COPING WITH TIME-VARYING DEMAND WHEN SETTING STAFFING REQUIREMENTS FOR A SERVICE SYSTEM

by

Linda V. Green

Graduate School of Business
Columbia University
lvg1@columbia.edu

Peter J. Kolesar

Graduate School of Business
Columbia University
pjk4@columbia.edu

Ward Whitt

IEOR Department
Columbia University
ww2040@columbia.edu

Submitted: April 2005; Revision Accepted: October 2005
Prepublication version: January 2006.

Abstract

We review queueing-theory methods for setting staffing requirements in service systems where customer demand varies in a predictable pattern over the day. Analyzing these systems is not straightforward, because standard queueing theory focuses on the long-run steady-state behavior of stationary models. We show how to adapt stationary queueing models for use in nonstationary environments so that time-dependent performance is captured and staffing requirements can be set. Relatively little modification of straightforward stationary analysis applies in systems where service times are short and the targeted quality of service is high. When service times are moderate and the targeted quality of service is still high, time-lag refinements can improve traditional stationary independent period-by-period and peak-hour approximations. Time-varying infinite-server models help develop refinements, because closed-form expressions exist for their time-dependent behavior. More difficult cases with very long service times and other complicated features, such as end-of-day effects, can often be treated by a modified-offered-load approximation, which is based on an associated infinite-server model. Numerical algorithms and deterministic fluid models are useful when the system is overloaded for an extensive period of time. Our discussion focuses on telephone call centers, but applications to police patrol, banking and hospital emergency rooms are also mentioned.

Keywords: staffing, call centers, time-varying demand, queues with time-varying arrival rate, nonstationary queueing models, police patrol, banking, hospital emergency rooms.

1. Introduction

A common feature of many service systems – ranging from telephone call centers to police patrol and hospital emergency rooms – is that the demand for service often varies greatly by time of day. This is illustrated by the plot of hourly arrival rates from a financial-services call center in Figure 1, taken from Section 4 of Green, Kolesar and Soares (2001). In this paper we discuss ways to cope with that time-varying demand when setting staffing requirements.

Since it helps to have a definite context in mind, we primarily focus on telephone call centers, where there already is a relatively high level of managerial control and sophistication, and extensive *information-and-communication-technology* (ICT) equipment, including *automatic call distributors* (ACD's), personal computers and assorted databases. In many call centers, staffing is performed by *workforce management* (WFM) software, which processes data and performs simple queueing analyses. For background on call centers, see the survey by Gans et al. (2003).

Many of our suggestions for call centers apply rather directly to other service systems, such as bank tellers, airlines ticket counters and tollbooths; e.g., see the classic toll-booth paper by Edie (1954). Moreover, the ideas apply in principle to other service systems, such as air-terminal queues (i.e., runways; Koopman 1972), police patrol (Green and Kolesar 1984a,b,

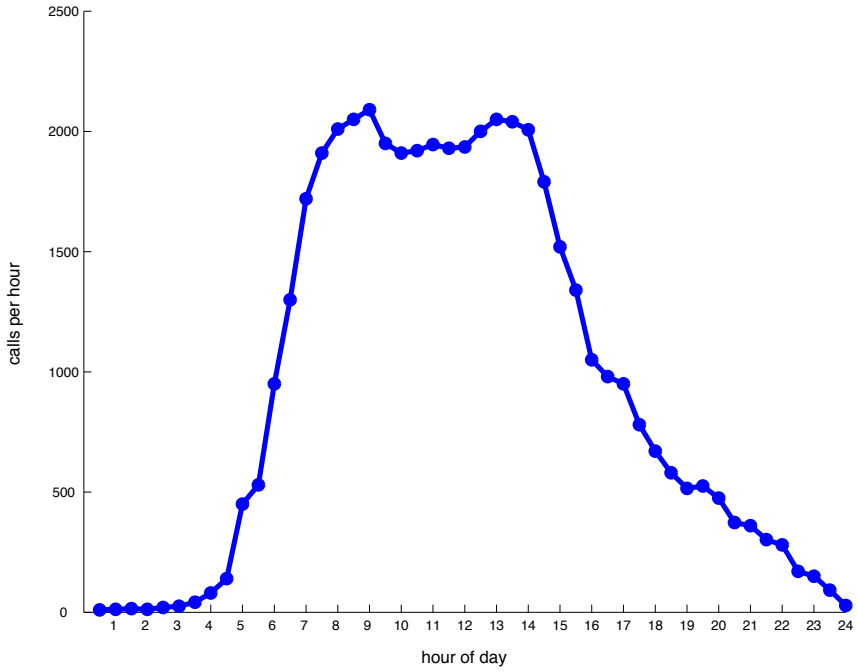

Figure 1: Arrivals per hour to a medium-sized financial-services call center.

1989, 2004) and hospital emergency rooms (Green et al. 2002, 2005), but the complexities of these systems invite more research. At the end of the paper we discuss the implications of our proposals for staffing in other service systems, including police patrol and hospital emergency rooms.

Organization of the Paper. In Section 2 we define the staffing problem and place it in context. In Section 3 we explain how stationary models can be used in a nonstationary manner to solve the staffing problem in the easiest cases - those systems with short service times and a high quality-of-service standard. In Section 4 we discuss refinements for harder cases with medium-to-long service times, but still with a high quality-of-service standard. We show how an associated infinite-server model can be used to develop and understand these refinements. In Section 5 we discuss the most difficult case, in which the system may be overloaded for an extensive period of time. In Section 6 we discuss staffing in other systems. In our final Section 7 we make concluding remarks, discussing extensions and other complications not addressed in the main paper, such as service systems with networks of facilities and systems with customer retrials.

2. The Staffing Problem

One Decision in a Hierarchy of Decisions. Setting staffing requirements is one in a hierarchy of decisions that must be made in the design and management of a service system. In a long-term planning horizon, managers set the system *capacity*. That usually involves hardware choices; e.g., in a call center managers determine the number of possible agent positions and the amount and capacity of supporting ICT equipment. In an intermediate-time planning horizon, managers set the overall size of the workforce, making important hiring and training decisions. The (daily) staffing decision specifies the number of customer service representatives (agents) needed to work during each staffing interval over the day.

After the staffing requirements are set, managers make agent *scheduling* decisions, specifying the number of agents to work on specific tours of duty, period by period, in conformance to the previously determined staffing levels, work rules and legal constraints. The scheduling decision is often determined by solving an *integer linear program* (Dantzig 1954, Segal 1974 and Kolesar et al. 1975). It is important to recognize that the staffing requirements could in principle be set by a larger algorithm that also addresses actual employee scheduling.

In real time, managers often make further adjustments – *flexing* decisions, which move

agents in and out of the line of duty (to and from "offline" work). This is accomplished by having extra agents on site doing alternative work or being trained, or by being able to use remote agents on short notice. If flexibility can be achieved, then it is often possible to efficiently provide a very high quality of service (Whitt 1999a).

In call centers, where the actual services required by customers are diverse and agent skills can be matched to them, we need to be concerned with the numbers of agents with different combinations of skills, not just the total number of agents. Telephone callers may speak different languages or may require special service. For example, we may need to ensure that enough agents are present to provide technical support in French and respond to billing inquiries in Spanish. When agents have different skills, staffing is intimately related to *call routing*. In this paper we act as if all agents can handle all calls, but at the end of the paper we indicate how the staffing methods for a single-skill call center can be applied to treat call centers with multi-skilled agents and skill-based routing. In many circumstances, the methods extend directly.

Other Common Characteristics of Service Systems. Service systems have several other common characteristics: First, a "service" usually must be performed relatively soon after the service request is made; there typically is *little opportunity to inventory* or backorder service requests, although a moderate kind of backordering – waiting – is usually allowed. It may be possible to prioritize jobs. For example, incoming calls to a 911 phone line saying "shots heard, man with a gun" get an immediate police response, while a "noisy party" gets queued. Setting priorities provides managers the opportunity to better meet service goals at less expense (Whitt 1999b).

Second, there is significant *variation in arrivals* around the temporal pattern; and significant *variation in service times*. One buys insurance against this uncertainty by overstaffing relative to the average demand and service rates. Hence it is natural to use a *queueing model* to compute the level of insurance needed.

A Single Basic Queueing Model. Our discussion will be centered around a single basic queueing model: the $M_t/GI/s_t + GI$ queue, which has a *nonhomogeneous Poisson arrival process* with a time-varying arrival-rate function $\lambda(t)$ (the M_t), independent and identically distributed (iid) service times following a general probability distribution (the first GI), a possibly time-varying number of servers on duty (agents, the s_t), an unlimited waiting space, a

first-come first-served queueing discipline, and customer abandonment with iid times to abandon following a general probability distribution (the $+GI$). The assumption of independence among customer times to abandon is realistic when, as happens in most telephone call centers, customers wait in *invisible queues*, where they cannot directly observe the state of the system.

As shown by Bolotin (1994) and Brown et al. (2005), service-time and time-to-abandon distributions tend to be non-exponential. They found the service-time distribution to be approximately lognormal, but the variability of the observed lognormal distributions was not too great. In particular, the squared coefficient of variation (SCV, variance divided by the square of the mean) was found to be between 1 and 2. (It is good to use the SCV instead of the variance, because it has meaning independent of the mean; the SCV of a random variable is unchanged if it is multiplied by a constant.) In such cases it is reasonable to use exponential distributions. In each new application the service-time distribution should be checked. We indicate what can be done if exponential distributions are not appropriate.

As emphasized by Garnett et al. (2002), Mandelbaum and Zeltyn (2004, 2005) and Feldman et al. (2004), in most call centers, some waiting customers abandon (leave without receiving service after joining the queue). A high level of customer abandonment may be a sign of poor service; indeed it often implies lost sales. On the other hand, a low level of abandonment, such as 1%, in a large call center may be a sign of proper staffing, where supply appropriately balances demand. Regardless of the interpretation, it can be useful to recognize the presence of customer abandonment and explicitly include its impact on performance and hence on staffing. Even a small amount of customer abandonment can significantly impact system performance and staffing requirements. In the past, abandonments were not included in staffing models, primarily because they appeared to make the model too complicated. However, we will show how the model with customer abandonment can be analyzed, and how its impact upon performance can be determined.

From a queueing-theory perspective, the $M_t/GI/s_t + GI$ model is quite complicated, primarily because of the time-varying arrival rate, but *actual call centers are often more complicated still*, since they can have multiple customer classes, agents with different skills and networks of work sites. Although the $M_t/GI/s_t+GI$ model does not address the full complexity of some actual call centers, it does permit us to analyze the impact of time-varying arrivals. Many of our ideas about how to cope with the time-varying arrival rate found from analyzing the $M_t/GI/s_t + GI$ queue apply more generally. (We discuss this briefly in the concluding Section 7.)

The Goal in Staffing. For us, then, the *staffing problem* is the specification of the *staffing-requirements function* s_t - the number of agents required to be on duty as a function of time t - which is the s_t term in the $M_t/GI/s_t + GI$ queue. However, changes in the staffing are usually allowed only at certain times, e.g., once every fifteen minutes, once every hour, or in some cases, only once every eight hours. Thus, we wish to determine a good staffing function subject to the constraint that changes are allowed only at the ends of prescribed *staffing intervals*.

Our goal is to minimize the total number of staff hours required over the day, while meeting a targeted level of service performance in each staffing interval. A common performance constraint is the *service level*: the requirement that $x\%$ of the calls be answered within y seconds. A commonly used standard is that 80% of the calls answered be within 20 seconds. Applied to a single time point, that means that the probability an arriving customer, who has unlimited patience, would have to wait no more than 20 seconds before starting service should be at least 0.80. (But, with customer abandonments, we want to properly account for the possible abandonment by customers waiting ahead of this waiting customer.)

Closely related to service level is the *delay probability*, i.e., the probability that an arriving customer has to wait at all before starting service. That is a special case of service level in which $y = 0$ seconds. The delay-probability constraint is generally easier to compute, tends to be a relatively robust performance measure (insensitive to model details) and tends to have a meaning independent of scale (typical number of servers), We elaborate on the independence of scale in Section 3.3.

Since customer abandonments are important and are measured automatically by modern call-center-management software, managers often place bounds on the abandonment rate, such as 4%. It is also common to constrain the expected waiting time (before starting service), which is called the *Average Speed to Answer* (ASA) by practitioners. (For a practice perspective on call center operations, see Cleveland and Mayben (1997).) Queueing theory helps link all four measures: service level, delay probability, abandonment rate and average speed to answer. Understanding these links can help diagnose difficulties in practice.

In many situations it is important to pay attention to non-congestion-related performance measures: Doing bad things with little delay does not constitute good service. It is vital to handle service requests properly as well as promptly; we seek *first-call resolution*. The goal of extracting the maximum value from the customer interaction by using an agent matched to the customer needs, with appropriate system support, leads to notions of *value-based routing* and *value-based staffing* (Sisselman and Whitt 2004).

A Daily Cycle. Here we assume that the total time period is a *day*. That might be an eight-hour day, conforming to normal business hours, or a full twenty-four-hour day, such as occurs with 911 phone lines and other continually-available (24/7) call centers, or something in between. In any event, the *common case* is to have significant variation in the arrival rate over the course of the day, so that the peak arrival rate is much greater than the average arrival rate. (Demand patterns often vary enough by day of the week so that this must be explicitly accounted for as well.)

A good example is the arrival-rate function depicted in Figure 1, based on empirical data from a financial-services call center. In Figure 1, each point is an average for a half-hour interval, multiplied by 2, to give the hourly rate. In that context, the average call holding time (service time) was about 6 minutes = 1/10 hour. Consequently, the instantaneous offered load (instantaneous arrival rate multiplied by the mean service time) would be $2000 \times (1/10) = 200$ if the arrival rate were 2000 calls per hour - which happens at about 9 am. Hence, to provide insurance against long delays, the required staffing at such times must exceed 200 agents.

There also can be somewhat predictable bursts of arrivals, e.g., as occur in a call center responding to planned television advertising promotions. There may also be sudden peaks or other anomalous behavior in the arrival-rate function at the beginning of the day, at noon time, or just prior to closing. There also can be unpredictable non-Poisson stochastic fluctuations, totally inconsistent with the nonhomogeneous-Poisson-process demand model we are assuming. Consider, for example, unexpected bursts of high demand in call centers serving brokerage customers when events of political or economic importance occur. Similarly, there are unexpected surges of demand for emergency responders in response to unanticipated large-scale incidents. While there is a need to plan for such eventualities, such phenomena will not be considered in this paper.

Model Fitting and Validation. The major step of *model fitting and validation* requires system *data*. However, service systems differ widely with regard to the availability of relevant data. Fortunately, call centers tend to fall into the data-rich category. We are in the midst of a technological revolution, making it ever more feasible and economical to be data-rich, but one still encounters some data-poor call centers. Data-poor environments are much more common in non-call-center applications. It can be important to facilitate economic transformation to a more data-rich environment by demonstrating the potential benefits of using better data in better models.

Here we do not focus on data analysis, but we do emphasize its importance. Unfortunately, there are few accounts of successful data analysis in the professional literature. However, Brown et al. (2005) is an excellent example of a statistical study of call-center data. Green and Kolesar (1984a, b, 1989) illustrate many data-analysis challenges in their use and validation of a queueing model of police patrol. Kolesar's (1984) study of automatic teller machines (ATM's) illustrates a case in the middle of the data-richness scale.

An important model fitting activity is *forecasting*, which itself is a hierarchial process. We will assume that a specified arrival-rate function forecast $\lambda(t)$ has been created. It is important to remember that forecasting is never perfect, so that there may be considerable uncertainty about the arrival-rate function $\lambda(t)$. When this is true, there are two fundamental sources of uncertainty: (i) stochastic fluctuations in arrivals and service times, for given model parameters, and uncertainty about the model parameters themselves. Here we only consider the first form of uncertainty, but it can be important to consider both. For recent research on models incorporating *both* forms of uncertainty, see Whitt (1999, 2005e), Harrison and Zeevi (2005) and Bassamboo et al. (2005a,b).

The Role of Simulation. A powerful time-tested approach to set staffing levels is to employ computer *simulation* (Anton et al. 1999, Brigandi et al. 1994 and Kwan et al. 1988). Simulation is especially useful to measure performance in systems that are so complex that they cannot be described by analytical queueing models. In data-rich environments, the simulation model can even be made an integral part of the data system, with specific models created and analyzed automatically as part of the data-analysis phase.

For any given staffing function s_t, evaluating the performance of the $M_t/GI/s_t + GI$ queue is relatively easy to do using computer simulation. But *choosing* a good staffing function is much more difficult, because there are usually a vast number of possibilities. For example, in a large call center with about 100 agents, there may be 20 available staffing-change points during a day and 20 reasonable candidate staffing levels at each of these times. That produces $20^{20} \approx 10^{26}$ different staffing functions to consider. So one cannot explore all staffing functions with simulation in a naive manner. Fortunately, there often are alternative simple analytical methods that make it possible to focus in on a small number of attractive alternatives. We will describe them in the rest of the paper. In practice it is a good idea to simulate the system in some detail after the staffing requirements have been identified by approximate analytical approaches in order to verify that the suggested staffing levels indeed produce the desired

performance.

3. Applying Stationary Models to Nonstationary Systems

Even though the arrival rate is highly time-varying, it may be possible to use stationary models to determine staffing requirements, but usually it is inappropriate to staff to the overall average arrival rate over the entire day; Green, Kolesar and Svoronos (1991) provide convincing numerical examples. On the other hand, it is often possible to use stationary models *in a nonstationary manner* - that is, chop time into segments and use a stationary model in each segment. This works well when the service times are short (e.g., $5-10$ minutes) and the quality-of-service standard is high. Under those conditions, systems are rarely overloaded and staffing requirements follow easily predictable patterns. The common case is when staffing intervals are short (e.g. $15-30$ minutes), but we will also briefly discuss longer staffing intervals (e.g, several hours or an entire day).

3.1. Short Staffing Intervals: PSA and SIPP

A Long History. There is a long history of using queueing models to set staffing requirements for groups of telephone operators. In the early days of telephony, a human telephone operator set up each telephone call, so the classic "call center" was a group of telephone operators. There was a strong daily pattern to demand, the service times were very short, and the average offered load was quite high. Because of the large system scale (and prevailing workforce rules), it was possible and economical to have short staffing intervals, e.g., 15 or 30 minutes (Segal 1974).

Similar situations emerged with the development of 800 numbers and telephone call centers. Even though the service times in these call centers are still short, they may experience high offered loads, and so may be even larger than the classical telephone operator group. For example, America On Line has a customer-support call center complex with over $10,000$ agents.

The Pointwise Stationary Approximation. The classic case with short service times, a high quality-of-service standard and short staffing intervals is a solved problem, in that an effective analytic strategy is to use what has been called a pointwise stationary approximation (PSA) - *PSA provides a time-dependent description of performance based on a stationary model, using the arrival rate and other parameters that prevail at each moment in time to describe the performance at that time.*

Adjustments for the Staffing Intervals: Segmented PSA and SIPP. However, direct application of PSA does not take account of the staffing intervals. As described, the PSA approach yields a time-dependent staffing function that does not restrict changes to be at the boundaries of staffing intervals. Experiments show that (with short service times and a high quality-of-service standard), if you could staff in that fully time-dependent manner, you would produce a good staffing function. However, we are typically constrained to hold the staffing level constant during each staffing interval.

Segmented PSA is a direct adjustment for the staffing-interval constraint - it works well when the staffing intervals are short: One generates the PSA-required staffing at each time t and then sets the staffing level to be the maximum of these staffing requirements over the staffing interval. Segmented PSA yields an upper bound on the required staffing, and tends to be effective for the case we are considering. Although segmented PSA may slightly overstaff, an initial staffing policy obtained by segmented PSA can easily be evaluated and refined by simulation.

In practice, many commercial call-center-management software packages use a different approach: *The arrival rate is first averaged over each staffing interval* and this average is used in a stationary model. Green, Kolesar and Soares (2001) refer to this as the *stationary independent period-by period* (SIPP) approach. A common idea underlies the segmented-PSA and SIPP approaches: Both use a stationary independent period-by-period approach. However, segmented PSA first determines the staffing level at each time point, whereas SIPP first averages the arrival rate over the staffing interval. When the arrival-rate function does not fluctuate too greatly over staffing intervals, SIPP and segmented PSA yield similar results. (Segmented PSA will produce somewhat higher staffing levels.)

For the classic case with short service times and short staffing intervals, SIPP does well provided that the arrival-rate function does not fluctuate too greatly within individual staffing intervals. Extensive experimental results evaluating the performance of SIPP as a function of model parameters are contained in Green, Kolesar and Soares (2001). They also propose several *refinements* to SIPP. One of these, SIPP Max, replaces the average arrival rate within each staffing interval by the maximum arrival rate within each staffing interval, which coincides with segmented-PSA.

In practice, it is common that call-center-management software only estimates the average arrival rate over staffing intervals, so that the arrival-rate function must be taken to be constant during each staffing interval. For such piecewise-constant arrival-rate functions, segmented

PSA (or SIPP Max) and SIPP will be identical. We can thus interpret the experiments from Green, Kolesar and Soares (2001) as providing a strong case for fitting a more realistic smooth estimate to the actual arrival-rate function in the cases where this refinement might be beneficial.

3.2. Long Staffing Intervals: Busy-Hour Engineering and SPHA

Now we discuss a second classic case, in which the service times are short and the quality-of-service standard is high, but the staffing interval is long. For example, the staffing interval might be eight hours or even an entire day. This case is not common in call centers, but it can occur. (Police staffing frequently uses three eight-hour tours of duty.) A classical example is determining the required number of trunk lines needed in a telephone exchange. In trunking there is no provision for waiting – so a multi-server loss model is used.

With long staffing intervals, there is an approach that reduces the problem to one with stationary demand. The idea lies in the requirement that satisfactory performance prevail at all times, so we staff (or set capacity) to *meet the peak demand* during the long staffing interval. That strategy led to *busy-hour engineering* (Bear 1980). Green and Kolesar (1995) refer to this approach as the *simple peak hour approximation* (SPHA).

In some circumstances, system managers may staff to meet *average performance* over the long staffing interval, instead of peak performance. But, while this is tempting, it is dangerous since focusing on average performance leads to understaffing at peak times. This produces complicated congestion, taking us to the difficult case discussed in Section 5.

3.3. Staffing for a Stationary Model

In the previous two subsections we observed that appropriate stationary models provide effective solutions to the classic staffing problems when service times are short and the quality-of-service standard is high. Thus, to address staffing for the $M_t/GI/s_t + GI$ model, it suffices to consider how to determine a staffing level s for the stationary $M/GI/s + GI$ model. We discuss ways to do that now.

Let λ denote the constant arrival rate. Let S denote a generic service time; let G denote its *cumulative distribution function* (cdf): $G(t) \equiv P(S \leq t)$, $t \geq 0$, with mean $\mu^{-1} \equiv E[S]$. An important quantity is the *offered load* $a \equiv \lambda E[S]$. (The notation a follows traditional usage in telephony (Cooper 1982).)

Numerical Methods. *We have made our original staffing problem less difficult by reducing*

it to a series of staffing problems in a stationary model. This reduced stationary problem can be solved by simulation. Since the staffing problem must be addressed for each of the staffing intervals during the day, it is natural to use a (stationary) simulation model just once to generate a table of the required staffing levels s as a function of the candidate arrival rates λ - given all other model parameters. Such tables can be periodically updated whenever the other model parameters change.

Instead of simulation, it is easier and, indeed, it is common practice to use the Erlang-C or $M/M/s$ model for this purpose. Let W_s denote the steady-state waiting time before starting service, when there are s servers. This random variable has an exponential distribution, except for an atom at 0. Using the standard service-level performance target, we would choose s to satisfy

$$P(W_s \leq 20 \quad \text{seconds}) \geq 0.80 > P(W_{s-1} \leq 20 \quad \text{seconds}) \ , \tag{3.1}$$

which is easy to compute numerically.

However, we may want to go beyond the $M/M/s$ model. Experience indicates that the next most important generalization to consider is usually abandonment. Fortunately, good algorithms also exist for the Erlang-A or $M/M/s+M$ model, which adds exponential abandonment times (Mandelbaum and Zeltyn 2005, Zeltyn and Mandelbaum 2005 and Whitt 2005a).

The Service-Time Distribution. It is easy to model exponential service times, for we need only match the observed sample mean (average) of the service-time data, because the exponential distribution has only one parameter, which can be taken to be the mean. Indeed, experience indicates that *the mean is the most important single parameter for any service-time distribution.*

But the exponential-service-time-distribution assumption should be validated by looking beyond the sample mean of the data. Experience indicates that *the second most important parameter is the SCV* c_s^2. We can estimate the SCV by estimating the variance as well as the mean, which in turn can be done by using the sample mean of the squares in addition to the sample mean. Since $c_s^2 = 1$ for an exponential distribution, if $\hat{c}_s^2 \leq 1$, the exponential-distribution assumption for the service times would be conservative. More generally, as a rough guideline, if $c_s^2 \leq 2$, then the exponential-distribution assumption for the service times tends to be a reasonable approximation.

In practice, it is not difficult to estimate the service-time distribution by its empirical distribution and use that in a simulation of the resulting $M/GI/s$ model to verify that the

performance captured by the exponential approximation is adequate: That procedure was illustrated by Kolesar (1984) in his study of queueing at automatic teller machines (ATM's). The data there were consistent with a gamma distribution having $c_s^2 = 0.5$. Simulation showed that the exponential assumption worked well.

When the exponential assumption does not fit well, it may be advantageous and not too difficult to calculate the service level in (3.1) for a more realistic model. When the average offered load (and thus the required number of servers) is low, it is possible to apply numerical algorithms to calculate all desired steady-state performance measures in approximating Markovian $M/Ph/s$ and $M/Ph/s + Ph$ models, which have both phase-type service-time and phase-type time-to-abandon distributions - usually with a small number of phases such as two (Seelen 1984, Seelen et al. 1985 and Takahashi and Takami 1976). Alternatively, we could use approximations, as in Whitt (1992, 1993). Again, these algorithms or formulas would be used only occasionally to produce tables of staffing levels as a function of arrival rates.

On the other hand, at high average offered load, with large numbers of servers, the computational complexity of numerical algorithms grows. Fortunately, *the service-time distribution beyond the mean tends not to matter much*, provided that the service-time-distribution SCV is not too far from 1 (Mandelbaum and Schwartz 2002, Mandelbaum and Zeltyn 2004 and Whitt 2000, 2004a, 2005a,b). In particular, the probability of delay tends to be relatively insensitive to the service-time distribution beyond its mean. Certainly, it is now well recognized that the impact of the service-time distribution beyond its mean in a many-server queue is very different than in a single-server queue (where the impact is significant). As important theoretical reference points, we remark that the infinite-server $M/GI/\infty$ and pure-loss $M/GI/s/0$ models have an *insensitivity property*: the steady-state distribution of the number of customers in the system depends on the service-time distribution only via its mean. Since abandonments tend to make the system behave like one of these models, the service-time distribution beyond the mean tends to matter less there as well.

However, it is important to be aware that high variability in the service-time distribution can significantly impact the conditional waiting-time distribution, given that the customer is delayed, and thus the service-level measure (Whitt 1992, 1993, 2000, 2004a, 2005a,b). We do not expect that to occur, but we should check.

The Time-to-Abandon Distribution. With abandonments in the model, we need to model the customer time-to-abandon (or patience) cdf, say $F(t)$, the probability that any

customer would abandon before time t if his service does not start before that time. In practice, customer abandonments are more difficult to model than service times since the time-to-abandon distribution is hard to observe and the time-to-abandon distribution affects performance in a more complicated way than the service-time distribution.

First, with abandonments there is *censored data* (Brown et al. 2005). Waiting customers end waiting in two different ways: by entering service or by abandoning. For the customers who enter service, we do not observe how long those customers would have been willing to wait before they abandoned; we only learn that they would have been willing to wait their observed waiting time before entering service. Thus we need to use statistical techniques for censored data

Even more important, however, is the different way that the time-to-abandon distribution affects performance, especially in systems with a large number of servers and/or a high quality-of-service standard. In those situations, few customers wait in queue extremely long. Thus *what is important about the time-to-abandon cdf is its value for small time arguments* (Mandelbaum and Zeltyn 2004, 2005, Zeltyn and Mandelbaum 2005 and Whitt 2005a,c). Perhaps surprisingly, the mean and the tail of the time-to-abandon distribution matter little. Thus we do *not* want to estimate the mean and higher moments to fit a distribution to those parameters.

Instead, from both perspectives - for effective estimation and for capturing what matters - it is preferable to estimate the time-to-abandon cdf F by estimating its *hazard-rate* (or failure-rate) function

$$r_F(t) \equiv \frac{f(t)}{1 - F(t)}, \quad t \geq 0 \ , \tag{3.2}$$

where f is the probability density function (pdf) associated with the cdf F (p. 276 of Ross 2003). It gives the conditional density of an abandonment at time t, given that the customer has not abandoned up to time t. The cdf F can be recovered from the hazard-rate function r_F by

$$F(t) = 1 - \exp\{-\int_0^t r_F(u)\, du\}, \quad t \geq 0 \ . \tag{3.3}$$

An important reference case is the exponential distribution, where the hazard rate function is constant.

To estimate the cdf F via its hazard-rate function r_F, suppose that time is divided into small subintervals, each containing an ample number of data points (abandonments). Then, let $\hat{n}_A(c,d)$ be the number of observed abandonments in the subinterval $[c,d) \equiv \{t : c \leq t < d\}$ and let $\hat{n}_W(c)$ be the number of customers that wait in queue at least time c before either

being served or abandoning. Then we estimate the hazard-rate function by

$$\hat{r}_F(t) \equiv \frac{\hat{n}_A(c,d)}{\hat{n}_W(c)(d-c)}, \quad c \le t < d \ . \tag{3.4}$$

It often suffices to work with an exponential time-to-abandon approximation, as was empirically justified by Brown et al. (2005). This works if the estimated hazard function does not fluctuate too much for small time arguments. We then fit a constant hazard rate θ by fitting a constant to the estimated hazard-rate function (3.4) above, using only data from the most relevant times, such as for the first 1 or 2 minutes. We then can use the $M/M/s+M$ model, using algorithms in Mandelbaum and Zeltyn (2005) or Whitt (2005a). The validity of this simplification can also be tested by simulation.

The Normal Approximation. To set staffing requirements, it is often not even necessary to calculate steady-state performance measures in the staffing interval. More elementary methods can be remarkably effective. When the offered load is not too small (say at least 5) and the targeted quality of service is high, the number of customers in the system is *approximately normally distributed*. A revealing derivation of the normal approximation is to first approximate the $M/GI/s+GI$ model by an infinite-server $M/GI/\infty$ model, having the same arrival rate and the same service-time distribution. The steady-state number of busy servers in the $M/GI/\infty$ model has (exactly) a Poisson distribution with mean equal to the offered load, $m = a \equiv \lambda E[S]$, independent of the service-time distribution beyond its mean (Example 5.16 of Ross 2003).

Given that the number of customers in the infinite-server model is Poisson, we immediately obtain the normal approximation from the well-known approximation discussed in most elementary probability textbooks, (e.g., p. 190 of Feller 1968). Since the actual distribution is Poisson, *the variance necessarily equals the mean*, so that there is only a *single parameter* in the normal approximation, namely, the offered load: $a \equiv \lambda E[S]$.

By the way, with abandonments there is additional justification for this infinite-server approximation: The number of customers in the system in the Markovian $M/M/s+M$ model, with customer abandonment, has *exactly the same distribution* as the number of customers in the associated Markovian infinite-server $M/M/\infty$ model when the abandonment rate equals the service rate.

In fact, there is a long history for using the infinite-server model. In classic telecommunications engineering it was used as an approximation for the Erlang-B (loss) model. In that setting it is called Molina's "lost-calls-held trunking" model (Molina 1922, Chapter XII of Fry 1965, p. 34 of Mina 1974 and p. 49 of Bear 1980).

The Square-Root-Staffing Formula. From the normal approximation, we immediately obtain the *square-root-staffing formula*:

$$s = a + \beta\sqrt{a} \ , \tag{3.5}$$

where $a \equiv \lambda E[S]$, the offered load, coincides with the mean number of busy servers in the infinite-server model, m, and β is a parameter reflecting the quality of service - in terms of delay and congestion - the quality of service (QoS) improves as β increases. A feasible integer staffing level is the least integer greater than or equal to s in (3.5).

With the normal approximation, we can directly relate the QoS parameter β in (3.5) to the desired steady-state delay-probability, which we denote by α. Letting Q be the number of busy servers in the infinite-server model, we approximate the steady-state delay probability α by

$$P(Delay) \equiv \alpha \approx P(Q \geq s) = P\left(\frac{Q-a}{\sqrt{a}} \geq \frac{s-a}{\sqrt{a}}\right) \approx 1 - \Phi(\beta) \ , \tag{3.6}$$

where Φ is the cdf of the standard (mean 0 and variance 1) normal distribution.

The analysis above supports a *simple rule of thumb*: if our aim is to avoid congestion without providing excessive capacity, then set $\beta = 2$. That produces a probability of delay of approximately 0.02. Applications of the normal approximation in a call-center context are discussed by Kolesar and Green (1998).

More Carefully Specifying the QoS parameter β. In some cases one choose the QoS parameter β in (3.5) to more-accurately satisfy performance constraints in the actual finite-server queueing model. We can often do substantially better without elaborate calculations, both in models with and without customer abandonment.

The derivation of refined methods focus on models with large numbers of servers (i.e., high offered load), but they in fact work for any number of servers at all; see Tables 1 and 2 of Jennings et al. (1996). The derivation relies on a *many-server heavy-traffic limit*, in which we let $s \to \infty$ and $\lambda \to \infty$, with the mean service time $1/\mu$ held fixed, so that

$$\frac{s-a}{\sqrt{a}} \to \beta \quad \text{where} \quad a \equiv \frac{\lambda}{\mu} \tag{3.7}$$

(Halfin and Whitt 1981, Whitt 1992, Garnett et al. 2002 and Borst et al. 2004).

From the defining limit in (3.7), we see that the many-server heavy-traffic regime also produces a square-root-staffing law, for in the limit we have $s \approx a + \beta \cdot \sqrt{a}$, which coincides with (3.5).

The many-server heavy-traffic limiting regime in (3.7) is shown to be appropriate as $s \to \infty$ and $\lambda \to \infty$; e.g., since the steady-state probability of delay approaches a limit strictly between 0 and 1. This limit will not hold for any other fundamentally different scaling. Thus, the steady-state delay probability has an interpretation independent of scale. It is noteworthy that the more popular service level does *not* have that property, because the conditional expected delay given that all servers are busy is asymptotically of order $1/\sqrt{s}$ as $s \to \infty$ (Halfin and Whitt 1981).

For the stationary Markovian $M/M/s$ model, without customer abandonment, there is a continuous strictly increasing function mapping the QoS parameter β into the limiting delay probability α, now commonly called the *Halfin-Whitt delay function*:

$$P(Delay) \equiv \alpha \approx HW(\beta) \equiv [1 + (\beta\Phi(\beta)/\phi(\beta))]^{-1} \ , \quad 0 < \beta < \infty \ , \tag{3.8}$$

where, again, Φ is the cdf and ϕ is the associated probability density function (pdf) of the standard normal distribution.

For the stationary Markovian $M/M/s + M$ model with customer abandonment and abandonment rate θ, Garnett et al. (2002) showed that a corresponding continuous strictly increasing function maps the QoS parameter β and the ratio of the abandonment rate to the service rate, $\theta_{rat} \equiv \theta/\mu$, into the limiting delay probability α. This is now commonly called the *Garnett delay function*:

$$P(Delay) \equiv \alpha \approx Garnett(\beta, \theta_{rat}) \equiv \left[1 + \sqrt{\theta_{rat}} \cdot \frac{h(\beta/\sqrt{\theta_{rat}})}{h(-\beta)}\right]^{-1} \ , \quad -\infty < \beta < \infty \ , \tag{3.9}$$

where $h(x) \equiv \phi(x)/(1 - \Phi(x))$ is the *hazard rate* of the standard normal distribution.

The Halfin-Whitt and Garnett delay functions in (3.8) and (3.9) apply to the $M/M/s$ and $M/M/s + M$ models. Extensions to other many-server models have been developed by Puhalskii and Reiman (2000), Jelenkovic et al. (2004), Zeltyn and Mandelbaum (2005) and Whitt (2004a, 2005b).

The QoS parameter β can be based on the targeted probability of delay, α, because they can be related, as shown in Figures 2 and 3. (See Borst et al. (2004) for a cost-benefit analysis, based on waiting and staffing cost rates.) We display plots of both the Halfin-Whitt and Garnett delay functions in Figures 2 and 3. We plot the Garnett delay function for five different values of $\theta_{rat} \equiv \theta/\mu$: 1/16, 1/4, 1, 4 and 16. When the abandonment rate is low, as in the case when θ_{rat} is equal to 1/16, the Garnett function is close to the Halfin-Whitt delay function, provided that β is not too small. Of course, without customer abandonment, the system is unstable for $\beta \leq 0$; hence $HW(\beta) = 1$ for $\beta \leq 0$.

The simple normal approximation in (3.6) *is also plotted in* Figures 2 and 3, because, as noted previously, the infinite-server model coincides exactly with the $M/M/s+M$ model when $\theta = \mu$, that is, when $\theta_{rat} = 1$, which is plotted in both Figures 2 and 3. The Halfin-Whitt

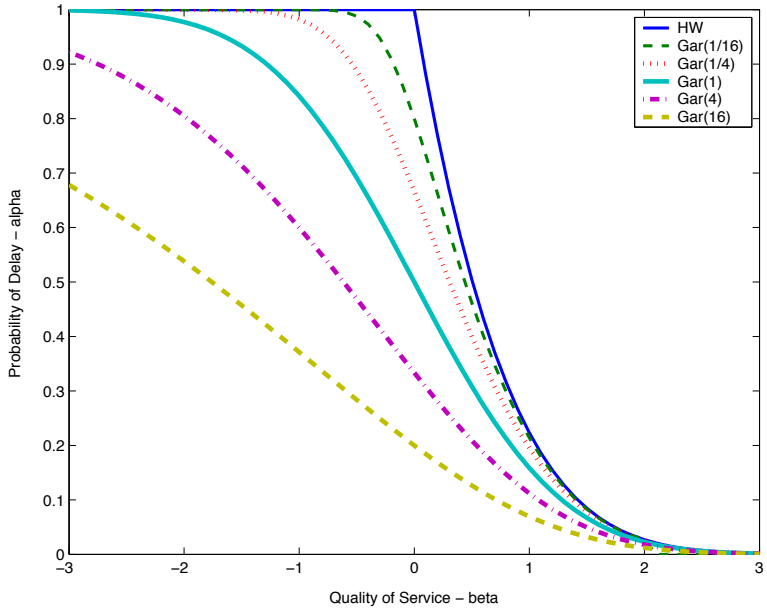

Figure 2: The Halfin-Whitt and Garnett functions mapping the QoS parameter β into the steady-state delay probability α. Five different values are considered for the parameter $\theta_{rat} \equiv \theta/\mu$: 1/16, 1/4, 1, 4 and 16.

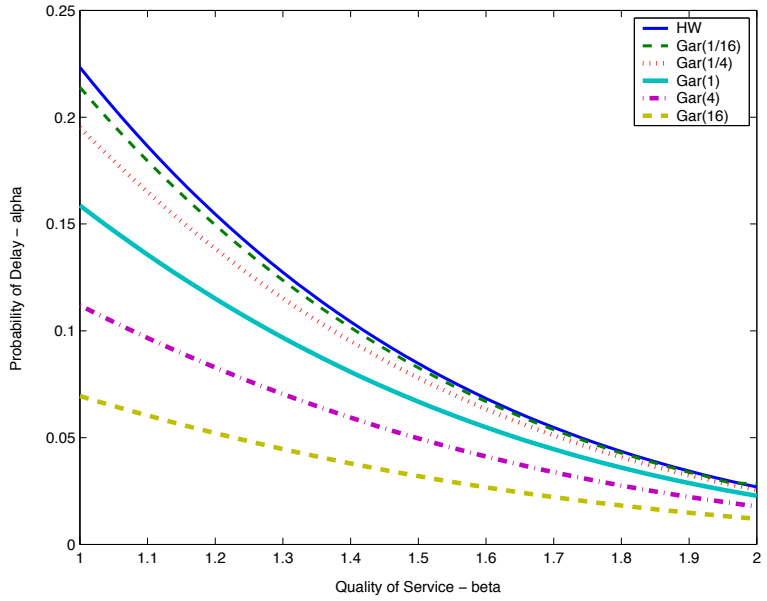

Figure 3: The Halfin-Whitt and Garnett functions mapping the QoS parameter β into the steady-state delay probability α, for β restricted to the interval $[1, 2]$. The same five different values are considered for the parameter $\theta_{rat} \equiv \theta/\mu$: 1/16, 1/4, 1, 4 and 16.

and Garnett delay-probability functions displayed in Figures 2 and 3 show the error caused by using formula (3.6) when $\theta_{rat} \neq 1$.

In our intended application to a call center, we would choose a target delay probability α, and then compute the appropriate QoS parameter β. For given model parameters and any selected target steady-state delay probability α, we can *invert* the function in (3.8) or (3.9) to obtain the desired β. The Halfin-Whitt function in (3.8) was first used to refine the normal approximation for the $M_t/M/s_t$ staffing problem by Jennings et al. (1996) and is summarized in Table 3 there; Feldman et al. (2004) used the Garnett function in (3.9) for the same purpose.

Looking at the inverse function in Figure 2, we see that the degree of abandonment, as measured by θ_{rat} can make a big difference in the staffing, when the quality of service is not too high. Figure 3 shows what happens for higher β, specifically for β restricted to the interval $[1, 2]$.

Example 3.1. *A typical situation with a high quality-of-service standard.* To illustrate a typical situation with a high quality-of-service standard, suppose that $a \equiv \lambda/\mu = 100$ and $\alpha = 0.10$, where the target delay probability is quite small. Then the square-root-staffing formula in (3.5) dictates that the number of agents should be the least integer greater than or equal to $s = 100 + 10\beta$, where β can be obtained from Figure 3. Given $\alpha = 0.10$, the appropriate value of β depends on $\theta_{rat} \equiv \theta/\mu$. A change from $\theta_{rat} = 4$ to $\theta_{rat} = 1/4$ increases β from 1.1 to 1.4, leading to a staffing change of 3 agents on a base of 111, which we regard as significant.

A change from $\theta_{rat} = 0$ (corresponding to the normal approximation) to $\theta_{rat} = 1/1000$ (corresponding to the case of no abandonment) increases β from about 1.28 to 1.42, leading to a staffing change of only 1.4 agents. We regard the required adjustment as minor, justifying use of the simple direct normal approximation based on the $M/M/\infty$ model. ∎

The infinite-server approximation, the normal approximation, the square-root-staffing formula and the many-server asymptotic regime all have long histories in the queueing literature. However, all these features and the linkages were initially not considered together; see Whitt (1992), Kolesar and Green (1998) and references therein. Indeed, the queueing pioneer A. K. Erlang (1948) even identified the most sophisticated aspect – the many-server heavy-traffic limiting regime – as early as 1924.

Other Performance Measures. So far, as a performance target to use in determining an appropriate staffing level, we have used the *probability of delay* α; see (3.6), (3.8) and (3.9). And, indeed, that is an important component of our suggested approach. However, management is often more directly concerned about other performance measures, such as the *proportion of customers that abandon*, $P(Ab)$, the expected delay, (*average speed to answer* - ASA) and the *service level* - having $x\%$ of all calls answered within y seconds, which requires that we know the proportion of customers that abandon and the conditional steady-state waiting time distribution given that a customer gets served. The service-level target can be expressed as

$$P(W_s \le y \quad \text{and} \quad \text{Served}) = P(\text{Served})P(W_s \le y|\text{Served}) \ge x \ ,$$

where W_s is the steady-state waiting time before starting service with s servers and $P(\text{Served}) = 1 - P(Ab)$.

From a practice perspective, the probability of delay is inconvenient because it is difficult to measure precisely. It is difficult to distinguish between entering service immediately upon arrival and having an extremely short delay. We recognize that and would advocate not making a strict definition. In practice we might regard any delay less than 2 seconds or 5 seconds as no delay at all.

In addition, we recognize that other performance measures are important. So it is significant that there are effective ways to calculate these other performance measures and relate them to the delay probability α and the basic model parameters. Through experience, we can learn how to relate the delay-probability target α to other performance targets we care more about.

For the Erlang-A model, convenient approximation formulas for these other performance measures are given along with the Garnett function (3.9) in Garnett et al. (2002) and Mandelbaum and Zeltyn (2005). Corresponding approximation formulas for the case of non-exponential time-to-abandon distributions are given in Zeltyn and Mandelbaum (2005). Simple fluid approximations for the $G/GI/s+GI$ model are given in Whitt (2005c). All these simple approximations provide a basis for directly considering other performance targets, but we do not expand upon that here.

We also can resort to numerical methods. An approximate numerical algorithm for the $M/GI/s+GI$ model is contained in Whitt (2005a). Finally, in practice all this can be confirmed and refined by, first, applying computer simulation and, second, comparing to system measurements.

Staffing to the Offered Load. The square-root-staffing formula (3.5) stemming from the normal approximation, possibly with refinements provided by (3.8) and (3.9), is remarkably easy and accurate. Yet there is an even more elementary formula that is often remarkably effective in larger systems, when customer abandonment is present and the targeted quality of service is not extremely high.

Paradoxically, in such cases it often suffices to simply apply a *naive deterministic approximation* and staff to the offered load; that is, just let $s = a \equiv \lambda E[S]$ (Feldman et al. 2004). Intuitively, we might anticipate a high abandonment rate, like 50%, in such a scenario. Amazingly, however, in large systems staffing to the offered load does not produce an extraordinary high rate of abandonment; the average abandonment rate might be 5% or less; e.g., see the examples in Feldman (2004). Even though the service goal may be to provide good quality of service, there is an advantage to having of a small amount of customer abandonment in the system, because the small amount of customer abandonment allow us to staff at a lower level. (And we would hope these abandoning customers would not be lost sales, but might retry later at a less congested time and get through.)

Staffing to the offered load tends to be appropriate when the target delay probability is about 0.5 (in the presence of customer abandonment). To see this, note that the Garnett function with $\theta_{rat} \equiv \theta/\mu = 1$ assumes the value $\alpha = 1/2$ at $\beta = 0$. Of course, if in fact $\theta_{rat} \equiv \theta/\mu << 1$, then such a staffing procedure will perform poorly.

Service-level Agreements. Many firms outsource the management of their call centers and an essential ingredient in such contracts are *service-level agreements* (SLA's), which specify requirements on the quality of service provided. It is useful to know how many agents are required to meet any given SLA.

To illustrate, we consider the following scenario: The customer arrival rate is $\lambda = 100$ calls per minute, while the mean service time is $\mu^{-1} = 4$ minutes (the offered load will be $a \equiv \lambda/\mu = 400$). Customers also abandon at rate $\theta = 1/4$ per minute ($\theta = \mu$). The outsourcer offers an SLA guaranteeing that 50% of all callers will not be delayed before starting service, that at most 2% of the customers will abandon, and that the average speed to answer will be at most 5 seconds. This proposal looks very good, but how many agents will the outsourcer need? Will he need 440, as dictated by the simple rule of thumb based on the square-root-staffing formula with $\beta = 2.0$? Surprisingly, the naive deterministic approximation, staffing to the offered load, that requires only 400 agents (40 agents less!) meets all those performance

requirements.

To understand why, it is helpful to consider what other assumptions would yield. If we had assumed no customer abandonment at all, and used the $M/M/s$ model, the square-root staffing formula with $\beta = 2.0$ would suggest 440 agents, which produces an excellent quality of service, with only about 2% of the customers delayed. In contrast, with the lower deterministic staffing level of 400, the queue length in the $M/M/s$ model would grow without bound. But, if customers do indeed have limited patience, with the mean abandonment above, the SLA would be met with 40 fewer agents.

Still, the simple rule of thumb based on the square-root staffing formula with $\beta = 2.0$ might actually be a better choice if the arrival-rate $\lambda = 100$ calls per minute is an imperfect estimate of the actual arrival rate. If for example there could be as much as 10% error in this estimate, the actual arrival rate might be 110 instead of 100 and the offered load would be 440 instead of 400, in which case the simple rule of thumb based on $\lambda = 100$ (luckily) produces just what is needed to be meet the naive deterministic staffing level at the higher offered load $a = 440$ associated with the higher arrival rate $\lambda = 110$; the two errors cancel!

Of course, management might well have recognized the data uncertainty and inflated the estimate of the arrival rate by say 10%. (Moreover, integer programming staffing-and-scheduling algorithms often introduce slack.) So, it could happen that on some days we may be overstaffing by 20% by using the simple rule of thumb based on the square-root staffing formula with $\beta = 2.0$.

Everything could work out by dumb luck, but everything could go wrong. A conclusion to draw from this example is that there is benefit from carefully analyzing the overall staffing process, using appropriate models, and paying attention to such issues as customer abandonment and model-parameter uncertainty. Proceeding in a haphazard manner can lead to understaffing or overstaffing by as much as $10\% - 20\%$ or possibly even more, even in the "easy" case with short service times.

4. A Harder Case: Medium-to-Long Service Times

The PSA-based approaches used with short service times in the previous section (segmented PSA or SIPP Max with short staffing intervals and SPHA or busy-hour engineering with long staffing intervals) will not perform well when the service times are longer. We consider this case now.

4.1. A Time Lag

Even in systems with a high quality-of-service standard, where overloading is to be avoided, medium-to-long service times modify the dynamics of the queueing system and can produce a significant impact on performance. In such cases we must modify the staffing algorithm to account for the concomitant strong time lags in congestion - that is, customer delays peak after the arrival-rate peaks. The reason is simple: Each arrival remains in the system for the length of his service time. Hence, even without considering complex congestion effects, the number of customers in the service system lags the arrival rate.

Hence, a natural adjustment to the previous methods – PSA, SIPP and SPHA – is to shift the arrival rate to the right by the mean service time before applying those simple stationary-model methods. When the service times are medium to long, this adjustment works well. Lagged versions of PSA were suggested by Eick et al (1993a,b) in the context of infinite-server models. Lagged versions of SIPP and SPHA were advocated and tested by Green, Kolesar and colleagues. In those tests, they used numerical methods to solve for time-dependent performance in the Markovian $M_t/M/s_t$ model. In particular, Green and Kolesar (1995, 1997) tested SPHA and lagged-SPHA for the case of long staffing intervals (SPHA), while Green, Kolesar and Soares (2001, 2003) tested SIPP and lagged SIPP for the case of short staffing intervals. Their experiments clearly show that the lag refinements, although not needed for short service times, are a significant improvement for medium-to-long service times. Their experimental results quantify the benefit provided by the lag method as a function of the mean service time and other model parameters in specific classes of models.

4.2. Insights from an Infinite-Server Model

A relatively simple model lets us drill down deeper to better understand staffing in the $M_t/GI/s_t + GI$ model with medium-to-long service times and a high quality-of-service standard. It is the corresponding infinite-server $M_t/GI/\infty$ model. Even though the arrival-rate is time-varying, the number of busy servers at each time in the $M_t/GI/\infty$ model has a simple probability distribution. The infinite-server model shows how many servers would actually be used if there were no resource constraints and thus no congestion (delay, loss or abandonment). Even though the actual system has only finitely many servers, the associated infinite-server model often gives a reasonable approximation to system performance - as long as there is a high quality-of-service standard.

Solution for the $M_t/GI/\infty$ Model. The number of busy servers at time t in the time-dependent $M_t/GI/\infty$ model has a Poisson distribution with a time-dependent mean $m_\infty(t)$, which is expressed in terms of the arrival-rate function $\lambda(t)$ and the service-time cdf G as follows:

$$m_\infty(t) = E[\lambda(t - S_e)]E[S] = E\left[\int_{t-S}^{t} \lambda(u)\, du\right] = \int_{-\infty}^{t} [1 - G(t-u)]\lambda(u)\, du\ , \tag{4.1}$$

where S_e is a random variable with the stationary-excess (or residual lifetime) cdf associated with the service-time cdf G, i.e.,

$$P(S_e \le t) \equiv \frac{1}{E[S]} \int_0^t [1 - G(u)]\, du, \quad t \ge 0\ , \tag{4.2}$$

with k^{th} moment

$$E[S_e^k] = \frac{E[S^{k+1}]}{(k+1)E[S]}\ , \tag{4.3}$$

and so mean $E[S_e] = E[S](c_s^2 + 1)/2$, where c_s^2 is SCV of the service time S; see Theorem 1 of Eick et al. (1993a), which we follow below in drawing implications from the exact representations in (4.1). Formally, Equation (4.1) applies to the situation in which the system began operation in the distant past. If we want to start at time 0, we just define $\lambda(u) = 0$ for $u < 0$.

Since a Poisson distribution is characterized by its mean, the time-dependent distribution of the number of busy servers in the $M_t/GI/\infty$ model is completely characterized by the deterministic time-dependent mean function $m_\infty(t)$ in (4.1). Moreover, this Poisson distribution supports the normal approximation and the square-root staffing formula in (3.5), but now with the modification that, Instead of using the PSA mean $m_{PSA}(t) \equiv \lambda(t)E[S]$, we use the exact time-dependent mean $m_\infty(t)$ in (4.1) in the $M_t/GI/\infty$ model.

Interpretation of the Mean Formula. We interpret the three components of equation (4.1) by relating them to the instantaneous offered-load, that is, the pointwise-stationary approximation (PSA) for the mean, $m_{PSA} \equiv \lambda(t)E[S]$, and the lagged PSA $m_{LaggedPSA} \equiv \lambda(t - E[S])E[S]$.

The first component of (4.1) shows that the PSA is correct except for a *random time lag*, with the random time lag being the stationary-excess variable S_e defined in (4.2), rather than just S. We note that this first representation is not identical to lagged PSA since the lag random variable is S_e instead of S and the expectation appears outside the arrival-rate function, not inside. Thus, the mean $E[S_e]$ is a natural candidate for the approximate lag, but this interpretation is not direct, because the expectation appears outside the arrival-rate function. We discuss how to *move the expectation inside* later in this section.

More insight into the random time lag can be gained from renewal theory. A service time in process in an endless succession of service times in equilibrium will have a residual remaining lifetime distributed as S_e. In equilibrium, the remaining service times of the customers in service in the stationary $M/GI/\infty$ model, conditioned on that number in service, are iid random variables, each distributed according to S_e (p. 161 of Takács 1962). When the service-time distribution is exponential, the random variable S_e has the same distribution as the service time S but not otherwise. In the exponential reference case, the time lag is approximately the mean service time, $E[S]$, and this suggests that, generally, the time lag should be about $E[S_e]$, as specified above. From this perspective, the approximate impact of different service-time distributions can thus be quantified; it was studied by Eick et al. (1993a,b) and Massey and Whitt (1997). In practice, it often suffices to act as if the service-time distribution were exponential, but that should be tested.

The second component of (4.1) shows that the mean is the integral of the arrival rate over a random interval before time t, specifically, over the interval $[t - S, t]$. The second formula can be interpreted as saying that PSA is correct except that $\lambda(t)$ should be replaced by an *average of the arrival rate in an interval before time* t, where the length of that interval should be about $E[S]$. That supports direct heuristic refinements of PSA proposed by Whitt (1991) and Thompson (1993). More importantly, the second formula also supports the notion of a time lag, showing that the extent of the lag is related to the service time S. (Here S appears instead of S_e, but the results are actually not inconsistent.)

Finally, the third component of (4.1), an integral, shows that the exact mean can be computed numerically, given the arrival-rate function $\lambda(t)$ and the service time cdf G.

An Idealized Mathematical Model. These methods apply to arbitrary arrival-rate functions, but one gains insight into system physics from a structured mathematical model that captures the spirit of typical arrival-rate functions. The dynamic character of the demand function is reasonably characterized by a *sinusoid*. The arrival rate per minute over the course of a 24-hour day, containing 1440 minutes, might be taken to be of the form

$$\lambda(t) = \bar{\lambda} + A \cdot \sin(\pi t/1440) \quad \text{arrivals per minute}, \quad 0 \le t \le 1440 \ , \tag{4.4}$$

so that there is one-half of a complete sinusoidal cycle (extending from 0 to π) over the day. In this model, there is a rise to a peak in the middle of the day and a decline to the original value, but no trough. That is roughly descriptive of Figure 1, as observed by Green, Kolesar and Soares (2001), and many other situations.

The sinusoidal arrival-rate function in (4.4), with its daily half cycle, fixes the *period* or the *frequency*, the *average arrival rate*, $\bar{\lambda}$, and the *amplitude* A. Instead of the amplitude, it is sometimes useful to focus on the *relative amplitude*, $A/\bar{\lambda}$.

In (4.4) a half sinusoidal cycle covers a full 24-hour day. Alternatively, we can parameterize the model to approximate other arrival-rate patterns found in practice. For example, in some settings (e.g., police patrol) there may be a full sinusoidal cycle within the day, including a decline to a trough at one time as well as a rise to a peak at another time. Some call centers have two peaks during the day, which might be roughly captured by having one and a half cycles over an six-hour period. Expressed differently, the three alternative scenarios suggest that a full sinusoidal cycle (2π) might occur over 48 hours, 24 hours or 4 hours. It should be understood that these idealized sinusoidal arrival-rate functions have been studied just to obtain insight; the actual arrival-rate functions would be used to determine staffing requirements.

A periodic arrival-rate function such as the sinusoidal function of (4.4), is consistent with the existence of a *dynamic steady state* for the $M_t/GI/s + GI$ model (with constant s); see Heyman and Whitt (1984) and references therein (although existence has only been rigorously proved in special cases). However, for practical purposes, the existence of a dynamic steady state and, more generally, the long-run average performance, are not too important, because in call centers (and many other service systems) service rarely extends beyond the day in which the service request arrives. Hence, we are primarily interested in the *shape of the arrival-rate function and the performance of the system within each day.*

However, to gain insight into *system physics*, Eick et al. (1993a,b) used the the $M_t/M/\infty$ model with a sinusoidal arrival-rate function in dynamic steady state (which is equivalent to assuming system operation began in the distant past), to derive closed-form analytical solutions. For the arrival-rate function in (4.4), the number of customers in the system at time t has a Poisson distribution with a time-dependent mean

$$\begin{aligned} m_\infty(t) &= \frac{\bar{\lambda}}{\mu} + \frac{A}{\mu(1+\gamma^2)}\left[\sin\left(\pi t/1440\right) - \gamma\cos\left(\pi t/1440\right)\right] \\ &= \frac{\lambda(t)}{\mu} - \frac{A}{\mu(1+\gamma^2)}\left[\gamma^2\sin\left(\pi t/1440\right) - \gamma\cos\left(\pi t/1440\right)\right] \ , \end{aligned} \tag{4.5}$$

where

$$\gamma \equiv \frac{\pi}{1440\mu} \ . \tag{4.6}$$

We now compare the exact mean $m_\infty(t)$ with three approximations for the $M_t/M/s_t + M$ model with $\theta = \mu$. Figure 4 illustrates the difficult case in which the mean service time is 5

hours. We let the overall average offered load be $\bar{a} = 100$, so that the average arrival rate is $\bar{\lambda} = 100/E[S] = 20$, and we let the amplitude be $50/E[S] = 10$.

The product of the arrival rate $\lambda(t)$ and the mean service time is the time-dependent offered load, which coincides with PSA: $m_{PSA}(t) = \lambda(t)E[S]$. We also consider a lagged version of PSA, in which the PSA is shifted by a mean service time: $m_{laggedPSA}(t) = m_{PSA}(t - E[S])$. Finally, we consider the exact value of $m_\infty(t)$ given in (4.5), which in this case coincides with the modified-offered-load approximation (MOL) to be discussed in the next section. Figure 4 clearly shows that there is both a time lag and magnitude shift in the peak of $m_\infty(t)$ compared to the peak of PSA. The lagged PSA captures the time lag well, but not the magnitude shift.

The explicit closed-form solution in (4.5) assumes that the system is in dynamic steady state. It is attractive because ir produces tractable expressions from which we gain insight. However, we would not make that assumption in practice, because it produces errors in the beginning of the day. We believe that the insights regarding the timing and magnitude of the peak are valid since experience has shown that convergence to the actual steady state usually occurs very quickly, usually within a few mean service times. To set staffing requirements in practice, one should work with the actual estimated arrival-rate function.

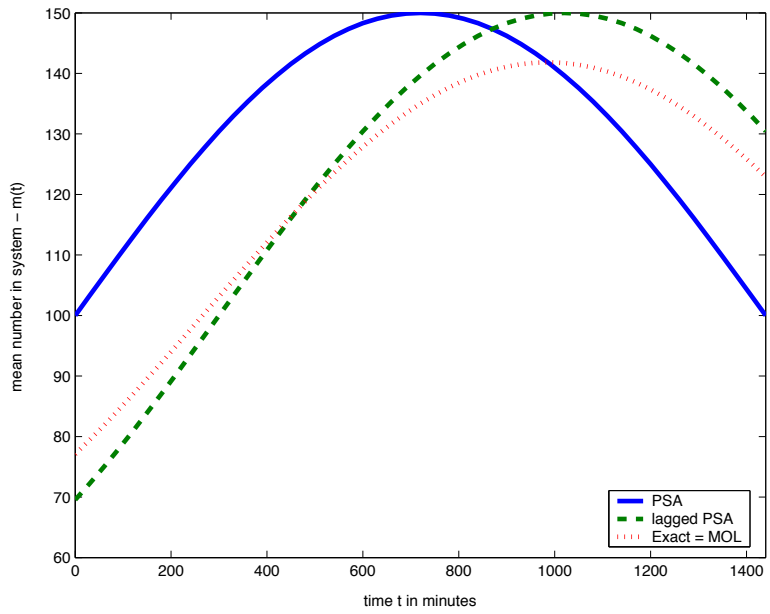

Figure 4: A comparison of PSA, lagged PSA and MOL for mean service time $ES = 300$ minutes (5 hours), with the sinusoidal arrival-rate function having $\bar{\lambda} = 100/ES = 1/3$, $A = 50/ES = 1/6$.

Taylor-Series Approximations. We can also obtain important insights without making such strong sinusoidal assumptions. The first representation $m_\infty(t) = E[\lambda(t - S_e)]E[S]$ in (4.1) is complicated since the random time lag S_e appears inside the general function $\lambda(t)$, inside the expectation. We could move the expectation inside to produce the deterministic time lag $E[S_e]$ if $\lambda(t)$ were linear and, more generally, we could directly express $m_\infty(t)$ in terms of moments of S_e if the arrival-rate function $\lambda(t)$ were a polynomial.

Of course, the arrival-rate function $\lambda(t)$ will usually not be a polynomial, but a smooth function can be approximately by polynomials in the neighborhood of individual arguments, by virtue of Taylor-series approximations. We proceed on this basis. Eick et al. (1993a) observed that the Taylor-series approximations provide great insight; also see Massey and Whitt (1997).

Suppose that we are interested in the performance at some time t. We can approximate the arrival-rate function in a time interval before time t by using a first-order Taylor-series approximation for $\lambda(t)$ centered at t:

$$\lambda(t - u) \approx \lambda(t) - \lambda^{(1)}(t)u \quad \text{for} \quad u \geq 0 \ , \tag{4.7}$$

where $\lambda^{(k)}(t)$ is the k^{th} derivative of $\lambda(t)$ evaluated at time t, from which we obtain from (4.1) the approximation

$$m_\infty(t) \approx \lambda(t - E[S_e])E[S] \ , \tag{4.8}$$

showing that $m_\infty(t)$ is approximately the PSA modified by the *deterministic time lag* $E[S_e]$, providing stronger support for the heuristic approximation introduced above.

Having confirmed that a natural approximation for the deterministic time lag is $E[S_e]$, we are even more interested in its value: $E[S_e] = E[S]((1 + c_s^2)/2)$. We see that $E[S_e]$ is itself well approximated by $E[S]$ when c_s^2 is near 1, which in most call centers appears to be the case. However, with medium-to-long service times, one should estimate c_s^2 as well as the mean $E[S]$, in order to incorporate the time-lag correction in staffing algorithms.

We can also consider a *second-order Taylor-series approximation* for the arrival-rate function $\lambda(t)$:

$$\lambda(t - u) \approx \lambda(t) - \lambda^{(1)}(t)u + \lambda^{(2)}(t)\frac{u^2}{2} \quad \text{for} \quad u \geq 0 \ , \tag{4.9}$$

from which Eick et al. (1993a, Theorem 9) obtain the approximation

$$m_\infty(t) \approx \lambda(t - E[S_e])E[S] + \frac{\lambda^{(2)}(t)}{2} Var(S_e)E[S] \ . \tag{4.10}$$

The first term in (4.10) is the first-order linear approximation given in (4.8), with the deterministic time lag, and the second term can be interpreted as a *deterministic magnitude shift*.

This quadratic approximation is approximately the PSA modified by *both* the deterministic time lag $E[S_e]$ and the deterministic magnitude shift

$$\frac{\lambda^{(2)}(t)}{2} Var(S_e) E[S] = \frac{\lambda^{(2)}(t)}{2} \left[\frac{E[S^3]}{3} - \frac{(E[S^2)^2}{4E[S]} \right] . \tag{4.11}$$

If time t is a local peak, then $\lambda^{(2)}(t) < 0$ and the magnitude shift will be negative, implying that the peak value of $m(t)$ will be below the peak value of $m_{PSA}(t)$, with the difference quantified by (4.11).

From (4.2), we see that the magnitude shift is directly proportional to the fluctuation of the arrival-rate function at time t, as measured by its second derivative $\lambda^{(2)}(t)$, and the variance of the service-time excess variable S_e. Experience shows that the magnitude shift does not matter much unless the service times are quite long.

An ODE for the Mean. With exponential service times, the infinite-server $M_t/M/\infty$ model becomes Markovian, and the mean $m_\infty(t)$ satisfies an *ordinary differential equation* (ODE):

$$m'_\infty(t) \equiv \frac{dm_\infty(t)}{dt} = \lambda(t) - \frac{m_\infty(t)}{E[S]} \tag{4.12}$$

for all t (Corollary 4 of Eick et al. 1993a), from which we can calculate $m_\infty(t)$ for any initial condition and any given (smooth) arrival-rate function. However, following Eick et al. (1993a), we emphasize important insights that can be gleaned from the ODE (4.12). Since $m_\infty(t)$ has an extreme point (maximum or minimum) at those points t for which $m'_\infty(t) = 0$, and (4.12) implies that $m'_\infty(t) = 0$ precisely at those time points t for which $m_{PSA}(t) \equiv \lambda(t)E[S] = m_\infty(t)$, *extreme points of the mean function occur where the mean function crosses the PSA mean function.* (This property holds in Figure 4.) This crossing property is useful to see whether the modeling and analysis are appropriate for a particular application, for if it fails, something is amiss. Low quality of service and a non-exponential service-time distribution are likely culprits.

In the common case where the arrival-rate function $\lambda(t)$ is unimodal with a single peak at $\hat{t}_\lambda$, the infinite-server mean $m_\infty(t)$ is also unimodal with a single peak $\hat{t}_m$, which occurs after $\hat{t}_\lambda$; see Corollary 2 of Eick et al. (1993a). We can thus represent the magnitude shift of the peak as

$$m_{PSA}(\hat{t}_\lambda) - m_\infty(\hat{t}_m) = m_{PSA}(\hat{t}_\lambda) - m_{PSA}(\hat{t}_m) \approx m_{PSA}(\hat{t}_\lambda) - m_{PSA}(\hat{t}_\lambda + E[S]) . \tag{4.13}$$

(Recall that $E[S_e] = E[S]$ for an exponential service-time distribution.) Thus, from the arrival-rate function and the mean service time alone, it is easy to estimate the magnitude shift of the peak.

4.3. The Modified-Offered-Load Approximation

The discussion above shows that the infinite-server $M_t/GI/\infty$ model provides insight into simple time-lag and magnitude-shift modifications of the PSA for the $M_t/GI/s_t+GI$ model. In particular, the infinite-server model helps us understand when the unlagged and lagged versions of PSA and SPHA will perform well. It also directly generates a full normal approximation and associated square-root staffing formula.

It is often possible to apply the infinite-server model as the *first step in a two-step procedure* to generate a better approximation for the time-dependent performance measures and the required staffing: This is the *modified-offered-load* (MOL) approximation, which also assumes that the system is never overloaded.

With MOL, we approximate the performance in the $M_t/GI/s_t+GI$ model at time t by PSA except we replace the instantaneous offered load $m_{PSA}(t) \equiv \lambda(t)E[S]$ by the exact infinite-server mean $m_\infty(t)$. In other words, we use a stationary finite-server $M/GI/s+GI$ model at each t with the "modified" time-dependent arrival rate

$$\lambda_{MOL}(t) \equiv \frac{m_\infty(t)}{E[S]} \ , \tag{4.14}$$

where $m_\infty(t)$ is the infinite-server mean in (4.1).

Example 4.1. *Comparing PSA, Lagged PSA and MOL.* Consider an example of the $M_t/M/s_t+M$ model with $\theta = \mu$ - for which the infinite-server approximation is exact. We compare three approximations for the time-dependent mean number of customers in the system: the PSA, the lagged PSA and the MOL. It is easy to see that the MOL approximation applied to the $M_t/GI/\infty$ model is exact, because the time-dependent number of customers in the system in both the stationary $M/G/\infty$ and nonstationary $M_t/GI/\infty$ models have Poisson distributions with common service-time distributions.

We assume a sinusoidal arrival-rate function, as in (4.4), where the average offered load is $\bar{a} = \bar{\lambda} \cdot E[S] = 100$ and the amplitude is $A = 50/E[S]$, and consider three different mean service times: (i) 300 minutes (5 hours) (ii) 30 minutes, and (iii) 3 minutes. We already displayed the results in the first case in Figure 4, and we plot the results for the second case in Figure 5. We see that both PSA and lagged PSA are clearly inadequate in Figure 4 when $E[S] = 300$ minutes. At the time of peak congestion or occupancy, which lags after the peak arrival rate, the lagged PSA errs by providing 10 extra agents, because it does not include the magnitude shift.

In Figure 5, where $E[S] = 30$ minutes, the curves for lagged PSA and the exact values of $m(t)$ fall on top of each, so clearly lagged PSA is sufficient; we also see the benefit of the lagged PSA over the plain PSA. To see the impact more clearly, we also plot the *difference* between the two functions in Figure 6. Note that the biggest discrepancies occur away from the peak; PSA overstaffs during the initial period of rising demand, but understaffs in the final period of declining demand, by as much as three agents. In the third case, where $E[S] = 3$ minutes,all three curves fall on top of each other, so that PSA is sufficiently accurate. ∎

The MOL approximation was first proposed by Jagerman (1975) for the time-dependent Erlang loss model $M_t/M/s/0$, but the approach generalizes to the $M_t/GI/s_t/r_t + GI$ model with time-dependent staffing s_t, possibly general time-dependent finite waiting room of size r_t and IID customer abandonments with a general time-to-abandon distribution. The MOL approximation for the time-dependent Erlang loss model was investigated further by Massey and Whitt (1994). They established bounds on the error and deduced when one time-dependent distribution stochastically dominates the other. The bounds imply that MOL is asymptotically correct when the arrival-rate function changes more slowly or when the load gets lighter. Those asymptotic results parallel previous results for PSA in Whitt (1991) and its generalization to the uniform acceleration asymptotic expansion (Massey 1981, Keller 1982 and Massey and Whitt 1998).

We have observed that the effectiveness of MOL approximation, just like PSA and lagged PSA, depends on the system not being too overloaded. A sinusoidal finite-server numerical example in Section 5 of Massey and Whitt (1997) shows how the approximations for the performance measures and lags degrade as the number of servers decreases. The tables there give a good sense for when it is appropriate to apply the MOL approximation. With a sufficiently high quality-of-service standard, there is little difficulty; MOL performs well.

The MOL approximation was applied to staffing in the $M_t/M/s_t$ and $M_t/M/s_t + M$ models by Jennings et al. (1996) and Feldman et al. (2004). They found that the MOL approximation is essentially equivalent to PSA and lagged PSA when they work well, and performs much better in the difficult examples with longer service times, when neither PSA nor lagged PSA performs well, as is illustrated by Figure 4.

End-of-the-day effects. The MOL approximation is also effective in coping with end-of-the-day effects. Data from actual systems frequently reveal anomalous behavior in the arrival-rate function, due to start-up and close-down effects. Clearly, simple time-lag refinements cannot

help at the beginning of an 8-hour working day if there is no arrival rate at all before the beginning of the day. Section 4 of Eick et al. (1993a) and Section 7 of Jennings et al. (1996) address staffing at the beginning of the day. This problem is taken care of by simply using MOL.

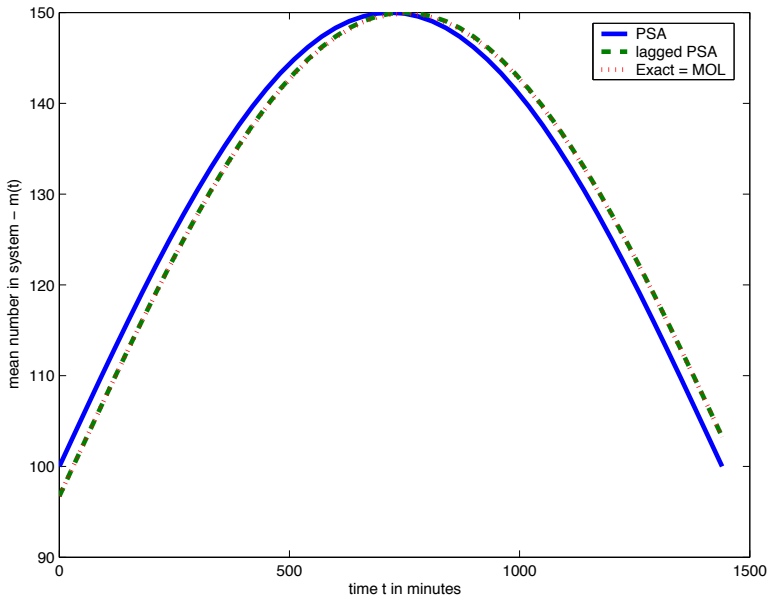

Figure 5: A comparison of PSA, lagged PSA and MOL for mean service time $ES = 30$ minutes, with the sinusoidal arrival-rate function having $\bar{\lambda} = 100/ES = 10/3$, $A = 50/ES = 5/3$.

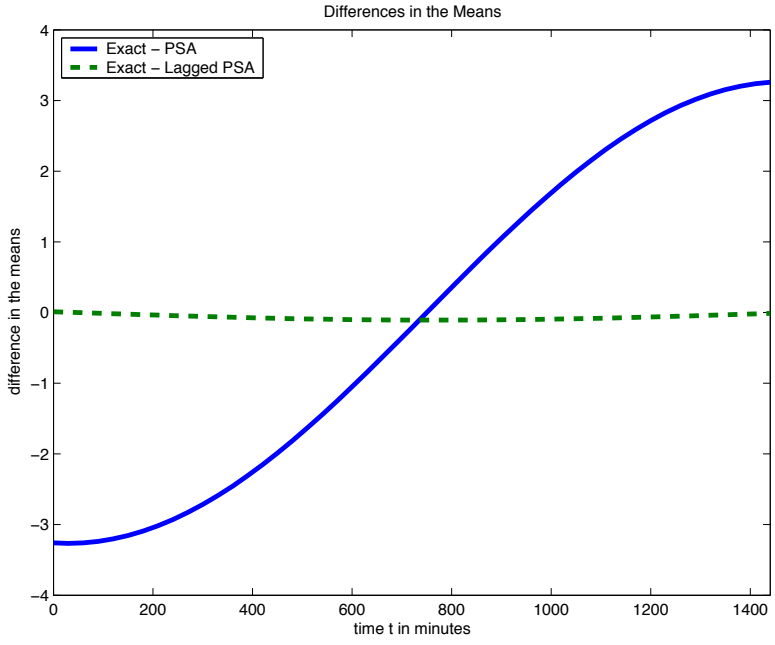

Figure 6: The differences between the exact mean number of servers and ,first, the PSA approximation and, second, the lagged PSA approximation, for the example in Figure 5.

4.4. A Simulation-Based Iterative Staffing Algorithm.

Feldman et al. (2004) developed a simulation-based iterative staffing algorithm (ISA) for the $M_t/GI/s_t+GI$ model. The ISA approach can be extended directly to more general models, for which analytic results are unavailable. It is also self-validating: that is, one can directly verify that the performance will be as desired (assuming of course that the simulation model itself is appropriate). It is substantially more computationally intensive than other approximations, but since the simulation is not done in real time, it may not be too great a burden.

The ISA ignores staffing intervals, and keeps staffing constant over small subintervals. It does a sequence of iterations, starting with an infinite-server system at iteration 0. (In practice, it suffices to pick a large finite number of servers.) Let $Q_n(t)$ be the number of customers in the system at time t in iteration n; and let $s_n(t)$ be the staffing function in iteration n, with $s_0(t) = \infty$ (or some large value) for all t. Given the staffing function $s_n(t)$ in the n^{th} iteration, we perform multiple (say 5000) independent replications of the full planning period (the day) in order to estimate the distribution of $Q_n(t)$, the number of customers in the system, at each time t. Given that estimated distribution of $Q_n(t)$, we then create a new staffing function $s_{n+1}(t)$, by choosing the value at each time t that just meets the specified delay-probability target at each time t:

$$P(Q_n(t) \geq s_{n+1}(t)) \leq \alpha < P(Q_n(t) \geq s_{n+1}(t) - 1) \ . \tag{4.15}$$

Having found the new staffing function $s_{n+1}(t)$, we simulate again to find the distribution of $Q_{n+1}(t)$ for each t. We continue to iterate until there is negligible change (e.g., at most a single agent) in the staffing function from one iteration to the next. For the special case of the $M_t/M/s_t + M$ model, Feldman et al. (2004) have proved that ISA (without estimation error) converges,

Feldman et al. (2004) show through experiments that the ISA is remarkably effective in achieving time-stable performance in face of time-varying demand, even with long service times. As with the earlier methods, the time-dependent performance constraints prevent the system from ever being overloaded, so that we stay in the class of manageable problems.

5. The Difficult Overloaded Case

In this section we briefly discuss the difficult case in which the system can be overloaded at times. This rarely occurs with properly managed call centers, but may occur more frequently with other service systems.

The Stabilizing Effect of Abandonments. If the targeted quality of service is not too high, it may happen that the arrival rate exceeds the maximum possible departure rate. In such circumstances, abandonments (unfortunate as they are) would prevent the buildup of congestion. This is illustrated by experiments reported in Feldman et al. (2004) in which the target delay probability was very high, e.g., 0.9, so that the resulting QoS parameter β is negative. Even then, because of abandonment, the staffing methods were effective in producing time-stable performance in face of strongly time-varying demand.

Whitt (2004b, 2005c) showed that deterministic fluid models can accurately describe performance in the stationary $M/GI/s+GI$ model when it is overloaded, and that it can be used as the stationary analysis along with PSA, SIPP and SPHA. In the fluid approximation, we think of arrivals coming deterministically at rate λ. Given that the system is overloaded, all s servers are constantly busy, and customers depart from the system deterministically at rate $s\mu$. Overloading means that $\lambda > \mu s$. Let the service-time cdf be G and time-to-abandon cdf be F.

In that fluid model, if customers are "reasonably" impatient, the approximate abandonment rate is just the arrival rate minus the total service rate, so that the abandonment probability is approximated by

$$P(Ab) \approx \frac{\lambda - \mu s}{\lambda} \ . \tag{5.1}$$

In the fluid model, all customers who do not abandon wait precisely a length of time w, where w satisfies the equation

$$F(w) = P(Ab) = \frac{\lambda - \mu s}{\lambda} \ , \tag{5.2}$$

where F is the time-to-abandon cdf. The total steady-state queue content is approximately the deterministic quantity

$$Q(w) = \lambda \int_0^w (1 - F(x))\, dx \ . \tag{5.3}$$

Numerical examples in Whitt (2005c) show that the deterministic fluid approximation is remarkably accurate in overloaded systems and that performance depends strongly upon the time-to-abandon cdf F but not upon the service-time cdf G. This is predicted by equations (5.1) and (5.2) above. Moreover, performance depends only on the time-to-abandon cdf F for small arguments - as is evident from equation (5.2). So we only need to know enough about the cdf F to determine the constant waiting time w, and that might be small, e.g., 20 seconds. As a consequence, the mean and other tail behavior of the time-to-abandon distribution play no role.

It is easy to set staffing requirements with the fluid model, assuming only moderate quality of service: For example, suppose that we require that 80% of all calls be answered within 30 seconds (1/2 minute) and that at most 5% of all arrivals should abandon. With the fluid model, the optimal staffing level is

$$s* = \frac{\lambda(1 - x*)}{\mu}, \quad \text{where} \quad x* \equiv \min\{F(1/2), 0.05\} \ . \tag{5.4}$$

If $F(1/2) < 0.05$, then the service-level constraint is binding, all served customers wait precisely 1/2 minute, and the abandonment probability is $F(1/2) < 0.05$. On the other hand, if $F(1/2) > 0.05$, then the abandonment constraint is binding, exactly 5% of the arrivals abandon, and all served customers wait precisely $w < 1/2$ minute, where $F(w) = 0.04$.

Note that the fluid staffing equation in (5.4) involves the "fluid" scaling, in which the staffing level is a fixed percentage of the offered load, unlike the square-root-staffing formula in (3.5) and the many-server heavy-traffic limiting regime in (3.7).

The deterministic fluid model provides such a tractable description of performance that we can do back-of-the-envelope calculations. Whitt (2005e) shows how the deterministic fluid model can help set staffing requirements in face of uncertain arrival rate and absenteeism. The deterministic fluid model also provides insight into sensitivity of performance to the model parameters; see Whitt (2005d). Whitt (2005c) also developed an algorithm to calculate the performance in a corresponding *time-varying* $M_t/GI/s + GI$ fluid model in a discrete-time setting. That discrete-time algorithm can be applied to calculate approximate performance descriptions in $M_t/GI/s_t + GI$ queueing models, which can be useful if the system is heavily loaded.

The Genuinely Overloaded Case. A genuinely overloaded case is most likely to arise when the arrival rate changes substantially during a long staffing interval, we tolerate only good *average performance*, and there is negligible customer abandonment. The good average performance can be achieved, at least on paper, even though the system is overloaded at some times, by averaging over periods of understaffing and overstaffing within the long staffing interval. Even though the average performance may be judged acceptable, there will be a buildup of congestion over subintervals, which is not well described by the previous methods in this paper.

Since all the approximations introduced so far fail badly in the genuinely overloaded case. it may be necessary to perform detailed calculations for the $M_t/GI/s_t + GI$ model, which

in turn may force us to make additional Markovian assumptions. One approach that is very effective is the numerical solution of *ordinary differential equations* (ODE's), as was employed for the $M_t/M/s_t$ model by Green, Kolesar and colleagues in their sequence of papers.

However, the numerical solution of ODE's is computationally intensive, so that it is useful to have more elementary approximations, such as the *closure approximations* of Rothkopf and Oren (1979), Ong and Taafe (1987) and others. Closure approximations solve the system by working with a smaller set of approximating differential equations; e.g., for the time-dependent mean and second moment only. These methods produce approximations paralleling the simple ODE in (4.12) that applies exactly for the infinite-server model. Such smaller sets of ODE's also can be obtained from *fluid and diffusion approximations* (Mandelbaum, Massey and Reiman 1998).

Severe overloading dominates other phenomena, so we may capture the essentials of performance by using methods that focus directly on the overloading. In particular, deterministic fluid models are remarkably effective for describing the performance in face of severe overloading (Sections 6.4 and 6.5 of Hall 1991 and Chapters 2 and 9 of Newell 1982).

Example 5.1. *Extreme overloading in Figure* 1. To illustrate how to work with a deterministic fluid model and how to make stochastic refinements when faced with severe overloading over an extended period of time, we now consider an artificial example based on the financial-services arrival-rate function in Figure 1. As before, we use a mean service time of 6 minutes = 1/10 hour. Consequently, the departure rate is $10k$ per hour when k customers are in service. The system will be severely overloaded if there is no abandonment at all and the staffing level is fixed at say $s = 180$ agents across the entire day. That makes the maximum possible total departure rate $s\mu = 1800$ calls per hour, which is less than the arrival rate for a long part of the day. Such extremely overloaded conditions are not typical of well-managed call centers, but might occur in some contexts - say if each agent is very costly or if demand has grown faster than new agents can be hired and trained.

This overloaded scenario has a significant buildup of congestion because the arrival rate first exceeds 1800 calls per hour at 7:30 am and stays above 1800 until 2:30 pm. The buildup of congestion is well described by a deterministic fluid analysis, in which both the arrival and service processes are treated as deterministic at the specified rates. The deterministic fluid approximation predicts a buildup of congestion from 7:30am until 2:30pm and then a gradual dissipation thereafter, as shown in Figure 7.

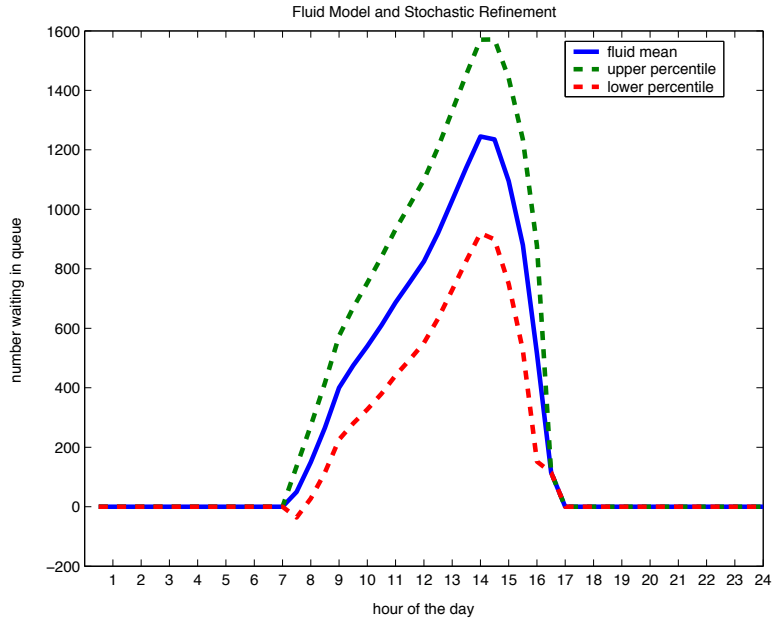

Figure 7: Fluid description of congestion buildup and subsequent recovery in an overloaded financial-services call center with arrival rate according to Figure 1 and a fixed staff of 180 agents and service rate of 10 per hour. Also shown are the 95% tails of the distribution based on a normal approximation.

It is also useful to apply a crude stochastic refinement to estimate the reliability of the fluid approximation. To do so, we observe that, during the overloaded period starting at 7:30am, the arrival process and departure process act approximately as two independent nonhomogeneous Poisson processes. Thus, after 7:00am, the queue length at the k^{th} half hour is approximately normally distributed with mean $\sum_{i=1}^{k}(\lambda_i - 900)$ and variance $\sum_{i=1}^{k}(\lambda_i + 900)$, where λ_i and 900 are the mean number of arrivals and departures, respectively, in the half-hour periods. We plot the time-dependent fluid mean and associated approximate 95% tails of the distributions (upper and lower bounds) in Figure 7.

Figure 7 indicates that the system will recover from congestion (the queue will first empty) at about 5:00pm. Allowing for stochastic fluctuations, the first-emptiness time should be somewhere between 4:30pm and 5:30pm. The expected waiting for an arrival at time t would be the displayed mean queue length divided by the maximum possible total service rate, $\mu s = 1800$ calls per minute. Thus, with this severe overloading, the maximum waiting time would be about 42 minutes, which occurs at about 2:00pm. The fluid approximation for the average waiting time among arriving customers is substantially less; it can be approximated by

$$\bar{W} \approx \frac{\sum_{i=1}^{48}(\lambda_i Q_i/1800)}{\sum_{i=1}^{48}\lambda_i} ,$$

where Q_i is the approximate queue length in half-hour period i. Simulations can confirm that

this is a quite accurate description. ∎

These overloaded situations have also received more mathematical attention via heavy-traffic analyses; e.g., see Section 5.3 of Whitt (2002). For other work in this direction, see Massey and Whitt (1994b), Choudhury, Mandelbaum, Reiman and Whitt (1997) and Choudhury, Lucantoni and Whitt (1998).

6. Other Applications

In this section, we discuss how the methods for treating nonstationary demand apply to service systems other than call centers. We first mention two categories of service systems that most resemble call centers, to which the classic PSA approaches - SIPP and SPHA - apply rather directly. We then discuss two other types of service systems which present distinct challenges.

6.1. Two "Easy" Service Systems

Airline airport support services and teller systems in banks both have short service times, short staffing intervals and high quality-of-service standards. Therefore, they are similar to many call centers and fall in the "easy" category discussed in this paper.

Airline Airport Support Services. Airlines use queueing theory to determine staffing requirements throughout the day for several categories of airport personnel (Holloran and Byrn 1986, Schindler and Semmet 1993 and Brusco et al. 1995). As is in call centers, 15-minute staffing intervals are often used. For example, staffing requirements must be set for the ticket-counter agents who are responsible for customer and baggage check-in, ticketing, and seat assignments. Arrival-rate functions, which have a strong time-varying component, are estimated based on flight departures and *contact ratios*, which provide estimates of the percentage of customer arrivals that will require service at individual ticket-counter queues (e.g., domestic, first class, etc.). Abandonments are essentially nonexistent, but special priority treatment may be given to customers with tight deadlines. The priority treatment need not be considered directly in the staffing, however.

Expected service times vary by the function performed and by time of day, and are generally short (a few minutes), but it is not evident that the distribution of these service times has been carefully studied. Unlike most call centers, the offered loads are relatively small, so staffing levels are much lower than in call centers. Given the need to insure that passengers make

their flights, the service standards, which also vary by type of queue and time of day, are fairly high (e.g., 85% of passengers served within 5 minutes), but clearly not as high as in most call centers. A SIPP approach is commonly used to determine staffing requirements.

A complication in setting staffing requirements for airline airport support services is the uncertainty about demand, due to disruptions in the planned airplane schedules, primarily due to bad weather conditions, locally or at distant airports. Of course, airlines do pay careful attention to weather forecasts, but significant uncertainty remains. Future research should assess and address the uncertainty in demand. When the uncertainty about demand is high, there is benefit from being able to make adjustments to staffing upon relatively short notice.

Teller Systems in Banks. Financial institutions introduced queueing-based teller staffing models in the late 1960s and early 1970s primarily to control increasing labor expenses (Brewton 1989). Until the introduction of automated teller machines (ATMs), transactions such as deposits, withdrawals, and cash-checking were handled exclusively by human tellers. Although automated banking and on-line banking have decreased the need for human tellers, many retail banks still rely on them to provide timely and personalized customer service. Staffing models are used in determining appropriate teller levels over the day. As with the airline ticket counters, these systems tend to be small, e.g., fewer than 10 servers. Service times are often calculated by identifying all routine transactions handled in the teller line, such as cashing checks or withdrawing savings, and collecting data on the respective time required to process each transaction type, the number of each type, and the number of transactions per customer over a given time period such as a month. These data are automatically recorded. Service times are short and service levels are fairly high. Staffing intervals are typically one hour. So again, SIPP staffing is likely an adequate approach for these systems.

For this application, a serious effort is currently being made to collect and analyze data, as illustrated by Brown et al. (2005). Previously, there was precious little data available to researchers.

Automatic Teller Machines (ATM's). ATM's present a somewhat different situation. There is no ATM staffing decision per se, but determining the proper capacity at a location - the number of ATMs there - is an important design decision and the critical issue is peak-hour congestion. In a detailed study of ATM operations, which included empirical data analysis, queueing modeling and development of managerial recommendations (Kolesar 1984), the goal

was to identify which 100 of some 600 urban ATM locations would benefit most from having additional machines. Since average service times are short, about one minute, PSA modeling works well.

Yet, this ATM study demonstrated that capacity specification could not be based on the customary delay calculations within the context of the classical Erlang-C model ($M/M/s/\infty$): There appeared to be little congestion from a delay viewpoint, but at many ATM locations the peak-hour counts of completed transactions were clearly at the physical capacity of the system and the arrival rate function appeared to be truncated. ATMs are often placed in vestibules of limited size and customer balking and customer abandonment are common. Based on both the extensive transaction data and somewhat limited field observations during peak hours, it appeared that so many customers were balking or abandoning the queue that the delay to those who remained was quite small. The actual data on balking and abandonments was very limited, since unlike call centers where each abandonment is captured and recorded electronically, attempted arrivals were not logged electronically. Also, unlike call centers where customers cannot see the length of the queue, arrivals to an ATM are likely to balk or abandon if the queue is too long, behaving somewhat as if there is a finite queue capacity. So, in order to refocus the bank management's thinking onto the more appropriate service measure of lost or inconvenienced customers, a finite-capacity M/M/s/k model was used to impute peak-hour abandonments at the 600 ATM locations. (Service times - actually occupancy times - which were considerably longer than computer recorded transaction times, were not exponential. However, predictions from exponential service-time models were shown to be sufficiently accurate in simulations.) These imputed peak-hour abandonments became the measure that guided the ATM investment decisions.

For ATM applications, future research could directly focus on customer balking by estimating the *state-dependent* (as well as time-dependent) arrival-rate function, e.g., with the aid of modern surveillance technology. An appropriate model might be the $M(n)/M/s_t + M$ queue, which has state-dependent arrival-rate. The stationary version $M(n)/M/s + M$ can be analyzed just like the $M/M/s + M$ model, because the number in system over time is still a birth-and-death process.

6.2. Police Patrol

The repetitive time-varying arrival pattern that has motivated and driven the methodological discussions of this paper is also found in police patrol systems, as illustrated by Figure 8, which

graphs the hourly average number of police responses to 911 emergency telephone calls for two New York City Police Department (NYPD) precincts. The police emergency call-and-response

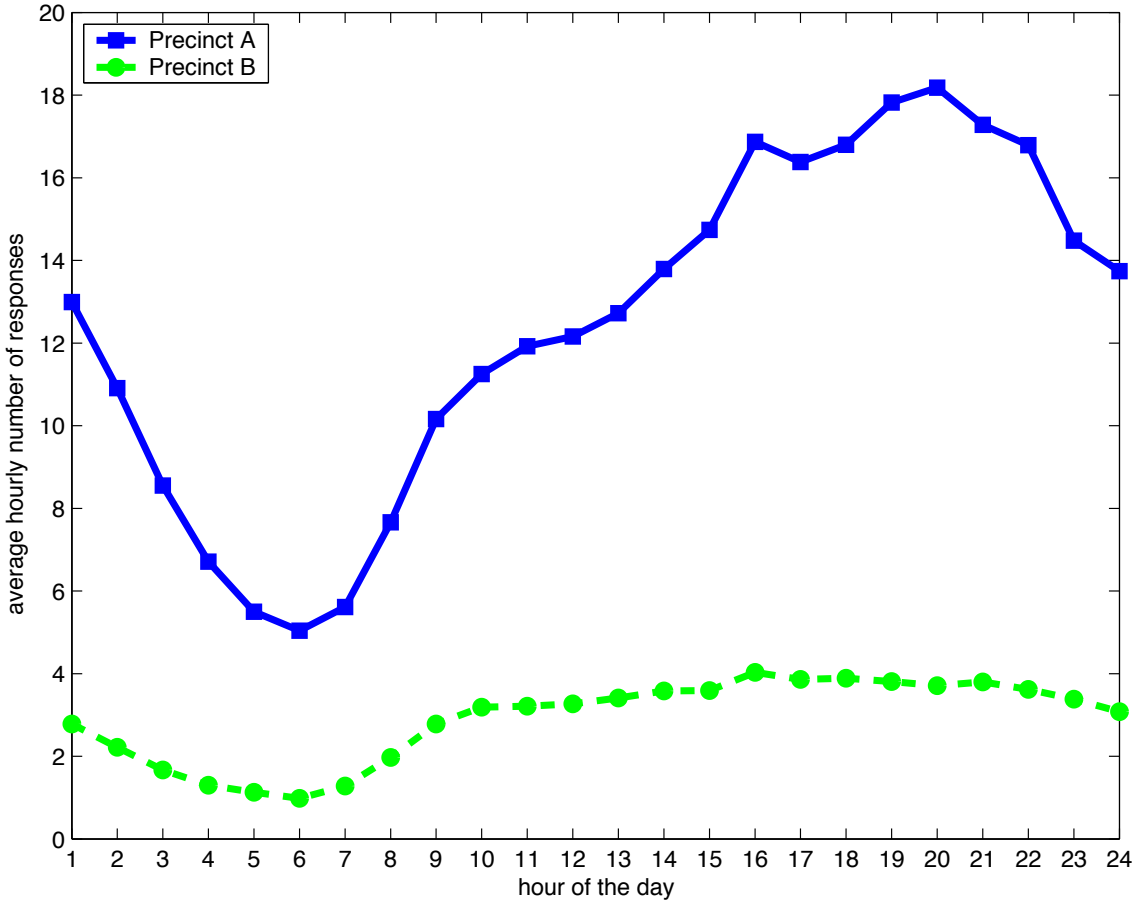

Figure 8: Average hourly number of police patrol emergency responses: two New York City precincts, 1999 to 2003.

system is a complex queueing network. Incoming 911 telephone calls are first screened by a bank of telephone operators, who pass them on to police dispatchers, each of whom handles a specific geography. Here the "jobs," as they are called in police jargon, are queued until the dispatcher identifies available and appropriate patrol cars. Queueing analysis has long been used at several stages of this process. We focus on the patrol-staffing problem and will show that it has characteristics that make the application of the previously described methods not straightforward.

Patrol force staffing is a complex and serious problem for large urban police departments as the cost, service and public-safety impacts of poor staffing can be substantial. Long delays to high-priority calls are simply not tolerable. Green and Kolesar (1984b,1989) describe the complexity of communications, decision-making and operations in the field - and how they can

be modelled.

Police patrol falls into our category of difficult problems: Average patrol-car service times are quite long - about 30 minutes, and the staffing intervals used by most police departments are also long - typically 8 hours. To make matters more complex, many calls for emergency service require more than one server (patrol car) at a time. In addition, calls for police service must be prioritized - a simple split being between crimes in progress versus all other jobs. Unlike call centers, patrol-car staffing is determined for individual command areas (e.g. precincts) which are relatively small scale. Figure 8 contains data for one of the heaviest (Precinct A) and lightest (Precinct B) call volumes among NYPD's 77 precincts. Assuming a half-hour average job service time, these precincts have average offered loads of 6.0 and 1.4, respectively.

Police patrol in most cities is, in our parlance, a "data rich" environment. All calls and assigned jobs are logged electronically, as are the identities and work times of the responding patrol cars, as well as the job's location and its final disposition. Thus, as in call centers, ample data is available for analysis and modeling. But, there is an important distinction and difficulty: The police-patrol world is not nearly as tightly managed as is the call-center world (Rosengrant 1993, Green and Kolesar 1984a,b, 1987, 1989.) Patrol cars are typically distributed over large geographical areas. They "patrol," that is, cruise around their sectors, and operate with a good deal of discretion and freedom of action. In contrast to many call centers - in which any service call can be monitored directly by several layers of management and managers may sit in a booth overlooking the work floor while tracking real time electronic displays of system status and recent queuing statistics - first-line supervision of police patrol is typically not intense and, in some jurisdictions, is really rather lax. The person with the most information about the state of the patrol system is arguably the dispatcher, who has no supervisory authority. Police unions are often politically powerful and a strong "anti-boss" culture frequently prevails among the front-line responders. In such a culture, work rules and standard operating procedures are merely taken as "suggestions." Thus, while there are rules and policies that can be modelled, actual behavior of the police on patrol may deviate substantially from the model in ways that are difficult to capture. All this makes the modeler's work challenging (Green and Kolesar 1987).

The Patrol Car Allocation Model (PCAM). Notwithstanding these problems, there has been remarkable success in using queueing to determine patrol staffing. In the late 1970's the RAND Corporation developed the *Patrol Car Allocation Model* (PCAM) (Chaiken and

Dormont 1978a, b, Chaiken et al. 1985), which has been used by dozens of cities across the nation. The scope of the patrol-staffing problem in New York City is illustrated by the fact that the city has some 77 police precincts, each of which has a population and number of patrol cars corresponding to a small city. Patrol cars must be allocated across this geography and across the three tours of duty. Unfortunately, the NYPD does not use the PCAM model on a daily or weekly basis, but they do use it annually to allocate some thousand new police recruits across the precincts upon their graduation from the police academy (Green and Kolesar 2004).

The most recent version of PCAM uses an $M/M/s_t$ model that explicitly incorporates the multiple-server-per-job aspect of police patrol and a non-preemptive priority structure (Green 1984). It also explicitly deals with the nonstationarity of demand via a SIPP approach. Though the Green queueing model improves substantially on the modified M/M/s model it replaced (Green and Kolesar 1984a) and has been shown to capture the essential nature of police patrol car behavior, there is room for improvement over PCAM's SIPP approach, given the long average service times and staffing intervals (Green and Kolesar 1989).

The impact of nonstationarity on police-patrol performance is significant. From Figure 8 we see that call rates drop to a minimum near 4 a.m. and then rise to their maximum about 10 p.m. and the relative amplitude is about 50% of the mean arrival rate. This dramatic rise and fall, coupled with very long staffing periods, makes the problem tough. Staffing for the average demand over the tour gives inadequate response, while staffing for the peak during the tour is uneconomical. The traditional tour designs and starting times - midnight to 8 a.m; 8 a.m to 4 p.m.; and 4 p.m. to midnight - do not align well with actual staffing needs. These factors lead to severe congestion at times, especially for the non-priority calls. This situation can be particularly bad at the end of the midnight tour, when staffing drops dramatically while the call rate is still high. This can lead to perverse behavior on the part of individual patrol cars. Cars going off duty at midnight apparently let jobs queue up to be handled by the next shift.

There are several possible approaches to deal with issues arising from the time-varying demand. Queueing analysis coupled with linear programming has been used to find tour starting times that better match the demand pattern, but the potential gains are limited as long as the 8-hour tour structure is maintained (Kolesar et al 1975). Overlay tours (sometimes called fourth platoons, Moore et al 1975) are another remedy used by some departments, in effect, creating shorter staffing intervals. The evaluation of these strategies can be aided by simulation. A detailed simulation model of police patrol designed to incorporate these characteristics as well as the specifics of geography, street patterns and inter-precinct patrol-

car-assignment strategies (Kolesar and Walker 1975) was used to validate the findings described from the queueing/linear-programming analysis of staffing in Kolesar et al (1975).

Given the underlying multiserver structure of this patrol-car queueing model, and the results demonstrated earlier in this paper for nonstationary versions of the Erlang and M/G/s model, it seems that better staffing levels could be identified by incorporating a time-lag into the current SIPP use of the Green police model or by doing other refinements, such as the modified-offered-load approximation. Yet, this is an unproven conjecture.

A key issue in police staffing is evaluating the right trade-off between understaffing - based on staffing according to the average arrival rate over a tour - and overstaffing - based on a peak-hour criterion. Clever combinations of new tour start times and break scheduling could improve performance. One might use MOL to find the appropriate staffing levels for each hour and then a combination of optimization and simulation to find appropriate tour start times and break times to get as close as possible to the MOL suggestion.

6.3. Hospital Emergency Departments

Timely access to an emergency provider is a critical dimension of quality for emergency departments (ED). Unfortunately, hospitals often struggle to provide adequate staffing to handle unpredictable demands for care that vary substantially over the day; Figure 9 shows the time-varying demand pattern.

Hospital managers, while aware of the variability over the day, have not used queueing models, but instead allocate staff based on general perceptions and intuition. One reason is a lack of data. Though arrival data is collected through the patient registration process, there are no systems in place to collect physician "service" times, which include taking patient histories, physical examinations, ordering and reading test results, consulting with other physicians, administering treatments, and writing up reports. Furthermore, these service times are very difficult to capture since they are often discontinuous due to the frequent need for the physician to order and wait for test results before making a discharge or treatment decision. While these tests are being done, physicians are able to handle other patients.

Abandonments, which are called *left without being seen* or LWBS, are an important feature of these systems, and can be 10% or more during times of peak congestion. While overall levels of LWBS are often documented, the exact times when abandonments occur are not known since they are only identified when a nurse fails to find the patient to begin examination and treatment. Another important characteristic of EDs is the use of triage to prioritize patients

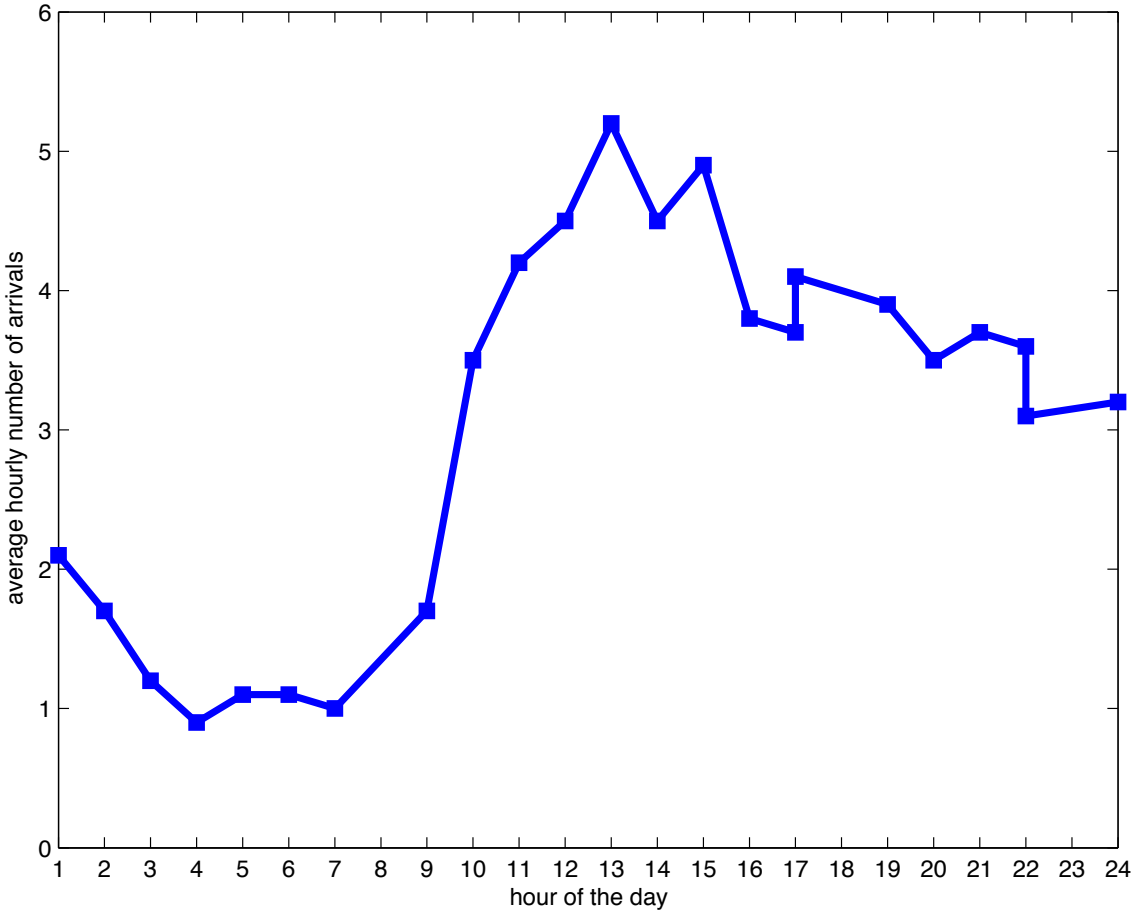

Figure 9: Arrivals per hour to a New York City Emergency Department.

into emergent, urgent and non-urgent categories.

Even in this data-poor and ill-defined service system, queueing models can be helpful in identifying better staffing levels. Green et al. (2005) describes a study of the physician staffing in an urban hospital ED that had high levels of congestion. Staffing levels, which were varied to roughly reflect the changing arrival rate, ranged between 1 and 3 physicians, making this a very small system. Arrivals were automatically recorded and a statistical analysis yielded strong support for the assumption of a time-varying Poisson arrival process. Staffing intervals were 2 hours long. As in most EDs, there was no system for collecting or estimating physician service times and no resources available to conduct detailed time-and-motion studies to capture them. Based on the experience of the ED director and data on the number of patients seen per physician during each shift, average service times were estimated to be about 30 minutes.

The only available performance data related to physician staffing was the fraction of patients who left without being seen, which has been shown to be highly correlated with ED delays

(Fernandes, et al. 1997, Green, Wyler and Giglio 2002). Assuming continuous exponentially distributed service times, a Lag SIPP approach was used to determine staffing levels needed to achieve a service level of 80% of patients experiencing a delay of no more than one hour – consistent with the time standards used in the National Hospital Ambulatory Medical Care Survey (McCaig and Burt 2004). The 80% criterion reflects the approximate percentage of emergent and urgent arrivals. Due to some substantial differences in arrival volumes across the days of the week, analyses were done to determine staffing requirements for individual days as well as for aggregated weekdays and weekends.

To gauge the potential benefit of using a lagged arrival function, the non-lagged SIPP approach was also used and the two suggested staffing functions were compared. Computations for both models involved numerically integrating the resulting ODEs and computing delays – the tail probabilities (Green and Soares 2005). The results confirmed that while SIPP's staffing levels often violated the target, Lag SIPP always identified levels that met the service target in each staffing period.

Both methods, however, suggested staffing levels that were higher than could be met with the available physicians, particularly for the high-volume weekdays. By looking at the delay probabilities, it was possible to identify some reduction in the staffing levels suggested by Lag SIPP, which still maintained the overall targeted performance. Though these adjusted staffing levels were still higher than the current physician availability, the modeling results were used to switch some staffing hours from the weekends to the weekdays and to obtain a 3% increase in hours. These changes minimized the disparity between the model-based recommendations and the actual staffing levels that were implemented as a result of the study. Comparing the actual volumes of LWBS for the same 9-month period before and after these staffing changes were implemented demonstrated an overall reduction in the fraction of LWBS by 23%. This improvement was achieved despite an increase in patient visits of over 6% between the before and after study periods.

Hospital EDs present many opportunities for further research. For one thing, the true nature of service times remains unclear: Experience indicates that the critical resource is usually the physicians, so the the service time should be the required physician treatment time. However, sometimes a single physician is serving more than one patient at a time, causing service interruptions for any one patient. What do we do about these service interruptions?

7. Concluding Remarks

In this paper we have covered a lot of ground, discussing several different techniques and several different service systems, but the situation is not difficult to summarize: *Setting staffing requirements tends to be manageable provided the $M_t/GI/s_t + GI$ model roughly fits and there is a high quality-of-service standard.* With long staffing intervals, that means using a peak-congestion criterion, as discussed in Section 3.2.

Assuming that the $M_t/GI/s_t + GI$ model is appropriate, *with a high quality-of-service standard, the modified-offered-load (MOL) approximation should be consistently effective.* After using MOL to set the staffing requirements at each time t, adjustments can be made to account for staffing intervals.

In many applications, it suffices to use more elementary methods, in particular, lagged versions of the basic PSA techniques - SIPP and SPHA, where accuracy can be confirmed by calculating the MOL approximation. Indeed, in many applications, it suffices to use the traditional SIPP and SPHA methods. These traditional methods - without lagged refinements - will be suitable if the service times are short (e.g. 3 minutes), while lagged refinements will be suitable if the service times are somewhat longer (e.g., 30 minutes), and MOL itself is needed if the service times are long (e.g., 300 minutes), as was illustrated in Example 4.1. When MOL does not agree with PSA or lagged PSA, it is better to use MOL. The MOL approximation is also important for treating anomalous behavior such as end-of-the-day effects.

When the targeted quality of service is not high, the service system is likely to experience substantial congestion during portions of the day. It is difficult to predict the human response to such congestion; the phenomena of balking, abandoning and retrying begin to play a critical role. All this makes management much more complicated. In Section 5 we described some analysis techniques that can be used with lower targeted quality of service, but the situation is inherently more difficult. *Providing a good quality of service not only pleases customers, but it is easier to manage.*

We conclude by discussing extensions to cope with time-varying demand when other complicating features are present, pointing to relevant literature.

Network Structure. In some service systems the customers may move through a complex network of queues, receiving several component units of service, before departing. The infinite-server analysis described in this paper has been generalized to networks of infinite-server queues

with time-varying demand (Massey and Whitt 1993 and Nelson and Taaffe 2004a,b). One conceptually easy network-approximation method applies the results for infinite-server networks given in these references, and then staffs using the MOL approximation at each queue separately, using the infinite-server network to generate the appropriate infinite-server offered loads.

Extensions also apply to other kinds of stochastic networks (Abdalla and Boucherie 2002, Jennings and Massey 1997, Leung et al. 1994, Massey 2002 and Massey and Whitt 1994c).

Multiple Customer Classes. Many modern call centers have multiple customer classes and skill-based routing. Thus it is important that the methods in this paper extend to that setting. The MOL approximation can be used in a first step to derive an appropriate arrival-rate function $\lambda_{MOL}(t)$ for each customer class, just as in (4.14). In this way we reduce the multi-class staffing problem with time-varying demand to the multi-class staffing problem with stationary demand. Just as in the single-class case, we can use a SIPP or segmented-PSA approach with a stationary model in each staffing interval. Then staffing algorithms such as in Wallace and Whitt (2005) can be applied. Wallace and Whitt show that the total number of agents can be nearly the same as with a single class when there is a limited amount of cross training.

New methods for treating multiple customer classes and skill-based routing have been proposed by Harrison and Zeevi (2004) and Bassamboo et al. (2005a,b). They focus on both the uncertainty in the arrival rate and its variation over time. Bassamboo et al. (2005a,b) obtain important model simplification by combining PSA and fluid approaches.

Retrials. We have ignored retrials. In some settings it may be important to consider them (Abdalla and Boucherie 2002, Aguir et al. 2004 and Hoffman and Harris 1986). Some of the many-server approximations apply there as well (e.g., Grier et al. 1997, Mandelbaum et al. 1998, 1999 and Massey 2002).

The Final Word. We end by reiterating that much remains to be done, especially for challenging service systems such as police patrol and hospital emergency departments.

References

[1] Abdalla, N., R. J. Boucherie. 2002. Blocking probabilities in mobile communications networks with time-varying rates and redialing subscribers. *Annals of Operations Research* 112 (1), 15–34.

[2] Aguir, M. S., F. Karaesman, O. Z. Aksin, F. Chauvet. 2004. The impact of retrials on call center performance. *OR Spectrum* 26 (3), 353–376.

[3] Anton, J., V. Bapat, B. Hall. 1999. *Call Center Performance Enhancement Using Simulation and Modeling*. Purdue University Press.

[4] Bassamboo, A., J. M. Harrison, A. Zeevi. 2005a. Dynamic routing and admission control in high-volume service systems: asymptotic analysis via multi-scale fluid limits. *Queueing Systems*, forthcoming.

[5] Bassamboo, A., J. M. Harrison, A. Zeevi. 2005b. Design and control of a large call center: asymptotic analysis of an LP-based method. *Operations Research*, forthcoming.

[6] Bear, D. 1980. *Principles of Telecommunication-Traffic Engineering*, Institution of Electrical Engineers, Peter Peregrinus.

[7] Bolotin, V. 1994. Telephone circuit holding-time distributions. J. Labetoulle, J. W. Roberts, eds., *Proceedings of International Teletraffic Congress, ITC 14*, North-Holland, Amsterdam, 125–134.

[8] Borst, S., A. Mandelbaum, M. I. Reiman. 2004. Dimensioning large call centers. *Operations Research* 52 (1), 17–34.

[9] Brewton, J.P. 1989. Teller staffing models. *Financial Managers' Statement* 11 (4), 22-24.

[10] Brigandi, A.,D. Dargon, M. Sheehan, T. Spencer, III. 1994. AT&T's call processing simulator (CAPS): operational design for inbound call centers. *Interfaces* 24 (1), 6–28.

[11] Brusco, M.J., L.W. Jacobs, R.J. Bongiorno, D.V. Lyons, B. Tang. 1995. Improving personnel scheduling at airline stations. *Operations Research* 43 (5), 741-751.

[12] Brown, L., N. Gans, A. Mandelbaum, A. Sakov, S. Zeltyn, L. Zhao, S. Haipeng. 2005. Statistical analysis of a telephone call center: a queueing-science perspective. *Journal of the American Statistical Association* (JASA) 100 (1), 36–50.

[13] Chaiken, J.M., P. Dormont. 1978a. A patrol car allocation model: background. *Management Science* 24 (12), 1280–1290.

[14] Chaiken, J.M., P. Dormont. 1978b. A patrol car allocation model: capabilities and algorithms. *Management Science* 24 (12), 1291–1300.

[15] Chaiken, J., R. Larson. 1972. Models for allocating urban emergency units: a survey. *Management Science* 19 (4) 110–130.

[16] Chaiken, J., W. Walker, and P. Dormont. 1985. Patrol car allocation model: executive summary. Rand Corporation, R-3087/3-NIJ, July 1985.

[17] Choudhury, G. L., K. K. Leung, W. Whitt. 1995. Efficiently providing multiple grades of service with protection against overloads in shared resources. *AT&T Technical Journal* 74 (4), 50–63.

[18] Choudhury, G. L., D. M. Lucantoni, W. Whitt. 1998. Numerical Solution of Mt/Gt/1 Queues. *Operations Research* 45 (3), 451–463.

[19] Choudhury, G. L., A. Mandelbaum, M. I. Reiman, W. Whitt. 1997. Fluid and Diffusion Limits for Queues in Slowly Changing Random Environments. *Stochastic Models* 13 (1), 121–146

[20] Cleveland, B., J. Mayben. 1997. *Call Center Management on Fast Forward*, Call Center Press, ICMI, Annapolis, MD.

[21] Cooper, R. B. 1982. *Introduction to Queueing Theory*, second edition, North Holland.

[22] Dantzig, G. B. 1954. A comment on Edie's "Traffic delays at toll booths." *Operations Research* 2 (3), 339–341.

[23] Davis, J. L., W. A. Massey, W. Whitt. 1995. Sensitivity to the service-time distribution in the nonstationary Erlang loss model. *Management Science* 41 (6) 1107–1116.

[24] Edie, L. C. 1954. Traffic delays at toll boooths. *Operations Research* 2 (2), 107–138.

[25] Eick, S., W. A. Massey, W. Whitt. 1993a. The Physics of The $M_t/G/\infty$ Queue. *Operations Research* 41 (4), 731–742.

[26] Eick, S., W. A. Massey, W. Whitt. 1993b. $M_t/G/\infty$ queues with sinusoidal arrival rates. *Management Science* 39 (2), 241–252.

[27] Erlang, A. A. 1948. On the rational determination of the number of circuits. In *The Life and Works of A. K. Erlang*, E. Brockmeyer, H. L. Halstrom, A. Jensen (eds.), Copenhagen Telephone Company: Second edition in 1960 by Danish Academy of Technical Sciences.

[28] Feldman, Z., A. Mandelbaum, W. A. Massey, W. Whitt. 2004. Staffing of time-varying queues to achieve time-stable performance. The Technion, Princeton University and Columbia University. Available at http://columbia.edu/∼ww2040.

[29] Feller, W. 1968. *An Introduction to Probability Theory and its Applications*, Vol. I, third edition, John Wiley and Sons.

[30] Fernandes C.M., A. Price, J.M. Christenson. 1997. Does reduced length of stay decrease the number of emergency department patients who leave without seeing a physician? *Journal of Emergency Medicine* 15 (3), 397-399.

[31] Fry, T. C. 1965. *Probability and its Engineering Uses*, van Nostrand, Princeton.

[32] Gans, N., G. Koole, A. Mandelbaum. 2003. Telephone call centers: tutorial, review and research prospects. *Manufacturing and Service Operations Management* (M&SOM) 5 (2), 79–141.

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

[34] Goldberg, J. 2004. Operations research models for the deployment of emergency services vehicles. *EMS Mgmnt. J.* 1 (1),20–39.

[35] Green, L. 1984. A multiple dispatch queueing model of police patrol operations. *Management Science* 30 (6), 653–664.

[36] Green, L. 1985. A queueing system with general-use and limited-use servers. *Operation Research* 33 (1), 168–182.

[37] Green, L., P. Kolesar. 1984a. The feasibility of one-officer patrol in New York City. *Management Science* 30 (8), 964–981.

[38] Green, L., P. Kolesar. 1984b. A comparison of the multiple dispatch and M/M/c priority queueing models of police patrol. *Management Science* 30 (6), 665–670.

[39] Green, L., P. Kolesar. 1987. On the validity and utility of queueing models of human service systems. *Annals of Operations Research* 9 (1), 469–479.

[40] Green, L., P. Kolesar. 1989. Testing the validity of a queueing model of a police patrol. *Management Science* 35 (2), 127–148.

[41] Green, L., P. Kolesar. 1991. The pointwise stationary approximation for queues with nonstationary arrivals. *Management Science* 37 (1), 84–97.

[42] Green, L. V., P. J. Kolesar. 1995. On the accuracy of the simple peak hour approximation for Markovian queues. *Management Science* 41 (8), 1353–1370.

[43] Green, L. V., P. J. Kolesar. 1997. The lagged PSA for estimating peak congestion in multiserver Markovian queues with periodic arrival rates. *Management Science* 43 (1), 80–87.

[44] Green, L. V., P. J. Kolesar. 1998. A note on approximating peak congestion in $M_t/G/\infty$ queues with sinusoidal arrivals. *Management Science* 44 (11), S137–S143.

[45] Green, L. V., P. J. Kolesar. 2004. Improving emergency responsiveness with management science. *Management Science* 50 (8), 1001–1014.

[46] Green, L., P. Kolesar, A. Svoronos. 1991. Some effects of nonstationarity on multiserver Markovian queueing systems. *Operation Research* 39 (3), 502–511.

[47] Green, L. V., P. J. Kolesar, J. Soares. 2001. Improving the SIPP approach for staffing service systems that have syclic demands. *Operations res.* 49 (4), 549–564.

[48] Green, L. V., P. J. Kolesar, J. Soares. 2003. In improved heuristic for staffing telephone call centers with limited operating hours. *Production and Operations Management* 12 (1), 46–61.

[49] Green, L.V., J. Soares. 2005. Computing time-dependent waiting time probabilities in nonstationary Markovian queueing systems. *Manufacturing and Service Operations Management*, forthcoming.

[50] Green, L.V., J. Soares, J. Giulio, R. Green. 2006. Using queueing theory to increase the effectiveness of physician staffing in the emergency department. *Academic Emergency Medicine* 13 (1), 61–68.

[51] Green R.A., P.C. Wyer, J. Giglio. 2002. ED walkout rate correlated with ED length of stay but not with ED volume or hospital census (abstract). *Academic Emergency Medicine* 9 (5), 514 .

[52] Grier, N., T. McKoy, W. A. Massey, W. Whitt. 1997. The time-dependent Erlang loss model with retrials. *Telecommunications Systems* R. B. Cooperand and R. Doverspike (eds.), 7 (1-3), 253–265.

[53] Halfin, S., W. Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Operations research* 29 (5), 567–587.

[54] Hall, R. W. 1991. *Queueing Methods for Services and Manufacturing* Prentice Hall.

[55] Harrison, J. M., A. Zeevi. 2005. A method for staffing large call centers based on stochastic fluid models. *Manufacturing and Service Operations Management* (M&SOM), 7 (1), 20–36.

[56] Heyman, D. P., W. Whitt. 1984. The Asymptotic Behavior of Queues with Time-Varying Arrival Rates. *Journal of Applied Probability* 21 (1), 143–156.

[57] Hoffman, K. L., C. M. Harris. 1986. Estimation of a caller retrial rate for a telephone information system. *European Journal of Operations Research* 27 (1), 207–214.

[58] Holloran, T.J., J.E. Byrn. 1986. United Airlines station manpower planning system. *Interfaces* 16 (1), 39-50.

[59] Jagerman, D. L. 1975. Nonstationary blocking in telephone traffic. *Bell System Tech. J.* 54 (3), 625–661.

[60] Jelenkovic, P., A. Mandelbaum, P. Moncilovic. 2004. Heavy-traffic limits for queues with many deterministic servers. *Queueing Systems* 47 (1), 53–69.

[61] Jennings, O. B., A. Mandelbaum, W. A. Massey, W. Whitt. 1996. Server staffing to meet time-varying demand. *Management Science* 42 (10), 1383–1394.

[62] Jennings, O. B., W. A. Massey. 1997. A modified offered load approximation for nonstationary circuit switched networks. *Telecommunication Systems* 7 (1-3), 253–265.

[63] Keller, J. B. 1982. Time-dependent queues. *SIAM Review* 24 (4), 401–412.

[64] Kolesar, P. 1984. Stalking the endangered CAT: a queueing analysis of congestion at automatic teller machines. *Interfaces* 14 (6), 16-26.

[65] Kolesar, P. J., L. V. Green. 1998. Insights on service system design from a normal approximation to Erlang's delay formula. *Production and Operations Management* 7 (3), 282–293.

[66] Kolesar, P., K. Rider, T. Crabill, W. Walker. 1975. A queueing linear programming approach to acheduling police cars. *Operations Research* 23 (6), 1045-1062.

[67] Kolesar, P., W.E. Walker. 1975. A simulation model of police patrol operations. Report R-1625-NYC/HUD, The Rand Corporation, Santa Monica, CA.

[68] Koopman, B. O. 1972. Air-terminal queues under time-dependent conditions. *Operations Research* 20 (6), 1089–1114.

[69] Kwan, S. K., M. M. Davis, A. G. Greenwood. 1988. A simulation model for determining variable worker requirements in a service operation with time-dependent customer demand. *Queueing Systems* 3 (2), 265–276.

[70] Leung, K. K., W. A. Massey, W. Whitt. Traffic models for wireless communication networks. *IEEE Journal on Selected Areas in Communication* 12 (8) 1353–1364.

[71] Mandelbaum, A., W. A., Massey, M. I. Reiman. 1998. Strong approximations for Markovian service networks. *Queueing Systems* 30 (1-2), 149–201.

[72] Mandelbaum, A., W. A., Massey, M. I. Reiman, R. Rider. 1999. Time-varyng multiserver queues with abandonments and retrials. In *Proceedings of the* 16^{th} *International Teletraffic Congress*, P. Key, D. Smith (eds.), 355–364.

[73] Mandelbaum, A., G. Pats. 1995. State-dependent queues: approximations and applications. F. P. Kelly, R. J. Williams, eds., *Stochastic Networks*, Institute for Mathematics and its Applications, Vol. 71, Springer, 239–282.

[74] Mandelbaum, A., R. Schwartz. 2002. Simulation experiments with $M/G/100$ queues in the Halfin-Whitt (QED) regime. Technical Report, The Technion, Haifa, Israel.
Available at: http://iew3.technion.ac.il/serveng/References/references.html

[75] Mandelbaum, A., S. Zeltyn. 2004. The impact of customer patience on delay and abandonment: some empirically-driven experiments with the $M/M/N+G$ queue. *OR Spektrum* 26 (3), 377–411.

[76] Mandelbaum, A., S. Zeltyn. 2005. Service engineering in action: the Palm/Erlang-A queue, with application to call centers.
Available at: http://iew3.technion.ac.il/serveng/References/references.html

[77] Massey, W. A. 1981. *Nonstationary Queues*, Ph.D. dissertation, Department of Mathematics, Stanford University.

[78] Massey, W. A. 2002. The analysis of queues with time-varying rates for telecommunication models. *Telecommunications Systems*, 21 (2-4) 173–204.

[79] Massey, W. A., G. A. Parker, W. Whitt. 1996. Estimating the parameters of a nonhomogeneous Poisson process with linear rate. *Telecommunication Systems* 5 (4), 361–388.

[80] Massey, W. A., R. B. Wallace. 2005. An optimal design of the $M/M/C/K$ queue for call centers. *Queueing Systems*, forthcoming.

[81] Massey, W. A., W. Whitt. 1993. Networks of infinite-server queues with nonstationary Poisson input. *Queueing Systems* 13 (1), 183–250.

[82] Massey, W. A., W. Whitt. 1994a. An analysis of the modified offered load approximation for the nonstationary Erlang loss model. *Annals of Applied Probability* 4 (4), 1145–1160.

[83] Massey, W. A., W. Whitt. 1994b. Unstable asymptotics for nonstationary queues. *Math. Oper. Res.* 19 (2), 267–291.

[84] Massey, W. A., W. Whitt. 1994c. A stochastic model to capture space and time dynamics in wireless communication systems. *Probability in the Engineering and Informational Sciences* 8, 541–569.

[85] Massey, W. A., W. Whitt. 1997. Peak congestion in multi-server service systems with slowly varying arrival rates. *Queueing Systems* 25 (1-4), 157–172.

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

[87] McCaig LF, C.W. Burt. 2002. *National Hospital Ambulatory Medical Care Survey: 2002 Emergency Department Summary. Advance data from vital and health statistics*, CDC: National Center for Health Statistics, Hyattsville, Maryland 340, 1-36.

[88] Mina, R. R. 1974. *Introduction to Teletraffic Engineering*, Telephony Publishing Corporation, Chicago.

[89] Molina, E. C. 1922. The theory of probabilities applied to telephone trunking problems. *Bell System Tech. J.* 1 (1), 69–81.

[90] Moore. M., G. Allison, T. Bates, J. Downing. 1975. The fourth platoon. Kennedy School of Government, Harvard University, Case C14-75-013, Cambridge MA.

[91] Nelson, B. L., M. R. Taaffe. 2004a. The $Ph_t/Ph_t/\infty$ queueing system: Part I–the single node. *INFORMS Journal on Computing* 16 (3), 266–274.

[92] Nelson, B. L., M. R. Taaffe. 2004b. The $[Ph_t/Ph_t/\infty]^K$ queueing system: Part II–the multiclass network. *INFORMS Journal on Computing* 16 (3), 266–274.

[93] Newell, G. F. 1982. *Applications of Queueing Theory*, second edition, Chapman and Hall.

[94] Ong, K. L., M. R. Taaffe. 1987. Approximating nonstationary $Ph(t)/M(t)/S/C$ queueing systems. *Annals of Operations Research* 8 (1), 103–116.

[95] Palm, C. 1943. *Intensity Variations in Telephon Traffic*, in German, *Ericsson Technics* 44. Translated into English, North-Holland, 1988.

[96] Puhalskii, A. A., M. I. Reiman. 2000. The multiclass $GI/PH/N$ queue in the Halfin-Whitt regime. *Advances in Applied Probability* 32 (2), 564–595.

[97] Rosengrant, S. 1993. A measure of delight: the pursuit of quality at AT&T universal card services (A). Harvard Business School Case No 9-694-077, Boston, MA.

[98] Ross, S. M. 2003. *Introduction to Probability Models*, eightth edition, Academic Press.

[99] Rothkoph, M. H., S. S. Oren. 1979. A closure approximation for the nonstationary $M/M/s$ queue. *Management Science* 25 (6), 522–534.

[100] Schindler, S., T. Semmel. 1993. Station staffing at Pan American World Airways. *Interfaces* 23 (3), 91-98.

[101] Seelen, L. P. 1984. An algorithm for $Ph/Ph/c$ queues. Research Report 131, Department of Actuariel Sciences and Econometrics, Free University, Amsterdam.

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

[103] Segal, M. 1974. The operator scheduling problem: a network-flow approach. *Operations Research* 24 (4), 808–823.

[104] Sisselman, M., W. Whitt. 2004. Value-based routing and preference-based routing in customer contact centers. Available at http://columbia.edu/∼ww2040.

[105] Syski, R. 1986. *Introduction to Congestion Theory in Telephone Systems*, second edition, North-Holland (first edition in 1960).

[106] Sze, D. Y. 1984. A queueing model for telephone operator staffing. *Operations Research* 32 (2), 229–249.

[107] Takács, L. 1962. *Introduction to the Theory of Queues.* Oxford University Press.

[108] Takahashi, Y., Y. Takami. 1976. A numerical method fo the steady-state probabilities of a $GI/G/c$ queueing system in a general class. *Journal of the Operations Research Society of Japan* 19 (2), 147–157.

[109] Thompson, G. M. 1993. Accounting for the multi-period impact of service when determining employee requirements for labor scheduling. *J. Oper. Management* 11 (3), 269–287.

[110] Wallace, R. B. 2004. *Performance Modeling and Design of Call Centers with Skill-Based Routing*, Ph.D. dissertation, the George Washington University, School of Engineering and Applied Science.

[111] Wallace, R. B., W. Whitt. 2005. A staffing algorithm for call centers with skill-based routing. *Manufacturing and Service Operations Management* (M&SOM) 7 (4), 276–294.

[112] Whitt, W. 1991. The pointwise stationary approximation for $M_t/M_t/s$ queues is asymptotically correct as the rate increases. *Management Science* 37 (3), 307–314.

[113] Whitt, W. 1992. Understanding the efficiency of multi-server service systems. *Management Science* 38 (5), 708–723.

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

[115] Whitt, W. 1999a. Dynamic staffing in a telephone call center aiming to immediately answer all calls. *Operations Res. Letters* 24 (5), 205–212.

[116] Whitt, W. 1999b. Using different reponse-time requirements to smooth time-varying demand for service. *Operations Research Letters* 24 (1), 1–10.

[117] Whitt, W. 2000. The impact of a heavy-tailed service-time distribution upon the $M/GI/s$ waiting-time distribution. *Queueing Systems* 36 (1), 71–87.

[118] Whitt, W. 2002. *Stochastic-Process Limits*, Springer.

[119] Whitt, W. 2004a. A diffusion approximation for the $G/GI/n/m$ queue. *Operations Research* 52 (6), 922–941.

[120] Whitt, W. 2004b. Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Science* 50 (10), 1449–1461.

[121] Whitt, W. 2005a. Engineering solution of a basic call-center model. *Management Science* 51 (2) 221–235.

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

[123] Whitt, W. 2005c. Fluid models for many-server queues with abandonments. *Operations Research*, forthcoming. Available at http://columbia.edu/∼ww2040.

[124] Whitt, W. 2005d. Sensitivity of performance in the Erlang A model to changes in the model parameters. *Operations Research*, forthcoming. Available at http://columbia.edu/∼ww2040.

[125] Whitt, W. 2005e. Staffing a call center with uncertain arrival rate and absenteeism. *Production and Operations Management* (POM), forthcoming.
Available at http://columbia.edu/∼ww2040.

[126] Wolff, R. W. 1989. *Stochastic Modeling and the Theory of Queues*, Prentice-Hall.

[127] Zeltyn, S., A. Mandelbaum. 2005. Call centers with impatient customers: many-server asymptotics of the $M/M/n + G$ queue. working paper, The Technion.
Available at: http://iew3.technion.ac.il/serveng/References/references.html