for QT

points for each Day-Interval combination).

These empirical results suggest strongly that the later hours of a day are not independent of the early hours and reinforce the findings of Avramidis et al. [2004]. Indeed, the findings on inter- and intra-day correlation, as illustrated here in Table 3 and Table 4 underscore the dangers associated with the independence assumption that is so central to SIPP.

Motivated by the above empirical results, we will use a specific model of call arrivals originally proposed in Whitt [1999] to illustrate our ideas. In this model, the arrival process is Poisson with arrival rate function $B(\lambda(s) : s \geq 0)$, where $(\lambda(s) : s \geq 0)$ is a "profile" describing the relative intensities of arrivals, and B is a random "busyness" parameter indicating how busy the day is. To simplify the analysis we assume that $\lambda(\cdot)$ is constant within each period.

While we utilize this model for its illustrative value, it is important to understand that our results are actually quite general. In particular, these results extend naturally to many more general arrival process models, including the autoregressive model given in Brown et al. [2002], the Poisson mixture model presented in (Jongbloed and Koole [2001]), and the recent models proposed in (Avramidis et al. [2004]).

3 Randomly Varying Arrival Rates

The key long-run performance measure is the long-run fraction of customers that receive satisfactory service in a given period. A customer receives satisfactory service if her delay in queue is at most τ seconds. Common choices for τ are 20 seconds (a moderate delay) and 0 seconds (no delay). For much of what follows we focus on a single period (e.g., 10am - 10.15am) in the day, arbitrarily representing this time period as time 0 through time t. With an abuse of notation, let Λ_i denote the real-valued random

arrival rate within this period on day i.

Let S_i denote the number of satisfactory calls (calls that are answered within the time limit τ) in the period on day i out of a total of N_i calls that are received. Notice that here we consider any call that abandons to be unsatisfactory. Some planners prefer to ignore calls that abandon within very short time frames. There is a difference, but it is not important for our discussion.

Over n days, the fraction of satisfactory calls is

$$\frac{\sum_{i=1}^n S_i}{\sum_{i=1}^n N_i}.$$

Assume that days are i.i.d., the staffing level is fixed throughout, and $\mathrm{E}N_1 < \infty$. (Assuming days are i.i.d. ignores the inter-day correlations seen in §2. More general dependence structures can be captured in essentially the same framework.) The last assumption holds if $\mathrm{E}\Lambda_1 < \infty$. Dividing both the numerator and denominator by n and taking the limit as $n \to \infty$, the strong law then implies that the long-run fraction of satisfactory calls is

$$\frac{\mathrm{E}S_1}{\mathrm{E}N_1}. \tag{1}$$

This ratio gives performance as a function of staffing level. But how do we compute it?

First note that

$$\begin{aligned} \mathrm{E}N_1 &= \mathrm{E}\mathrm{E}[N_1|\Lambda_1] \\ &= \mathrm{E}[\Lambda_1 t] \\ &= t\mathrm{E}\Lambda_1, \end{aligned} \tag{2}$$

so that $\mathrm{E}N_1$ is easily computed. Computing $\mathrm{E}S_1$ is more difficult. We again condition on Λ_1 to obtain $\mathrm{E}S_1 = \mathrm{E}s(\Lambda)$, where $s(\lambda)$ is the conditional expected number of satisfactory calls in the period, conditional on $\Lambda_1 = \lambda$. Our initial goal is an expression for $s(\lambda)$.

Fix the arrival rate to be deterministic and equal to λ (for now). Let $X(\cdot;\lambda) = (X(s;\lambda) : s \geq 0)$ be a Markov process used to model the call center when there is a fixed arrival rate λ. In specialized cases one can take X to be the process giving the number of customers in the system, but it may be more complicated. Suppose that a customer arriving at time s will receive satisfactory service if and only if $X(s;\lambda) \in B$ for some distinguished set of states B.

Example 1 *A common model of a call center is an $M/M/c+M$ queue, i.e., the Erlang-A model. There are c servers, service times are exponentially distributed, and the arrival process is Poisson. Customers are willing to wait an exponentially-distributed amount of time (the "patience time") in the queue, and*

abandon if they do not reach a server by that time. Here we take $X(s;\lambda)$ to be the number of customers in the system at time s. Then X is a continuous-time Markov chain (CTMC). Suppose that a service is considered satisfactory if and only if the customer immediately reaches a server. Then we can take $B = \{0, 1, 2, ..., c-1\}$, i.e., a service is satisfactory if and only if the number of customers in the system is $c-1$ or less when the customer arrives.

Example 2 *Consider the same model as in the previous example, but now define a service to be satisfactory if and only if the customer reaches a server in at most $\tau > 0$ seconds so long as she doesn't abandon. The state space of the CTMC defined in the previous example is no longer rich enough to determine, upon a customer arrival, whether that customer will receive satisfactory service or not. We turn to a different Markov process in such a case. Without loss of generality, suppose that as soon as a customer arrives, the patience and service times for that customer are sampled and therefore known. Since customers are served in FIFO order we can determine, for every customer that has arrived by time s, whether that customer will abandon or not, and if not which agent the customer will be served by. Let $V_i(s;\lambda)$ denote the "work in process" for agent i at time s, $i = 1, \ldots, c$. The quantity $V_i(s;\lambda)$ gives the time required for agent i to complete the service of all customers in the system at time s that are, or will be, served by agent i. Let $X(s;\lambda)$ be the vector $(V_i(s;\lambda) : 1 \leq i \leq c)$. The process $X(\cdot;\lambda) = (X(s;\lambda) : s \geq 0)$ is a Markov process, albeit a rather complicated one, and we can take $B = \{v : \min_{i=1}^{c} v_i \leq \tau\}$, so that a service is satisfactory if and only if at least one server will be available to answer a call within τ seconds of a customer's arrival.*

Let $P_\varphi(\cdot)$ denote the probability measure when the Markov process has initial distribution φ. Let ν and π be, respectively, the distribution of the Markov process at time 0 and the stationary distribution (assumed to exist and be unique). Proposition 1 serves as a foundation for the use of steady-state approximations for performance measures in both the deterministic and random arrival rate contexts.

Proposition 1 *Under the conditions above,*

$$s(\lambda) = \lambda \int_0^t P_\nu(X(s;\lambda) \in B)\, ds.$$

If $\nu = \pi$, so that the Markov process is in steady-state at time 0, then

$$s(\lambda) = \lambda t f(\lambda),$$

where $f(\lambda) = P_\pi(X(0;\lambda) \in B)$ is the steady-state probability that the system is in state B. We can interpret $f(\lambda)$ as the long-run fraction of customers that receive satisfactory service.

Proof: For notational simplicity we suppress the dependence on λ. For $s \geq 0$, let $U(s) = I(X(s) \in B)$,

where $I(\cdot)$ is the indicator function that is 1 if its argument is true and 0 otherwise. Note that X can be defined such that U is left continuous and has right hand limits. Let $L = (L(s) : s \geq 0)$ be the arrival process. Then L is a Poisson process with rate λ. For arbitrary $v \geq 0$, $(L(v+u) - L(v) : u \geq 0)$ is independent of $(U(s) : 0 \leq s \leq v)$ and $(L(s) : 0 \leq s \leq v)$. Then $s(\lambda) = \lambda \mathrm{E}_\nu \int_0^t U(s)\,\mathrm{d}s$ by the PASTA result (e.g., Wolff [1989, §5.16]). By Fubini's theorem, for arbitrary $v \geq 0$, $\mathrm{E}_\nu \int_0^v U(s)\,\mathrm{d}s = \int_0^v \mathrm{E}_\nu U(s)\,\mathrm{d}s$. Therefore

$$\mathrm{E}_\nu \int_0^v U(s)\,ds = \int_0^v P_\nu(X(s) \in B)\,ds. \tag{3}$$

Taking $v = t$, it follows that $s(\lambda) = \lambda \int_0^t P_\nu(X(s) \in B)\,ds$.

For the second result the system is in steady state at time 0 so that $\nu = \pi$. But $P_\pi(X(s) \in B) = P_\pi(X(0) \in B)$ for all $s \geq 0$. Defining $f(\lambda) = P_\pi(X(0) \in B)$, it follows from (3) that

$$\mathrm{E}_\pi \int_0^v U(s)\,ds = v f(\lambda), \tag{4}$$

and so $s(\lambda) = \lambda t f(\lambda)$.

To see that $f(\lambda)$ can be interpreted as the long-run fraction of customers that receive satisfactory service, define the stochastic process $A = (A(s) : s \geq 0)$, where $A(s) = \int_0^s U(u)\,dL(u)$. Then the fraction of customers that have received satisfactory service up to time v is given by $A(v)/L(v)$. It is assumed that as $v \to \infty$, $A(v)/L(v)$ converges to some constant p, where p is the long-run fraction of customers that receive satisfactory service. We show that $f(\lambda) = p$. From the PASTA result (e.g., Wolff [1989, §5.16]), since $A(v)/L(v)$ converges to p, $\int_0^v U(s)\,ds/v$ also converges to p as $v \to \infty$. But $p = \mathrm{E}_\nu p = \mathrm{E}_\nu \lim_{v\to\infty}(1/v)\int_0^v U(s)\,ds$. By the bounded convergence theorem, $\mathrm{E}_\nu \lim_{v\to\infty}(1/v)\int_0^v U(s)\,ds = \lim_{v\to\infty}(1/v)\mathrm{E}_\nu \int_0^v U(s)\,ds$. By (4), $\lim_{v\to\infty}(1/v)\mathrm{E}_\nu \int_0^v U(s)\,ds = f(\lambda)$. Therefore $f(\lambda) = p$. □

3.1 Steady-state approximations

Suppose that we adopt the steady-state approximation $s(\lambda) \approx \lambda t f(\lambda)$. Here λt is the expected number of customer arrivals in the period and $f(\lambda)$ is the long-run fraction of customers that receive satisfactory service. From (1) and (2), we see that

$$\frac{\mathrm{E}S_1}{\mathrm{E}N_1} = \frac{\mathrm{E}s(\Lambda_1)}{t\mathrm{E}\Lambda_1} \approx \frac{\mathrm{E}[\Lambda_1 f(\Lambda_1)]}{\mathrm{E}\Lambda_1}. \tag{5}$$

The fact that one should weight $f(\Lambda)$ by the arrival rate in (5) is well known. It is implicit (and at times explicit) in the work of Harrison and Zeevi [2005] and Whitt [2004] for example. Chen and Henderson

[2001] did *not* perform this weighting in their analysis. So their results do not directly apply to the RVAR case, in contrast to what is claimed there. (But their results may apply in the UAR case considered in §4.)

What are the consequences of ignoring a randomly-varying arrival rate when predicting performance in a call center? In that case we would first estimate a deterministic arrival rate. The most commonly used estimates converge to $\mathrm{E}\Lambda_1$ as the data size increases. We then estimate performance as $f(\mathrm{E}\Lambda_1)$.

Together with (5), Proposition 2 below establishes that if f is decreasing and concave over the range of Λ_1, then we will overestimate performance if a random arrival rate is ignored. The function f is, in great generality, decreasing in λ. For many models it is also concave, at least in the region of interest; see Chen and Henderson [2001].

Proposition 2 *Suppose that f is decreasing and concave on the range of Λ_1. Then*

$$\frac{\mathrm{E}[\Lambda_1 f(\Lambda_1)]}{\mathrm{E}\Lambda_1} \leq f(\mathrm{E}\Lambda_1).$$

Proof: We have that

$$\begin{aligned} \mathrm{E}[\Lambda_1 f(\Lambda_1)] &\leq (\mathrm{E}\Lambda_1)(\mathrm{E}f(\Lambda_1)) && (6) \\ &\leq (\mathrm{E}\Lambda_1) f(\mathrm{E}\Lambda_1) && (7) \end{aligned}$$

establishing the result. The inequality (6) follows since f is decreasing (see, e.g., Whitt [1976]), and (7) uses Jensen's inequality. □

For certain models and distributions of Λ_1, we may be able to compute (5) exactly. In general though, this will not be possible. In such a case we can use some numerical integration technique. The problem is quite straightforward since f is typically easily computed and the integral $\mathrm{E}[\Lambda_1 f(\Lambda_1)]$ is one-dimensional.

We now turn from long-run performance to short-run performance. We want to determine the distribution of S_1/N_1, the fraction of satisfactory calls in a single period $[0, t]$ of a single day. There is a positive probability that $N_1 = 0$, but it is vanishingly small for most call centers. In any case, we can just define $0/0 = 1$ arbitrarily to ensure that S_1/N_1 is a proper random variable. Our approach is once again to condition on the arrival rate Λ_1 in the period. We reason heuristically (non-rigorously) as follows.

Let $(X(s;\lambda) : s \geq 0)$ denote the underlying Markov process conditioned on $\Lambda_1 = \lambda$. Let $T_i(\lambda)$ denote the time of the ith customer arrival when the arrival rate is λ. Define $Z_i(\lambda) = X(T_i(\lambda);\lambda)$ to be the state of the Markov process at the time of the ith customer arrival. The ith customer receives satisfactory service

if and only if $Z_i(\lambda) \in B$. So conditional on $\Lambda_1 = \lambda$, S_1/N_1 has the same distribution as

$$\frac{1}{N(t;\lambda)} \sum_{i=1}^{N(t;\lambda)} I(Z_i(\lambda) \in B),$$

where $N(s;\lambda)$ is a Poisson random variable with mean λs giving the number of arrivals in $[0, s]$.

The strong Markov property for $X(\cdot;\lambda)$ ensures that $(Z_i(\lambda) : i \geq 1)$ is a Markov chain. We can then apply a central limit theorem (e.g., Meyn and Tweedie [1993, Chapter 17]) to assert that under appropriate conditions

$$\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n} I(Z_i(\lambda) \in B) - f(\lambda)\right] \Rightarrow \sigma(\lambda)\mathcal{N}(0,1),$$

as $n \to \infty$, where $\Rightarrow$ denotes convergence in distribution, $\mathcal{N}(a, b)$ is a normal random variable with mean a and variance b, and $\sigma^2(\lambda)$ is a variance constant. Again under appropriate conditions a random-time-change theorem ensures that

$$N^{1/2}(s;\lambda)\left[\frac{1}{N(s;\lambda)} \sum_{i=1}^{N(s;\lambda)} I(Z_i(\lambda) \in B) - f(\lambda)\right] \Rightarrow \sigma(\lambda)\mathcal{N}(0,1)$$

as $s \to \infty$. A converging-together argument then ensures that

$$(\lambda s)^{1/2}\left[\frac{1}{N(s;\lambda)} \sum_{i=1}^{N(s;\lambda)} I(Z_i(\lambda) \in B) - f(\lambda)\right] \Rightarrow \sigma(\lambda)\mathcal{N}(0,1). \tag{8}$$

The limit result (8) then ensures that, so long as t is "large enough", conditional on $\Lambda_1 = \lambda$,

$$\frac{S_1}{N_1} \stackrel{\mathcal{D}}{=} \frac{1}{N(t;\lambda)} \sum_{i=1}^{N(t;\lambda)} I(Z_i(\lambda) \in B) \stackrel{\mathcal{D}}{\approx} \mathcal{N}\left(f(\lambda), \frac{\sigma^2(\lambda)}{\lambda t}\right),$$

where $\stackrel{\mathcal{D}}{=}$ and $\stackrel{\mathcal{D}}{\approx}$ denote equality and approximate equality in distribution respectively. Unconditioning, we then assert that

$$\frac{S_1}{N_1} \stackrel{\mathcal{D}}{\approx} \mathcal{N}\left(f(\Lambda), \frac{\sigma^2(\Lambda)}{\Lambda t}\right), \tag{9}$$

so that the realized fraction of acceptable calls is approximately a mixture of normal random variables.

Computing or approximating $\sigma^2(\lambda)$ even for simple models of call centers is a challenging problem that we will address elsewhere. One might use simulation to estimate it, but for now we adopt the "zeroth order" assumption that $\sigma^2(\lambda)/(\lambda t) \approx 0$ so that

$$\frac{S_1}{N_1} \approx f(\Lambda).$$

In other words, the strong law provides a reasonable approximation for performance in the period. We do not address the question of whether taking the variance to be 0 is a reasonable approximation because that is a somewhat involved question. The random variable $f(\Lambda)$ is quite tractable. It represents a simple transformation of the distribution of Λ, and if f and the distribution of Λ are available then its distribution function and/or density can be readily computed through standard results for transformations of random variables.

3.2 Simulation-based estimates

The approximations for long-run and short-run performance described above may be inappropriate, either because the steady-state approximations for time-dependent quantities may be inaccurate for a non-negligible set of arrival rates, or because the true system is not well modelled by simple models for which steady-state results are readily computed. It is natural to then turn to simulation to compute performance measures.

In terms of long-run performance, we have already noted that the problem reduces to computing $\mathrm{E}S_1$, the expected number of satisfactory calls in a particular period. This is straightforward using simulation. One can simply generate the arrival rate process, Λ say, and then conditional on the realized value, simulate the call center for the day, giving a realization of S_1. Repeating this process in i.i.d. fashion gives $S_1, \ldots, S_n$ say, which can be averaged to give an estimate of $\mathrm{E}S_1$. But we can develop more efficient (in the sense of lower variance) estimators of $\mathrm{E}S_1$ by taking advantage of structure.

For definiteness, suppose we adopt the model that the arrival rate is given by $B(\lambda(s) : s \geq 0)$, where $\lambda(\cdot)$ is constant in each period, and B is a random "busyness" factor. If we know $\mathrm{E}B$ then we can use $B - \mathrm{E}B$ as a control variate, i.e., we use

$$\frac{1}{n}\sum_{i=1}^{n}(S_i - \beta(B_i - \mathrm{E}B))$$

to estimate $\mathrm{E}S_1$, where $((S_i, B_i) : i = 1, \ldots, n)$ are i.i.d. and distributed as (S_1, B), and β is a constant that is chosen to maximize the variance reduction; see, e.g., Law and Kelton [2000]. However, we will typically know much more than just the mean of B. If we know its distribution then, as discussed in Glasserman [2004, p. 220], stratifying on B should yield larger variance reduction than using $B - EB$ as a control variate. See Glasserman [2004, §4.3] for details on how to implement stratification.

For short-run performance we wish to compute the distribution of S_1/N_1. This random variable does not have a (Lebesgue) density since it is supported on the rationals. Its probability mass function is also uninformative. Therefore, we would probably estimate a moderately coarse histogram (say, with bins of

width $\Delta x = 0.01$). The height of the bin $[x, x + \Delta x]$ is proportional to $F(x + \Delta x) - F(x)$, where F is the distribution function of S_1/N_1. Hence, estimating this histogram is equivalent to estimating the distribution function at the fixed set of points $\Delta x, 2\Delta x, \ldots, 1$. This estimation is straightforward based on i.i.d. observations (S_1, N_1), and one can apply standard results (e.g., Ross [1996, pp. 360–363]) to compute tolerance bounds for F. As with estimating $\mathrm{E}S_1$, one can stratify on the busyness parameter B to reduce variance in the estimation of the quantities $F(x) = P(S_1 - xN_1 \leq 0)$.

4 Uncertain Arrival Rates

Suppose that the arrival rate function does not vary from day to day and is given by the fixed function $(\lambda(s) : s \geq 0)$ say, but we do not know this function with certainty. This situation can arise, for example, when a call center is opening for the first time, when a new product is added to an existing portfolio of products, or when a new marketing promotion comes into effect. It corresponds to what we usually interpret as "forecast uncertainty," and commonly arises in dynamic business environments.

Just as in the RVAR case, in the long run we are interested in $\mathrm{E}S_1/\mathrm{E}N_1$, the long-run fraction of satisfactory calls in a given period, and in the short run we are interested in the distribution of S_1/N_1, the fraction of satisfactory calls in a single period in a single day. In the long-run we will eventually learn the true arrival rate through observation, but decisions need to be made before that eventuates, which helps to explain our interest in this case.

We focus on a single period $[0, t]$ of the day and assume that the true arrival rate takes on the constant value λ^* in this period. Let Λ denote a random variable representing our knowledge of the value λ^*. In the RVAR case we obtained a new sample from the distribution of Λ every day. In contrast, in the UAR case, although we cannot directly observe it, Λ takes on the true value λ^* on the first day and then remains constant.

Arguing as in the previous section, conditional on $\Lambda = \lambda$ the long-run fraction of satisfactory calls is

$$\frac{\mathrm{E}[S_1|\Lambda = \lambda]}{\mathrm{E}[N_1|\Lambda = \lambda]} = \frac{s(\lambda)}{\lambda t}, \tag{10}$$

where we have used the notation $s(\lambda)$ for the conditional expectation of S_1 given $\Lambda = \lambda$. The unconditional long-run performance is therefore $s(\Lambda)/(\Lambda t)$, which is random because it depends on the unknown Λ. We might then select the staffing level so that, with high probability, this fraction is larger than some specified level. But how do we compute $s(\Lambda)$?

4.1 Steady-state approximations

Under conditions that are explored in §5 we can employ the steady-state approximation $s(\lambda) = \lambda t f(\lambda)$. In that case we see from (10) that long-run performance is simply $f(\Lambda)$. The expected long-run performance is then $\mathrm{E}f(\Lambda)$, which differs from the RVAR case in that it does not weight the function f by Λ, as noted in the introduction. The random variable $f(\Lambda)$ can be analyzed reasonably easily once f and the distribution of Λ are known, as noted earlier.

Turning to short-run performance, the argument leading to (9) is directly relevant, and so we can approximate the distribution of S_1/N_1 as $\mathcal{N}(f(\Lambda), \sigma^2(\Lambda)/(\Lambda t))$. This distribution is an amalgam of *parameter uncertainty* in that the true arrival rate is unknown, and *process uncertainty* that is exhibited through the normal distribution for any given Λ.

4.2 Simulation-based estimates

If steady-state approximations are deemed inappropriate then we may turn to simulation. Recall from (10) that long-run performance is given by the random quantity $s(\Lambda)/(\Lambda t)$. We can write this as

$$\mathrm{E}\left[\left.\frac{S_1}{\Lambda t}\right| \Lambda\right]. \tag{11}$$

The distribution function of the conditional expectation (11) is relevant for computing the probability that long-run performance is satisfactory. The density is of interest in understanding how the uncertainty modelled by Λ translates into uncertainty about performance. Methods for estimating the distribution function and density of a conditional expectation can be found in Lee [1998] and Steckley and Henderson [2003] respectively. These simulation methods involve a combination of "macro replications" that sample observations of Λ, and "micro-replications" that estimate the conditional expectation for a sampled value of Λ.

One may prefer to simply determine summary statistics of (11) such as the mean. In this case, the discussion given in §3.2 about the use of stratified sampling is directly relevant.

Recall that the short-run performance measures in the UAR case coincide exactly with those for the RVAR case, and so the methods sketched in §3.2 are directly relevant.

5 Accuracy of Steady-State Approximations

The use of steady-state approximations rests on the assumption that they are reasonably accurate. In this section we consider a few simple models that are both mathematically tractable and practically relevant to get a sense of the error in steady-state approximations. Here the term "error" refers to the difference

between steady-state and time-dependent performance measures for a given model, and not to the difference between performance for a real system and the performance predictions based on a simplified model of the system. In the case we examine there is a single model. In the latter case that we do not examine there are two.

In this section we focus on the case where the arrival rate function is deterministic. In the random case, we can view the approximations below as holding conditional on the realized arrival rate function. So long as the approximations are reasonable over the range of the random/uncertain arrival rate function $\Lambda(\cdot)$ we can be confident that the results derived earlier using steady-state approximations are relevant.

We begin by sketching the key ideas, which are then applied to a succession of models. Suppose that the arrival process is Poisson with constant arrival rate λ. From Proposition 1 recall that the expected number of satisfactory calls in the period $[0, t]$ is

$$\mathrm{E}S_1 = \lambda \int_0^t P_\nu(X(s) \in B)\, ds, \tag{12}$$

where $P_\nu(X(s) \in B)$ gives the probability under initial distribution ν that a customer arriving at time s would receive satisfactory service. (We again drop the dependence of the process X on λ for ease of notation.) Our approximation replaces (12) with $\lambda t f(\lambda)$. Thus, the error is given by

$$\begin{aligned} \lambda \int_0^t P_\nu(X(s) \in B)\, ds - \lambda t f(\lambda) &= \lambda \int_0^t P_\nu(X(s) \in B) - f(\lambda)\, ds \\ &\approx \lambda \int_0^\infty P_\nu(X(s) \in B) - f(\lambda)\, ds. \end{aligned} \tag{13}$$

The approximation (13) simplifies our calculations. One might consider refinements to this approximation, but we do not do so here.

The expression (13) is closely related to *Poisson's equation* for Markov processes. To make that connection, first recall from Proposition 1 that so long as the Markov process is appropriately positive recurrent, $f(\lambda) = P_\pi(X(s) \in B)$. (Notice that the latter expression doesn't depend on s since π is a stationary distribution.) Then (13) can be written as

$$\lambda \int_0^\infty [P_\nu(X(s) \in B) - P_\pi(X(s) \in B)]\, ds = \lambda \int g(x)\nu(dx), \tag{14}$$

where

$$g(x) = \int_0^\infty [P_x(X(s) \in B) - P_\pi(X(s) \in B)]\, ds$$

and P_x is the probability when the Markov chain is started in the deterministic initial state x. Hence, the

error is a mixture of the function values $g(\cdot)$. The function g is known to solve *Poisson's equation*

$$Ag(x) = -[I(x \in B) - P_\pi(X(0) \in B)] \quad \forall x,$$

where $I(\cdot)$ is the indicator function that is 1 if its argument is true and 0 otherwise, and A is the *generator* of the Markov process. For example, when X is a continuous-time Markov chain on a discrete state space A is the rate matrix, and when X is a diffusion process A is a differential operator.

For many processes one can assert, using coupling theory or otherwise, that g is π-integrable, i.e., $\int |g(x)|\pi(dx) < \infty$. Furthermore, notice that

$$\begin{aligned} \int g(x)\pi(dx) &= \int \int_0^\infty [P_x(X(s) \in B) - P_\pi(X(s) \in B)]\, ds\, \pi(dx) \\ &= \int_0^\infty \int [P_x(X(s) \in B) - P_\pi(X(s) \in B)]\, \pi(dx)\, ds \qquad (15) \\ &= \int_0^\infty [P_\pi(X(s) \in B) - P_\pi(X(s) \in B)]\, ds \\ &= 0. \end{aligned}$$

Thus, if the interchange (15) is valid, then g has mean zero under the stationary distribution π.

Therefore, our agenda for each model is as follows:

1. Solve Poisson's equation to obtain $g(x)$ using the fact that g is π-integrable and integrates to 0.
2. Compute $\lambda\nu g \stackrel{\triangle}{=} \lambda \int g(x)\nu(dx)$ as an approximation for the error and hence compute the relative error by dividing by the steady-state approximation $\lambda t f(\lambda)$.

These quantities then lend insight into when the approximation can be expected to be reasonable. But what should we take for the initial distribution ν? A reasonable candidate, and the one that we adopt, is to take ν to be the stationary distribution associated with the parameters of the previous period. Again one can imagine refinements to this approximation, but this one should capture the key behavior. For mathematical tractability we assume that only the arrival rate changes between periods, and the other parameters including service rate and number of agents remain the same. Of course, it is typically the case that agent levels change between periods. A more complete analysis might also consider such changes. However, our approach dramatically simplifies the analysis and we believe that it captures the main effects.

We consider several models, and for each model consider different regimes, which correspond to the relative values of the customer arrival rate and the maximum possible service rate. We focus on two main regimes: the "efficiency-driven" regime (very high arrival rates) and the "quality and efficiency driven" regime (a careful balance between arrival rate and service capacity). The study of these regimes also

provides insight into the quality-driven regime where the arrival rate is small relative to the maximum service rate, showing that in that setting steady-state approximations are usually very accurate. For simplicity we assume that in both the previous period and the current one we remain in the same regime. This will not always be the case in practice, but it makes the calculations more tractable.

5.1 Efficiency-Driven Models

We first consider the $M/M/1$ queue, partly to demonstrate the methodology in a transparent setting, and partly because the conclusions we draw from this model can be extrapolated to more realistic models. We restrict attention to the $\tau = 0$ case here, so that the goal is to immediately answer calls. (It is possible to treat the $\tau > 0$ case for the $M/M/1$ queue, but for more complicated models the analysis appears to be difficult.)

5.1.1 The $M/M/1$ queue

An appropriate Markov process is $X = (X(s) : s \geq 0)$, where $X(s)$ gives the number of customers in the system (including in service) at time s. This is a CTMC on state space $\{0, 1, 2, \ldots\}$ with rate matrix A having non-zero off-diagonal elements $A_{i,i+1} = \lambda$ and $A_{i+1,i} = \mu$, $i \geq 0$. An arriving call is immediately answered if the system is empty, so the set $B = \{0\}$. If $\rho \stackrel{\triangle}{=} \lambda/\mu < 1$ then X has a steady-state distribution π where $\pi_i \stackrel{\triangle}{=} \pi(\{i\}) = (1-\rho)\rho^i$, so that $f(\lambda) = 1 - \rho$. Poisson's equation is then

$$Ag(x) = -(I(x = 0) - (1 - \rho)) \quad \forall x \geq 0.$$

This set of difference equations has π-integrable solution $g(x) = \kappa - x/\mu$, and since $\pi g = 0$ the constant $\kappa = \rho/\mu(1-\rho)$.

This expression for g confirms the intuitive notion that the error is smallest when the initial state is close to typical steady-state values ($g(x) = 0$ for $x = \mathrm{E}_\pi X(0)$). Also, $g(x)$ is positive for smaller values of x, so that if the initial state is small relative to steady-state conditions, then the true performance level is greater than the steady-state approximation. This again is consistent with our intuition that suggests that performance should be better when the system has less customers in it. The reverse applies when x is larger than typical steady-state values.

But how significant are these biases? To answer that question we compute $\lambda\nu g$, where ν is the stationary distribution for the parameters associated with the previous period. Suppose that the previous arrival and

service rates are λ_0 and μ_0. Let $\rho_0 = \lambda_0/\mu_0$. A direct calculation then shows that

$$\lambda\nu g = \frac{\rho(\rho - \rho_0)}{(1-\rho)(1-\rho_0)}$$

so that the relative error is

$$\frac{\lambda\nu g}{\lambda f(\lambda)t} = \frac{1}{\mu t}\frac{(\rho - \rho_0)}{1-\rho_0}\frac{1}{(1-\rho)^2}. \tag{16}$$

From (16) we see that if ρ is close to ρ_0 then the relative error is negligible (as expected since the system then remains in steady state). The sign of the error is the same as the sign of $\rho - \rho_0$, which means that true performance is better than the approximation when $\rho_0 < \rho$, again as expected. When μt is large, i.e., when the expected number of service completions when the server remains busy over the entire period is large, the error is reduced. This observation reinforces the results of Whitt [1991]. The error is magnified by the factor $(1-\rho)^{-2}$, suggesting that when the system enters heavy traffic the errors can be significant. Interestingly, when the system *leaves* heavy traffic (so that ρ_0 is close to 1 but ρ is not), the error is of the order $(1-\rho_0)^{-1}$, which is an order of magnitude smaller than when the system enters heavy traffic. We see exactly this behavior in Figure 2.

5.1.2 The $M/M/c$ queue

The results for the $M/M/1$ queue offer insight, but are they representative of multi-server systems? Let us now consider the $M/M/c$ queue with c servers with a large load of customers. Here we look at the "efficiency driven" regime where one tries to keep a very high utilization of agents.

Again the number of customers in the system is a CTMC, with a well-known rate matrix (e.g., Wolff [1989]). A customer immediately enters service if the number of customers seen on arrival is $c-1$ or less. Accordingly we solve Poisson's equation for $B = \{0, \ldots, c-1\}$, so that $f(\lambda) = \pi_0 + \cdots + \pi_{c-1}$. It is possible to show that the solution to Poisson's equation is of the form $\kappa - x/(c\mu)$ for $x > c$, with a more complicated form for $x \leq c$. The expressions are cumbersome and difficult to extract meaning from. We instead perform the calculations outlined above numerically. Some care is needed in the calculations due to numerical instabilities.

A representative plot of g is given in Figure 1. Without loss of generality we took $c\mu = 1$. (This is merely a choice of time scale.) We see similar behavior to the $M/M/1$ solution in that g is decreasing and crosses 0 near the steady-state mean, which is approximately 56 for this example. Notice also the clear linear growth beyond c. The slight "kink" near $x = 0$ is due to the use of a "guess" to replace unreliable values due to numerical difficulties.

A representative plot of the relative error for various parameter values is given in Figure 2. As with

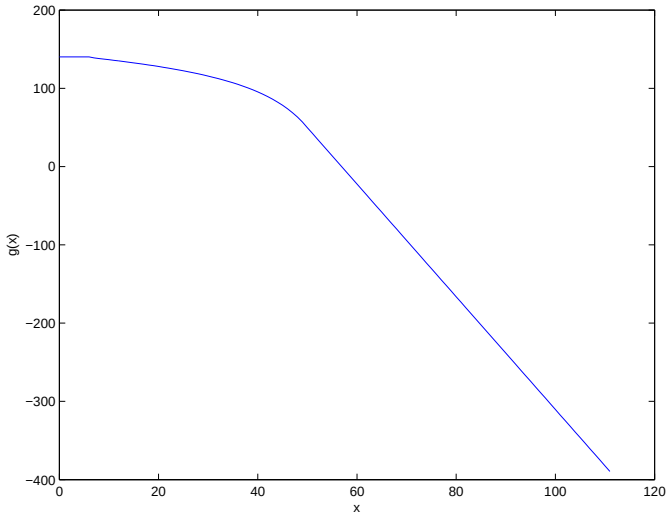

Figure 1: The solution to Poisson's equation for the $M/M/c$ queue where $\lambda = 0.94$, $\mu = 0.02$ and $c = 50$

the $M/M/1$ model we see dramatic increases in relative error as $\rho \to 1$ that are tempered near the line $\rho = \rho_0$. Furthermore, the sign of the error coincides with the sign of $\rho - \rho_0$ in line with intuition. Finally note that for fixed ρ and increasing ρ_0 the (negative) error increases in absolute value, again agreeing with intuition that says that if the queue is very long at the start of the current period then we can expect true performance to be poor even if the steady-state values indicate otherwise.

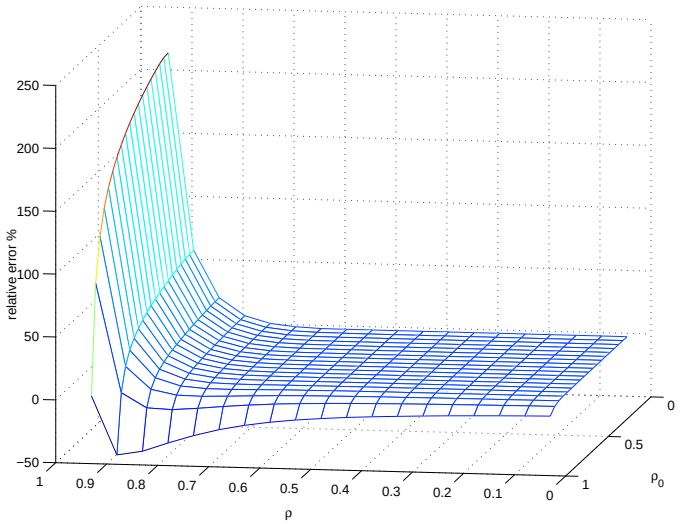

Figure 2: The relative error for the $M/M/c$ queue expressed as a percentage with $\mu = 0.02$, $c = 50$, and values of $\lambda = \rho$ and $\lambda_0 = \rho_0$ in $[0.04, 0.94]$

As noted earlier, the explosive behavior of the relative error in heavy traffic is similar to that seen in the $M/M/1$ queue. We can understand this phenomenon using diffusion approximations. The process

giving the number of customers in the system over time in a heavily-loaded $M/G/c$ queue with finite service-time variance can be approximated by a regulated (reflected) Brownian motion with appropriate parameters. This approximation is established through a rigorous limit theorem that arises as the arrival rate approaches the maximum service capacity $c\mu$. This is known as the "efficiency-driven regime" since it reflects the notion that agents are very heavily utilized. Agents are only very rarely free, and so the behavior of the approximation is dominated by the behavior of the original birth-death process above the level c. We now study the diffusion process to see if further insight can be developed.

The approximation is $X(\cdot) \approx Y(\cdot; -\gamma, \sigma^2)$, where $\gamma = c\mu - \lambda$, $\sigma^2 = c\mu(c_s^2 + \rho)$, $\rho = \lambda/(c\mu)$ and c_s^2 is the squared coefficient of variation of the service times (see, e.g., Whitt [2002b, §10.2]). Here $Y(\cdot; -\gamma, \sigma^2)$ is a regulated Brownian motion with drift $-\gamma$ and infinitesimal variance σ^2.

The process Y lives on the state space $[0, \infty)$ and so we take $B = [0, b]$ for some $b \approx c - 1$. A reasonable choice is $c - 1/2$ where the value $1/2$ is a "continuity correction."

When $\gamma > 0$, Y has a stationary distribution that is exponential with mean $\eta^{-1} \stackrel{\triangle}{=} \sigma^2/2\gamma$. Therefore $f(\lambda) = P_\pi(Y(0) \in B) = 1 - e^{-\eta b}$. Poisson's equation is then

$$\begin{aligned} \frac{\sigma^2}{2} g''(x) - \gamma g'(x) &= -[I(x \leq b) - f(\lambda)] \\ g'(0) &= 0. \end{aligned} \tag{17}$$

See Karlin and Taylor [1981, Chapter 15] and Glynn [1990] for background on diffusions and equations of this type.

The solution to (17) is [Glynn and Torres, 1996]

$$g(x) = \frac{b}{\gamma} e^{-\eta b} + \begin{cases} \frac{e^{-\eta b}}{-\gamma\eta}(e^{\eta x} - \eta x - 1) & 0 \leq x \leq b, \\ \frac{e^{-\eta b}}{-\gamma\eta}(e^{\eta b} - \eta b - 1) + \frac{1 - e^{-\eta b}}{-\gamma}(x - b) & x > b, \end{cases}$$

where we have used the fact that $\pi g = 0$ to compute an additive constant. Notice that the solution is decreasing, and beyond b is linear, exactly as in the $M/M/c$ model.

Next we compute the relative error by first computing $\lambda\nu g$, where ν corresponds to the stationary distribution associated with the parameters of the previous period. Letting η_0^{-1} denote the steady-state mean occupancy in the previous period, algebra reveals that

$$\lambda\nu g = \frac{\lambda b e^{-\eta b}}{\gamma} + \frac{\lambda\eta}{\gamma\eta_0} \frac{e^{-\eta_0 b} - e^{-\eta b}}{\eta_0 - \eta}.$$

The approximations $e^{-\eta b} \approx 1$ and $e^{-\eta_0 b} - e^{-\eta b} \approx (\eta - \eta_0)b$ give

$$\begin{aligned} \lambda\nu g &\approx \frac{\lambda b}{\gamma}\left(1 - \frac{\eta}{\eta_0}\right) \\ &= \frac{b(c_s^2+1)}{c_s^2+\rho}\frac{\rho(\rho-\rho_0)}{(1-\rho)(1-\rho_0)}, \end{aligned} \tag{18}$$

where (18) follows since $\sigma^2 = c\mu(c_s^2 + \rho)$ and $\sigma_0^2 = c\mu(c_s^2 + \rho_0)$.

Finally,

$$\begin{aligned} \lambda f(\lambda)t &= \lambda(1 - e^{-\eta b})t \\ &\approx \lambda\eta b t \\ &= \frac{2\lambda b t(1-\rho)}{c_s^2+\rho}, \end{aligned}$$

so that

$$\frac{\lambda\nu g}{\lambda f(\lambda)t} \approx \frac{1+c_s^2}{2}\frac{1}{c\mu t}\frac{\rho-\rho_0}{1-\rho_0}\frac{1}{(1-\rho)^2}. \tag{19}$$

Notice the close correspondence between the expression (19) and (16). This suggests that the same observations made for the $M/M/1$ model also apply for models that can be approximated by RBM. This includes a large class of multi-server queues that includes the $M/G/c$ family of queues with finite service time variance. Service time variability is exhibited through the factor $(1 + c_s^2)/2$, so that the relative error increases with service time variability.

5.2 Quality and Efficiency Driven Models Without Abandonment

A criticism of the efficiency-driven regime explored in the previous section is that the approximations are most valid when agents are busy almost constantly and customer waiting times are large or excessive. Given a choice, one would usually prefer to operate in the so-called "quality and efficiency driven" regime, which is also known as the Halfin-Whitt regime in honor of Halfin and Whitt [1981]. In this regime, not only are the servers highly utilized, but also servers are free for a nontrivial amount of time. In this regime the agents are highly utilized *and* customers receive a high level of service.

While a study of Figure 2 reveals some insights about this case, further insight can be obtained by studying a diffusion approximation that is relevant in the Halfin-Whitt regime. The study of this model can be further motivated by noting that the RBM approximation given in the previous section is a special case of the approximation given here when the arrival rate is excessively high relative to the maximum service rate. These appealing properties of the Halfin-Whitt regime come at a cost: the diffusion approximation is tractable for only a small family of service-time distributions. We restrict attention to exponential service

times here.

The approximation is [Halfin and Whitt, 1981]

$$X(\cdot) \approx c + \sqrt{c}Y(\cdot;\beta), \tag{20}$$

where $\beta = \sqrt{c}(1-\rho) > 0$ and $\rho = \lambda/(c\mu)$. Here $Y(\cdot;\beta)$ is a diffusion on $(-\infty,\infty)$ with drift function

$$\mu(x) = \begin{cases} -\mu\beta & x \geq 0 \\ -\mu(\beta + x) & x < 0 \end{cases}$$

and constant infinitesimal variance 2μ. This approximation is justified by taking a limit as the number of servers c and the arrival rate λ increase so that β converges to some positive value. It is therefore most relevant when the number of servers c is large.

Let $Y(\infty)$ denote a random variable distributed according to the steady-state distribution of Y. Then

$$\begin{aligned} P(Y(\infty) > 0) &= \alpha \stackrel{\triangle}{=} \left[1 + \frac{\beta\Phi(\beta)}{\phi(\beta)}\right]^{-1}, \\ P(Y(\infty) > x | Y(\infty) > 0) &= e^{-\beta x}, \quad x > 0, \text{ and} \\ P(Y(\infty \leq x | Y(\infty) \leq 0) &= \frac{\Phi(x+\beta)}{\Phi(\beta)} \quad x \leq 0, \end{aligned}$$

where ϕ and Φ are, respectively, the density and cumulative distribution function of a (standard) normal random variable with mean 0 and variance 1. Hence, the steady-state distribution π is a mixture of an exponential distribution on $(0,\infty)$ and a truncated normal distribution on $(-\infty, 0]$.

We want to solve Poisson's equation for the process $c + \sqrt{c}Y$ and function $I(x \leq b)$, where $b \approx c - 1$. For simplicity we take $b = c$. We first solve Poisson's equation for the process Y and the function $I(x \leq 0)$ to give h say, and then set

$$g(x) = h\left(\frac{x-c}{\sqrt{c}}\right).$$

We have that $f(\lambda) = P(Y(\infty) \leq 0) = 1 - \alpha$, and Poisson's equation is

$$\mu h''(x) + \mu(x)h'(x) = -[I(x \leq 0) - (1-\alpha)] \quad \forall x. \tag{21}$$

We can solve (21) analytically. Along the way we exploit the fact that h is π-integrable to establish that certain constants equal 0. The solution is

$$h(x) = \kappa + \begin{cases} \frac{-(1-\alpha)x}{\mu\beta} & x > 0 \\ \frac{\alpha}{\mu}\int_x^0 \frac{\Phi(s+\beta)}{\phi(s+\beta)}\, ds & x \leq 0. \end{cases}$$

The constant κ is chosen to ensure that $\pi h = 0$, and is given by

$$\kappa = \frac{\alpha(1-\alpha)}{\mu\beta^2} - \frac{\alpha(1-\alpha)}{\mu\Phi(\beta)} \int_{-\infty}^{\beta} \frac{\Phi^2(s)}{\phi(s)}\, ds.$$

Next we compute the relative error by first computing $\lambda\nu g$, where ν corresponds to the stationary distribution associated with the parameters of the previous period. We append a suffix of 0 to parameters for the previous period, and again assume for simplicity that only the arrival rate changes. Some algebra reveals that

$$\frac{\lambda\nu g}{\lambda f(\lambda) t} = \frac{1}{\mu\beta t}\left(\frac{\alpha}{\beta} - \frac{\alpha_0}{\beta_0}\right) + \frac{\alpha}{\mu t\Phi(\beta)} \int_{-\infty}^{\beta} \frac{\Phi(s)}{\phi(s)} \left(\frac{(1-\alpha_0)\Phi(\beta)}{(1-\alpha)\Phi(\beta_0)}\Phi(s+\beta_0-\beta) - \Phi(s)\right) ds.$$

Plots of this expression, again evaluated numerically, are essentially identical to Figure 2. To obtain some sense of the magnitude of the relative errors when $\rho_0 \approx \rho$, we use two-term Taylor expansions around β. The result after some effort is

$$\frac{1}{\mu t}\alpha(\beta_0 - \beta)\left[\frac{1}{\beta} + \frac{1-\alpha}{\beta^3} + \frac{1}{\Phi(\beta)} \int_{-\infty}^{\beta} \Phi(s)\, ds + \frac{\alpha}{\beta\Phi(\beta)} \int_{-\infty}^{\beta} \frac{\Phi^2(s)}{\phi(s)}\, ds\right].$$

Notice that the coefficient of $\beta_0 - \beta$ is positive. As a quick check we see that as $\beta \to 0$ (i.e., we approach heavy traffic), the dominant term in the coefficient is $(1-\alpha)/\beta^3$ which is of the order $\beta^{-2} = [c(1-\rho)^2]^{-1}$ as expected from previous results. Furthermore, as $\beta \to \infty$ (i.e., we approach light traffic conditions) the error decreases to 0 extremely rapidly; it is asymptotically of the order

$$\frac{1}{\mu t}\phi(\beta)(\beta_0 - \beta) \int_{-\infty}^{\beta} \frac{\Phi(s)}{\beta}\, ds.$$

(The integral in this expression converges to 1 as $\beta \to \infty$.)

5.3 Models With Abandonment

The results for the models considered up to now suggest, among other things, that for large traffic intensities the error can be significant. However, those models omit an important aspect of call centers, namely customer abandonment, that one might suspect may at least reduce the heavy-traffic effect. In this section we consider both the $M/M/c+M$ model, and a diffusion approximation for that model that is valid for large numbers of servers.

The $M/M/c+M$ model is identical to the $M/M/c$ model except that customers have limited patience. Customers are willing to wait an exponentially distributed amount of time (independent of all else) and,

if that time passes before they reach service, they depart without receiving service.

Abandonment has the effect of stabilizing the queue lengths, since even if customers arrive faster than they can be served, abandonment rates increase and keep waiting times in the queue small. Again our performance measure is the fraction of arriving customers who immediately receive service, so that unsatisfactory services include customers who abandon, or who wait in the queue for a positive amount of time before reaching a server.

Again the process giving the number of customers in the system over time is a CTMC, with rate matrix A given as follows. Let λ, μ and θ denote the arrival rate, the service rate for a single server, and the abandonment rate for a single customer. The nonzero off-diagonal entries of A are

$$\begin{aligned} A_{i,i+1} &= \lambda \qquad i \geq 0 \\ A_{i,i-1} &= i\mu \qquad 1 \leq i \leq c \\ A_{i,i-1} &= c\mu + (i-c)\theta \qquad i > c \end{aligned}$$

The stationary distribution π associated with this CTMC is easily computed numerically using standard birth-death results; see, e.g., Ross [1996, p. 253]. Then $f(\lambda)$ is given by $\pi_0 + \cdots + \pi_{c-1}$, the steady-state probability that there are $c-1$ or fewer customers in the system. Poisson's equation is again of the form

$$Ag(x) = -[I(x \leq c-1) - f(\lambda)].$$

We solve this equation numerically. Again care is required due to numerical issues. A representative plot of g is given in Figure 3. Notice that it decreases at a sublinear rate; this rate is clarified through the calculations for a diffusion approximation below. It is worthwhile comparing this plot to the solution to Poisson's equation for the $M/M/c$ queue in Figure 1. Notice the large reduction in scale, suggesting that the errors are much reduced when abandonment is taken into account.

This suspicion is confirmed when we look at the plot of the numerically-computed relative error in Figure 4. There is a dramatic reduction in relative error relative to the $M/M/c$ case. Again the error is greatest when ρ is large and ρ_0 is small, i.e., when the system becomes very busy after being less so.

As with the $M/M/c$ queue we can obtain further insight by considering an appropriate diffusion approximation. Garnett et al. [2002] show that the process X giving the number of customers in the system over time in the $M/M/c+M$ queue can be approximated by

$$X(\cdot) \approx c + \sqrt{c}Y(\cdot;\beta),$$

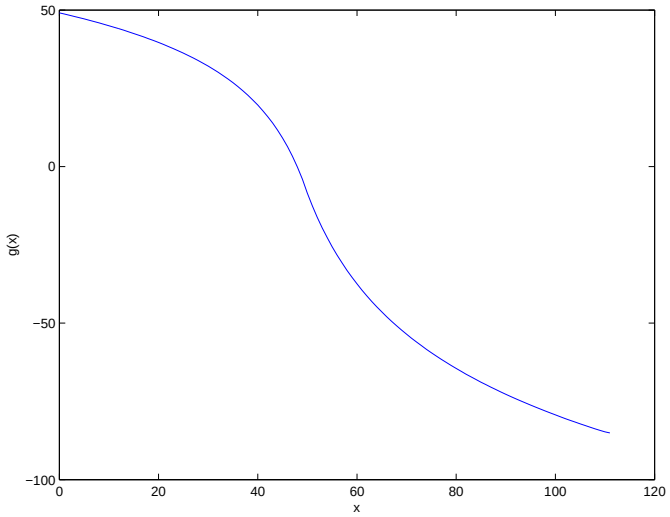

Figure 3: The Solution to Poisson's Equation for the $M/M/c+M$ Queue with $\lambda = 0.94$, $\mu = 0.02$, $c = 50$ and $\theta = 0.02$.

where $\beta = \sqrt{c}(1-\rho) > 0$, and $\rho = \lambda/(c\mu)$. Here $Y(\cdot;\beta)$ is a diffusion on $(-\infty, \infty)$ with drift function

$$\mu(x) = \begin{cases} -(\mu\beta + \theta x) & x \geq 0 \\ -\mu(\beta + x) & x < 0 \end{cases}$$

and constant infinitesimal variance 2μ. This approximation is justified by taking a limit as the number of servers c and the arrival rate λ increase so that β converges to a value that, unlike the $M/M/c$ case, is not restricted to be positive. (This reflects the fact that customers abandon if they wait too long, so a steady-state exists for any set of parameter values.)

Let $Y(\infty)$ denote a random variable distributed according to the steady-state distribution of Y. Let $r = (\theta/\mu)^{1/2}$. Then [Garnett et al., 2002]

$$\begin{aligned} P(Y(\infty) > 0) &= \alpha \stackrel{\triangle}{=} \left[1 + \frac{H(\beta r^{-1})}{r^{-1}H(-\beta)}\right]^{-1}, \\ P(Y(\infty) > x | Y(\infty) > 0) &= \frac{\bar{\Phi}(rx + \beta r^{-1})}{\bar{\Phi}(\beta r^{-1})}, \quad x > 0, \text{ and} \\ P(Y(\infty \leq x | Y(\infty) \leq 0) &= \frac{\Phi(x+\beta)}{\Phi(\beta)} \quad x \leq 0, \end{aligned}$$

where $\bar{\Phi}(\cdot) = 1 - \Phi(\cdot)$ is the complementary cdf, and $H(\cdot) = \phi(\cdot)/\bar{\Phi}(\cdot)$ is the hazard function of a standard normal random variable. Hence, the steady-state distribution π is a mixture of two truncated normal distributions.

We want to solve Poisson's equation for the process $c + \sqrt{c}Y$ and function $I(x \leq b)$, where $b \approx c - 1$. For simplicity we take $b = c$. We first solve Poisson's equation for the process Y and the function $I(x \leq 0)$

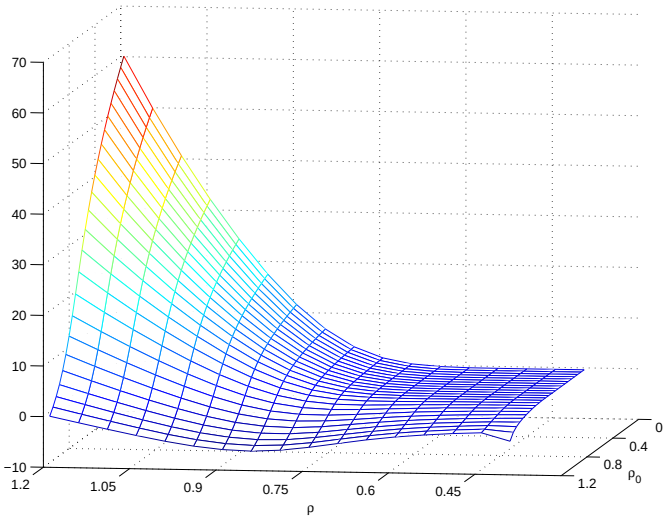

Figure 4: The relative error for the $M/M/c + M$ queue with $c = 50$, $\mu = 0.02$, $\theta = 0.2$, $\lambda \in [0.4, 1.2]$ and $\lambda_0 \in [0.04, 1.2]$. (Abandonment stabilizes the system so we do not require that $\lambda < c\mu$. Numerical problems arose for $\lambda < 0.4$)

to give h say, and then set

$$g(x) = h\left(\frac{x-c}{\sqrt{c}}\right).$$

We have that $f(\lambda) = P(Y(\infty) \leq 0) = 1 - \alpha$, and Poisson's equation is

$$\mu h''(x) + \mu(x)h'(x) = -[I(x \leq 0) - (1-\alpha)] \quad \forall x. \tag{22}$$

We can solve (22) analytically. Along the way we exploit the fact that h is π-integrable to establish that certain constants equal 0. The solution is

$$h(x) = \kappa + \begin{cases} \frac{-(1-\alpha)}{\mu r} \int_0^x \frac{\bar{\Phi}(rs+\beta r^{-1})}{\phi(rs+\beta r^{-1})}\, ds & x > 0 \\ \frac{\alpha}{\mu} \int_x^0 \frac{\Phi(s+\beta)}{\phi(s+\beta)}\, ds & x \leq 0. \end{cases}$$

The constant κ is chosen to ensure that $\pi h = 0$, and is given by

$$\kappa = \frac{\alpha(1-\alpha)}{\mu r^2 \bar{\Phi}(\beta r^{-1})} \int_{\beta r^{-1}}^{\infty} \frac{\bar{\Phi}^2(s)}{\phi(s)}\, ds - \frac{\alpha(1-\alpha)}{\mu \Phi(\beta)} \int_{-\infty}^{\beta} \frac{\Phi^2(s)}{\phi(s)}\, ds.$$

The growth rate of h is now clear: since $x\bar{\Phi}(x)/\phi(x) \to 1$ as $x \to +\infty$, h decreases at a logarithmic rate in the right tail. Similarly, the left-hand tail increases at a logarithmic rate.

Next we compute the relative error by first computing $\lambda \nu g$, where ν corresponds to the stationary distribution associated with the parameters of the previous period. We append a suffix of 0 to parameters

for the previous period, and again assume for simplicity that only the arrival rate changes. Algebra reveals that

$$\begin{aligned}\frac{\lambda\nu g}{\lambda f(\lambda)t} &= \frac{\alpha}{\mu t\Phi(\beta)}\int_{-\infty}^{\beta}\frac{\Phi(s)}{\phi(s)}\left(\frac{(1-\alpha_0)\Phi(\beta)}{(1-\alpha)\Phi(\beta_0)}\Phi(s+\beta_0-\beta)-\Phi(s)\right)ds \\ &\quad+\frac{\alpha}{\mu tr^2\bar{\Phi}(\beta r^{-1})}\int_{\beta r^{-1}}^{\infty}\frac{\bar{\Phi}(s)}{\phi(s)}\left(\bar{\Phi}(s)-\frac{\alpha_0\bar{\Phi}(\beta r^{-1})}{\alpha\bar{\Phi}(\beta_0 r^{-1})}\bar{\Phi}\left(s+\frac{\beta_0-\beta}{r}\right)\right)ds.\end{aligned}$$

Consider the case when $r = 1$, i.e., $\theta = \mu$. This leads to large simplifications, most likely since the queue process is then identical to that of an $M/M/\infty$ queue, which has a Poisson stationary distribution with mean λ/μ. A linear Taylor expansion of $\Phi(s+\beta_0-\beta)$ shows that

$$\frac{\lambda\nu g}{\lambda f(\lambda)t} \approx \frac{\alpha}{\mu t}(\beta_0-\beta)\left[\frac{1}{\bar{\Phi}(\beta)}\int_{\beta}^{\infty}\bar{\Phi}(s)\,ds+\frac{1}{\Phi(\beta)}\int_{-\infty}^{\beta}\Phi(s)\,ds\right]. \tag{23}$$

In light traffic, i.e., as $\beta\to\infty$, the bracketed term in (23) is of the order β, while $\alpha=\Phi(-\beta)$ is of the order $\phi(\beta)/\beta$, and hence the error is of the order $\phi(\beta)(\beta_0-\beta)/\mu t$, which converges to 0 very rapidly. In heavy traffic, i.e., as $\beta\to-\infty$, the bracketed term is of the order $-\beta$, while α converges to 1, and hence the error is of the order $(-\beta)(\beta_0-\beta)/\mu t$.

6 Conclusions and Future Research

We have described two settings where a random arrival rate arises. The settings differ in terms of whether the arrival rate is randomly varying, or simply unknown. We have given empirical evidence that suggests the RVAR case is essentially the rule rather than the exception.

The performance measures one should use in the two settings are similar but not identical. We have shown how to approximate these performance measures using steady-state approximations based on simple models, and also sketched how to estimate them using simulation. We will typically prefer to use steady-state approximations, but if the models they are based on represent too large a departure from reality then simulation may be preferred. We may also prefer simulation if the approximations do not accurately reflect the time-dependent performance measures we seek, even for the simple model they are computed from. This is a question we have explored in some depth. When the underlying model does not include abandonment, the error can be enormous under heavy loads. But in the more realistic case when the underlying model *does* include abandonment the errors are much smaller. The errors are still large for some parameter regimes that again coincide with heavy loads, so care still needs to be exercised in using the approximations.

This work immediately suggests a number of avenues for research, several of which we intend to pursue.

The "mixture of normals" distribution described in §3.1 is appealing due to its clear breakdown of the variability in performance due to both parameter uncertainty and process uncertainty. It remains to determine expressions or approximations for the function $\sigma^2(\cdot)$ for some common and useful models. This would go a long way to helping understand which form of uncertainty dominates, if any, and under what circumstances. This work will probably involve understanding multi-server workload models better than we do now. We have focused on performance measures related to the fraction of calls answered on time, since this is the industry standard. But other performance measures may make more practical sense and, in time, be adopted. It would be instructive to see whether the conclusions reached here apply more broadly to other performance measures, such as the one considered in Koole [2003], or perhaps some variant of the conditional value-at-risk measure that is receiving a great deal of attention in the risk management community.

Acknowledgements

This work was partially supported by NSF Grant DMI-0400287.

References

A. N. Avramidis, A. Deslauriers, and P. L'Ecuyer. Modeling daily arrivals to a telephone call center. *Management Science*, 50(7):896–908, 2004.

A. Bassamboo, J. Harrison, and A. Zeevi. Design and control of a large call center: Asymptotic analysis of an LP-based method. 2004. Submitted.

L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao. Statistical analysis of a telephone call center: A queueing-science perspective. 2002. Submitted.

L. D. Brown and L. H. Zhao. A test for the Poisson distribution. *Sankhya*, 64(A-3):611–625, 2002.

B. P. K. Chen and S. G. Henderson. Two issues in setting call center staffing levels. *Annals of Operations Research*, 108:175–192, 2001.

A. Deslauriers, P. L'Ecuyer, J. Pichitlamken, A. Ingolfsson, and A. N. Avramidis. Markov chain models of a telephone call center in blend mode. Submitted, 2004.

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

O. Garnett, A. Mandelbaum, and M. Reiman. Designing a call center with impatient customers. *Manufacturing & Service Operations Management*, 4(3):208–227, 2002.

P. Glasserman. *Monte Carlo Methods in Financial Engineering*. Springer-Verlag, New York, 2004.

P. W. Glynn. Diffusion approximations. In D. P. Heyman and M. J. Sobel, editors, *Handbooks in Operations Research and Management Science Volume 2: Stochastic Models*, pages 145–198. Elsevier (North Holland), 1990.

P. W. Glynn and M. Torres. Nonparametric estimation of tail probabilities for the single-server queue. In P. Glasserman, K. Sigman, and D. Yao, editors, *Stochastic Networks: Stability and Rare Events*, pages 109–138. Springer-Verlag, 1996.

W. K. Grassmann. Finding the right number of servers in real-world queuing systems. *Interfaces*, 18(2): 94–104, 1988.

L. V. Green and P. J. Kolesar. The pointwise stationary approximation for queues with nonstationary arrivals. *Management Science*, 37(1):84–97, 1991.

L. V. Green, P. J. Kolesar, and J. Soares. Improving the SIPP approach for staffing service systems that have cyclic demands. *Operations Research*, 49(4):549–564, 2001.

S. Halfin and W. Whitt. Heavy-traffic limits for queues with many exponential servers. *Operations Research*, 29(3):567–588, 1981.

J. M. Harrison and A. Zeevi. A method for staffing large call centers using stochastic fluid models. *Manufacturing & Service Operations Management*, 2005. To appear.

G. Jongbloed and G. Koole. Managing uncertainty in call centers using Poisson mixtures. *Applied Stochastic Models in Business and Industry*, 17:307–318, 2001.

S. Karlin and H. M. Taylor. *A Second Course in Stochastic Processes*. Academic Press, Boston, 1981.

G. Koole. Redefining the service level in call centers. Working paper, 2003.

A. M. Law and W. D. Kelton. *Simulation Modeling and Analysis*. McGraw-Hill, New York, 3rd edition, 2000.

S. H. Lee. *Monte Carlo Computation of Conditional Expectation Quantiles*. PhD thesis, Stanford University, Stanford, CA, 1998.

W. A. Massey and W. Whitt. Uniform acceleration expansions for Markov chains with time-varying rates. *Annals of Applied Probability*, 8(4):1130–1155, 1998.

S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Springer-Verlag, London, 1993.

A. M. Ross. *Queueing Systems with Daily Cycles and Stochastic Demand with Uncertain Parameters*. PhD thesis, University of California, Berkeley, Berkeley, California, USA, 2001.

S. M. Ross. *Stochastic Processes*. Wiley, New York, 2nd edition, 1996.

S. G. Steckley and S. G. Henderson. A kernel approach to estimating the density of a conditional expectation. In S. E. Chick, P. J. Sánchez, D. J. Morrice, and D. Ferrin, editors, *Proceedings of the 2003 Winter Simulation Conference*, pages 383–391, Piscataway, NJ, 2003. IEEE.

G. M. Thompson. Server staffing levels in pure service environments when the true mean daily customer arrival rate is a normal random variate. Unpublished manuscript, 1999.

W. Whitt. Bivariate distributions with given marginals. *The Annals of Statistics*, 4:1280–1289, 1976.

W. Whitt. The pointwise stationary approximation for $M_t/M_t/s$ queues is asymptotically correct as the rates increase. *Management Science*, 37:307–314, 1991.

W. Whitt. Dynamic staffing in a telephone call center aiming to immediately answer all calls. *Operations Research Letters*, 24:205–212, 1999.

W. Whitt. Stochastic models for the design and management of customer contact centers: some research directions. 2002a. Working Paper.

W. Whitt. *Stochastic-Process Limits*. Springer Series in Operations Research. Springer, New York, 2002b.

W. Whitt. Staffing a call center with uncertain arrival rate and absenteeism. 2004. Submitted.

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