STAFFING OF TIME-VARYING QUEUES
TO ACHIEVE TIME-STABLE PERFORMANCE

by

Z. Feldman

Technion Institute
Haifa 32000
ISRAEL
zoharf@tx.technion.ac.il

A. Mandelbaum

Technion Institute
Haifa 32000
ISRAEL
avim@ie.technion.ac.il

W.A. Massey

Princeton University
Princeton, NJ 08544
U.S.A.
wmassey@princeton.edu

W. Whitt

Columbia University
New York, NY 10027-6699
U.S.A.
ww2040@columbia.edu

November 2004, Revision: November 25, 2005

Abstract

This paper develops methods to determine appropriate staffing levels in call centers and other many-server queueing systems with time-varying arrival rates. The goal is to achieve targeted time-stable performance, even in the presence of significant time-variation in the arrival rates. The main contribution is a flexible simulation-based *iterative-staffing algorithm* (ISA) for the $M_t/G/s_t + G$ model - with nonhomogeneous Poisson arrival process (the M_t) and customer abandonment (the $+G$). For Markovian $M_t/M/s_t + M$ special cases, the ISA is shown to converge. For that $M_t/M/s_t+M$ model, simulation experiments show that the ISA yields time-stable delay probabilities across a wide range of target delay probabilities. With ISA, other performance measures - such as agent utilizations, abandonment probabilities and average waiting times - are stable as well. The ISA staffing and performance agree closely with the modified-offered-load (MOL) approximation, which was previously shown to be an effective staffing algorithm without customer abandonment. While the ISA algorithm so far has only been extensively tested for $M_t/M/s_t + M$ models, it can be applied much more generally, to $M_t/G/s_t + G$ models and beyond.

Keywords: Contact centers; call centers; staffing; non-stationary queues; queues with time-dependent arrival rates; many-server queues; capacity planning; queues with abandonment; time-varying Erlang models.

1. Introduction

In this paper we develop methods to determine appropriate staffing levels in call centers and other many-server queueing systems with time-varying arrival rates. For background on call centers, see Gans et al. (2003).

In setting staffing levels, we are faced with two sources of variability: *predictable variability* – time-variations of the expected load – and *stochastic variability* – random fluctuations around this time-dependent average. (There may also be model uncertainty, but we do not consider it.) Most available staffing algorithms are designed to cope only with stochastic variability, avoiding the predictable variability in various ways. For example, when the service times are relatively short, as in many call centers when service is provided by a telephone call, it is customary to use a *pointwise stationary approximation* (PSA), i.e., to act as if the system at time t were in steady-state with the arrival rate occurring at that instant (or during that half hour); see Green and Kolesar (1991) and Whitt (1991). In call centers, staffing typical is held constant over staffing intervals of $15 - 30$ minutes. The effect of staffing intervals can be important, see Green et al. (2001), but here we do not consider staffing intervals.

However, service times are not always short, even in call centers. If relatively lengthy interactions are not uncommon or if arrival rates change quite rapidly, then PSA can produce poor performance. As a consequence, some parts of the day may be overstaffed, while others are understaffed. For a review of staffing methods to cope with time-varying arrivals, see Green et al. (2005).

In this paper we address the staffing problem with *both* predictable and stochastic variability: Given a daily performance goal, and faced with both predictable and stochastic variability, we seek to find the minimal staffing levels that meet this performance goal stably over the day. We aim to understand when PSA is appropriate and to do significantly better than PSA when it is not appropriate. We emphasize the importance of achieving time-stable performance. With time-stable performance, the nearly-constant quality of service is easily adjusted up or down, as desired. Moreover, our experience suggests that customers tend to prefer consistent performance even at the expense of some service level.

Our main contribution in this paper is a flexible simulation-based *iterative-staffing algorithm* (ISA). We develop the ISA for the many-server $M_t/G/s_t + G$ queueing model, which has a nonhomogeneous Poisson arrival process (the M_t) with time-varying arrival-rate function $\lambda(t)$, independent and identically distributed (i.i.d.) random service times with a general cumulative distribution function or cdf (the first G), a time-varying number of servers s_t, which is for us to set, and i.i.d. random times to abandon (before starting service) with a general cdf (the final $+G$). Allowing non-exponential service-time and time-to-abandon distributions is important, because they have been found to occur in practice; see Bolotin (1994) and Brown et al. (2005).

We show that the ISA staffing function s_t^{ISA} yields time-stable delay probabilities across a wide range of delay-probability targets for the Markovian $M_t/M/s_t{+}M$ special case, where the service-time and time-to-abandon cdf's are exponential with means μ^{-1} and θ^{-1}, respectively. Even though we only report results for ISA applied to Markovian $M_t/M/s_t + M$ models, the method is developed for more general $M_t/G/s_t + G$ models. (Indeed, we obtained similar results for log-normal and deterministic service-time distributions.) Moreover, the ISA applies much more generally, so that it has the potential of far-reaching applications. Indeed, by being based on simulation, ISA has two important advantages: First, by using simulation, we achieve *generality*: We can apply the approach to a large class of models; we are not limited to models that are analytically tractable. We are able to include realistic features, not ordinarily considered in analytical models. For example, we can carefully consider what happens to agents who are in the middle of a call when their scheduled shift ends. Second, by using simulation, we achieve *automatic validation*: In the process of performing the algorithm, we directly confirm that ISA achieves its goal; we directly observe the performance of the system under the final staffing function $\{s_t^{ISA} : 0 \leq t \leq T\}$. Of course, in other settings the effectiveness of the ISA still needs to be verified.

Here is how the rest of this paper is organized: In §2 we specify the ISA. Then in §3 we review the infinite-server and modified-offered-load (MOL) approximations from Jennings et al. (1996). We will show that the ISA staffing levels and performance agree closely with MOL and that both perform well. In §4 and §5 we illustrate the performance of ISA by considering $M_t/M/s_t + M$ examples, first with a stylized sinusoidal arrival-rate function and long service times, and then with a realistic arrival-rate function from a medium-sized financial-services call center, taken from Green et al. (2001) and shorter (customary) service times. In §6 we present some supporting theory for the case $\theta = \mu$. In §7, we discuss the dynamics of the iterative algorithm, establishing convergence of the ISA in the $M_t/M/s_t + M$ special case (for all μ and θ). Finally, in §8 we draw conclusions and indicate some directions for further research.

We present additional material in a longer unabridged version available on line as an Internet Supplement. There we consider the $M_t/M/s_t$ model (without abandonment) with the same sinusoidal arrival-rate function used for the $M_t/M/s_t+M$ model in §4, and show that ISA also works well for it. We also revisit the "challenging example" in Jennings et al. (1996), again showing that ISA performs well, just like MOL. We expand the analysis of the $M_t/M/s_t+M$ example in §4 by considering different abandonment rates, in particular, $\theta = 0.2$ and $\theta = 5.0$ with $\mu = 1$, representing relatively patient and impatient customers, respectively. We present additional material for the realistic example discussed in §5. We also provide additional theoretical perspective for the square-root-staffing algorithm from a uniform-acceleration perspective, as in Mandelbaum et al. (1998) and Massey and Whitt (1998) and references therein.

2. The Simulation-Based Iterative-Staffing Algorithm (ISA)

In this section we specify the ISA. For our implementation of the algorithm, we assume that we have an $M_t/G/s_t+G$ model, but it will be evident that the method applies much more generally. To start, we specify a time-horizon $[0,T]$, an arrival-rate function $\{\lambda(t); 0 \le t \le T\}$, a service-time cdf and a time-to-abandon cdf. The algorithm is iterative, continuing until the observed error is negligible. Let $s_t^{(n)}$ be the staffing level at time t in iteration n and let $N_t^{(n)}$ be the total number of customers in the system at time t under this staffing function. The final iteration yields the ISA staffing s_t^{ISA} and the stochastic process N_t^{ISA} representing the number of customers in the system with that staffing function.

Although our algorithm is time-continuous, we make staffing changes only at discrete times. That is achieved by dividing the time-horizon into small intervals of length Δ. In all experiments presented in this paper, we use $\Delta = 0.1/\mu$, where $1/\mu$ is the mean service time. We then let the number of servers be constant within each of these intervals. For any specified staffing function, the system simulation can be performed in a conventional manner. We generate a continuous-time sample path for the number in system by successively advancing the next generated event. The candidate next events are of course arrivals, service completions, abandonments and ends of shifts (the times at which the staffing function is allowed to change). For non-stationary Poisson arrival process, we generated arrival times by thinning a single Poisson process with a constant rate λ^* exceeding the maximum of the arrival-rate function $\lambda(t)$ for all t, $0 \le t \le T$. Then an event in the Poisson process at time t (a potential arrival time) is in an actual arrival in the system with probability $\lambda(t)/\lambda^*$, independent of the history up to that time; see Section 5.5 of Ross (1990). We estimate the distribution of $N_t^{(n)}$ for each n and t by performing multiple (5000) independent replications. We think of starting off with infinitely many servers. Since this is a simulation, we choose (after experimenting) a large finite number, ensuring that the probability of delay (i.e., of having all servers busy upon arrival) is negligible for all t.

The algorithm iteratively performs the following steps, until convergence is obtained. Convergence means that the staffing levels do not change more than some threshold τ after an iteration, which we take to be 1.

1. Given the i^{th} staffing function $\{s_t^{(i)} : 0 \le t \le T\}$, evaluate the distribution of $N_t^{(i)}$ for all t using simulation.

2. For each t, $0 \le t \le T$, let $s_t^{(i+1)}$ be the least number of servers so that the delay-probability constraint is met at time t; i.e., let

$$s_t^{(i+1)} = \arg\min \{k \in \mathbb{N} : P\{N_t^{(i)} \ge k\} \le \alpha\} \ .$$

3. If there is negligible change in the staffing from iteration i to iteration $i+1$, then stop; i.e., if

$$\max \{|s_t^{(i+1)} - s_t^{(i)}| : 0 \le t \le T\} \le \tau \ ,$$

then stop and let $s_t^{(i+1)}$ be the proposed staffing function, denoted by s_t^{ISA}. Otherwise, advance to the next iteration, i.e., replace i by $i+1$ and go back to step 1. ■

As indicated before, s_t^{ISA} denotes the final staffing level at time t and N_t^{ISA} denotes the number in system at time t with that staffing function. If the algorithm converges, then necessarily $P(N_t^{ISA} \ge s_t^{ISA}) \approx \alpha$, $0 \le t \le T$.

Our implementation of ISA was written in C++. For the special case of the Markovian $M_t/M/s_t + M$ model with individual service rate $\mu = 1/E[S]$ and individual abandonment rate θ, we rigorously establish convergence of the algorithm in §7. Experience indicates that the algorithm consistently converges relatively rapidly. Experience also indicates that the final time-dependent delay probabilities, and other performance measures, are remarkably stable. The number of iterations required depends on the parameters, especially the ratio $r \equiv \theta/\mu$. If $r = 1$, corresponding to an infinite-server queue - see §6, then no more than two iterations are needed, since the distribution of the number in system does not depend upon the number of servers in that special case. As r departs from 1, the number of required iterations typically increases. For example, when $r = 10$, the number of iterations can get as high as $6 - 12$. When r is very small and the traffic intensity is very high, so that we are at the edge of stability, the number of iterations can be very large. For more discussion, see §7.

3. Infinite-Server and Modified-Offered-Load Approximations

In this section we review staffing algorithms based on *infinite-server* (IS) and *modified-offered-load* (MOL) approximations from our (with Otis B. Jennings) previous paper Jennings et al. (1996). These approximations were developed for the $M_t/G/s_t$ model without customer abandonment, but the methods extend directly to the corresponding model with customer abandonment. The effectiveness of these methods with abandonments was not demonstrated previously, though. Our simulation experiments here will show that ISA produces essentially the same results as MOL, with and without customer abandonment, and that both are effective. (Our reported experiments are limited to Markovian $M_t/M/s_t + M$ models, but limited experimentation for other $M_t/G/s_t + G$ models indicate that excellent results hold there too.)

To describe our goal in staffing, let N_t be the number of customers in the $M_t/G/s_t + G$ system at time t, either waiting or being served. We focus on the probability of delay (of a potential arrival, i.e., $P(N_t \geq s_t)$), aiming to choose the time-dependent staffing level s_t such that

$$P(N_t \geq s_t) \leq \alpha < P(N_t \geq s_t - 1) \quad \text{for all} \quad t \ , \tag{3.1}$$

where α is the target delay probability.

The Infinite-Server Approximation. We discuss the MOL and infinite-server approximations together, because the MOL approximation builds on the infinite-server approximation. We start by considering the infinite-server approximation. Why would anyone consider an infinite-server approximation? From a mathematical perspective, the reason is that the finite-server $M_t/G/s_t + G$ model of interest is analytically intractable, whereas the corresponding infinite-server $M_t/G/\infty$ model is remarkably tractable. From an engineering perspective, the reason is that the infinite-server model can be used to show the amount of capacity that would actually be used (and is thus needed) if there were no capacity constraints (i.e., a limited number of servers). For the Markovian $M_t/M/s_t + M$ model, where $\theta = \mu$, there is even a stronger connection: In that special case, the distribution of the number of customers in the infinite-server $M_t/M/\infty$ model actually *coincides* with the distribution of the number of customers in the $M_t/M/s_t + M$ model, as we explain in §6, so there is additional strong motivation for considering the infinite-server approximation.

So what does the infinite-server approximation do? The infinite-server approximation for the $M_t/G/s_t + G$ model approximates the random variable N_t by the number N_t^{∞} of busy servers in the associated $M_t/G/\infty$ model, having infinitely many servers but the same arrival process and service times. *The infinite-server staffing function s_t^∞ is obtained by applying* (3.1) *with N_t^∞ instead of N_t.* As we now explain, that approximation provides great simplification because (i) the tail probability $P(N_t^\infty \geq s_t)$ at time t depends on the staffing function $\{s_t : t \geq 0\}$ only through its value at the single time t and (ii) the exact time-dependent distribution of N_t^∞ is known.

The first simplification follows from the fact that the distribution of the stochastic process $\{N_t^\infty : t \geq 0\}$ is totally independent of the staffing function $\{s_t : t \geq 0\}$. When we calculate $P(N_t^\infty \geq s_t)$, the staffing level s_t just serves as the argument of the tail-probability function. The second simplification stems from basic properties of $M_t/G/\infty$ queues. In particular, as reviewed in Eick et al. (1993a), for each t, N_t^∞ has a *Poisson distribution* whenever the number in the system at the initial time has a Poisson distribution. (Being empty is a degenerate case of a Poisson distribution.) That Poisson distribution is fully characterized by its mean m_t^∞.

As in previous work, such as Eick et al. (1993a,b) and Jennings et al. (1996), our work reported here shows that the time-dependent mean m_t^∞ is the crucial quantity. We regard this exact time-dependent mean m_t^∞ in the $M_t/G/\infty$ model as the (*time-dependent*) *offered load* for the $M_t/G/s_t + G$ model.

We now observe that convenient formulas exist for the offered load m_t^∞. Eick et al. (1993a) showed that the offered load has the tractable representation

$$m_t^\infty \equiv E\left[N_t^\infty\right] = \int_{-\infty}^{t} G^c(t-u)\lambda(u)\, du = E\left[\int_{t-S}^{t} \lambda(u)\, du\right] = E\left[\lambda(t-S_e)\right] E[S] \ , \qquad (3.2)$$

where $\lambda(t)$ is the arrival-rate function, S is a generic service time with cdf G, $G^c(t) \equiv 1-G(t) \equiv P(S > t)$, and S_e is a random variable with the associated *stationary-excess cdf* (or equilibrium-residual-lifetime cdf) G_e associated with the service-time cdf G, defined by

$$G_e(t) \equiv P(S_e \leq t) \equiv \frac{1}{E[S]} \int_0^t G^c(u)\, du, \quad t \geq 0 \ , \qquad (3.3)$$

with k^{th} moment $E[S_e^k] = E[S^{k+1}]/((k+1)E[S])$; see Theorem 1 of Eick et al. (1993a) and references therein.

The different expressions in (3.2) provide useful insight; see Eick et al. (1993a, b) and Section 4.2 of Green et al. (2005). For the special case in which $\lambda(t)$ is constant, $m_t^\infty \equiv m^\infty = \lambda E[S]$. Accordingly, the PSA approximation for m_t^∞ in the $M_t/G/\infty$ model is $m_t^{PSA} \equiv \lambda(t)E[S]$. We call m_t^{PSA} the PSA (time-dependent) offered load for the $M_t/G/s_t + G$ model.

In addition, there are convenient explicit formulas for m_t^∞ in special cases as well as useful approximations. We will use the explicit formula for sinusoidal arrival-rate functions in §4.

Based on a second-order Taylor-series approximation for λ about t, the offered load can be approximated by

$$m_t^\infty \approx \lambda(t - E[S_e])E[S] + \frac{\lambda^{(2)}(t)}{2} Var(S_e)E[S] \; , \tag{3.4}$$

where $\lambda^{(2)}(t)$ is the second derivative of the function λ evaluated at time t; see Theorem 9 of Eick et al. (1993a). Approximation (3.4) shows that the approximate offered load in (3.4) coincides with the PSA offered load $m_t^{PSA} \equiv \lambda(t)E[S]$ except for a *time shift* by $E[S_e]$ and a *space shift* by $\lambda^{(2)}(t)Var(S_e)E[S]/2$. The time shift is especially important. A simple refinemnt of PSA based on (3.4) suggested by Eick et al. (1993a) is *lagged PSA*, where we ignore the space shift and approximate m_t^∞ by $\lambda(t - E[S_e])E[S]$.

We now continue, exploiting the established Poisson distribution with a known time-dependent mean m_t^∞. Assuming that m_t^∞ is not extremely small, we can apply a *normal approximation* for the Poisson distribution, obtaining first $P(N_t \geq s_t) \approx P(N_t^\infty \geq s_t)$ and then

$$P(N_t^\infty \geq s_t) \approx P(N(m_t^\infty, m_t^\infty) \geq s_t) = P\left(N(0,1) \geq \frac{s_t - m_t^\infty}{\sqrt{m_t^\infty}}\right) = 1 - \Phi\left(\frac{s_t - m_t^\infty}{\sqrt{m_t^\infty}}\right) \; , \tag{3.5}$$

where $N(m, \sigma^2)$ denotes a normally distributed random variable with mean m and variance σ^2, and $\Phi(x) \equiv P(N(0,1) \leq x)$ is the standard normal cdf.

From (3.5), we see that we can obtain a stable approximate delay probability if we can choose the staffing function s_t^∞ to make $(s_t^\infty - m_t^\infty)/\sqrt{m_t^\infty}$ stable in the final term of (3.5). Accordingly, we obtain the *square-root-staffing formula*:

$$s_t^\infty = \lceil m_t^\infty + \beta\sqrt{m_t^\infty} \rceil, \quad 0 \leq t \leq T, \tag{3.6}$$

where $\lceil x \rceil$ is the least integer greater than or equal to x and the constant β is a measure of the *quality of service*. Combining the target in (3.1) and the normal approximation in (3.5), we see that the quality-of-service parameter β in (3.6) should be chosen so that $1 - \Phi(\beta) = \alpha$.

The normal approximation and the square-root-staffing formula for *stationary* many-server queues are classic results, see Whitt (1992) and references therein. What is less well understood is the role of the offered load m_t^∞ with time-varying arrivals. The notation s_t^∞ means that we staff according to the infinite-server approximation. In doing so, we not only apply the normal approximation and the square-root-staffing formula, but we also use the infinite-server mean m_t^∞ as the offered load.

The MOL Approximation. Section 4 of Jennings et al. (1996) also introduced a refinement of the infinite-server approximation for the time-dependent delay probabilities, which is tantamount to a *modified-offered-load* (MOL) approximation, as in Jagerman (1975) and Massey and Whitt (1994, 1997). The MOL approximation for N_t in the $M_t/G/s_t + G$ model at time t, denoted by N_t^{MOL}, is the limiting steady-state number of customers in the system in the corresponding stationary $M/G/s + G$ model (with the same service-time and time-to-abandon distributions and the same number of servers s_t at time t), but using m_t^{∞} as the stationary offered load operating at time t. Since the stationary offered load is $\lambda E[S]$, that means letting the homogeneous Poisson arrival process in the stationary $M/G/s + G$ model have time-dependent arrival rate

$$\lambda_t^{MOL} \equiv \frac{m_t^{\infty}}{E[S]} = m_t^{\infty}\mu \quad \text{at time} \quad t\ . \tag{3.7}$$

The MOL staffing function s_t^{MOL} is obtained by applying (3.1) *with N_t^{MOL} instead of N_t.*

The important insight is that the "right" time-dependent offered load in the $M_t/G/s_t + G$ model should be the time-dependent mean number of busy servers in the associated infinite-server model - m_t^{∞}. Since the right offered load for the stationary model is $\lambda E[S]$, the "obvious" direct time-dependent generalization is the PSA offered load $m_t^{PSA} \equiv \lambda(t)E[S]$. However, $\lambda E[S]$ is also the mean number of busy servers in the associated stationary infinite-server model. It turns out that the mean number of busy servers in the infinite-server model is a better generalization of "offered load" than the PSA time-dependent offered load for most time-varying many-server models. Indeed, it may be considered exactly the right definition for the infinite-server model itself.

The MOL approximation in §4 of Jennings et al. (1996) was not applied directly. Instead of calculating the steady-state delay probability for the stationary $M/M/s$ model, we exploited an approximation for the delay probability based on a many-server heavy-traffic limit in Halfin and Whitt (1981). That produces a simple formula relating the delay probability α and the service quality β. Moreover, the heavy-traffic limit provides an alternative derivation of the square-root staffing formula in (3.6), without relying on an infinite-server approximation or a normal approximation. We will do the same thing here with customer abandonments, relying on the heavy-traffic limits for the $M/M/s + M$ model established by Garnett et al. (2002).

Jennings et al. (1996) showed that the method for setting staffing requirements in the $M_t/G/s_t$ model outlined above is remarkably effective. This was demonstrated by doing numerical comparisons for the $M_t/M/s_t$ special case. For any given staffing function, the time-dependent distribution of N_t in that Markovian model can be derived by solving a system of time-dependent ordinary differential equations (ODE's). We too could have exploited ODE's for the $M_t/M/s_t + M$ model, but we wanted to develop a method that applies to much more general models.

The most important conclusion from those previous experiments in Jennings et al. (1996) is that it is indeed possible to achieve time-stable performance for the $M_t/M/s_t$ model by an appropriate choice of a staffing function s_t, even in the face of a strongly time-varying arrival-rate function. Here we show the same is true for the $M_t/M/s_t + M$ model. And we provide a means to go far beyond these Markovian models.

4. An $M_t/M/s_t + M$ Example with a Sinusoidal Arrival-Rate Function

We demonstrate the performance of ISA by considering $M_t/M/s_t + M$ examples. We start in this section with a sinusoidal arrival-rate function

$$\lambda(t) = a + b \cdot \sin(ct), \quad 0 \le t \le T \ , \tag{4.1}$$

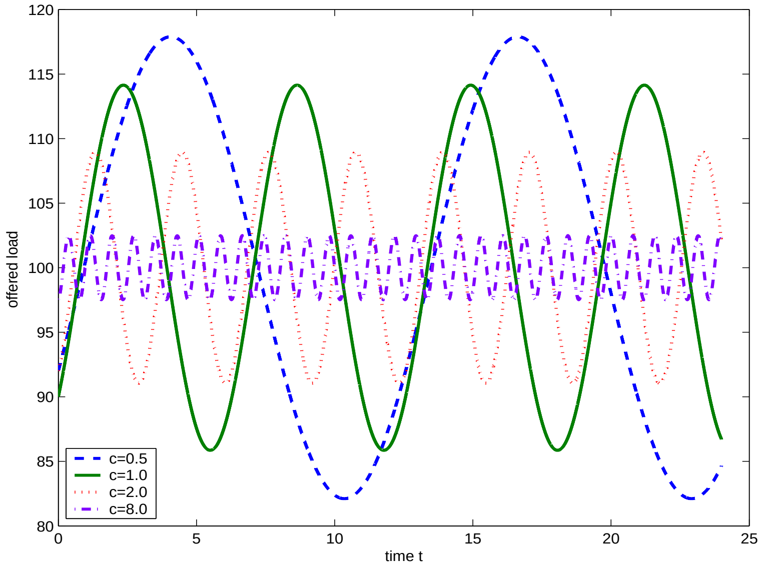
letting $a = 100$, $b = 20$ and $c = 1$. Here we let the individual service rate μ and the individual abandonment rate θ both be 1. Letting $\mu = 1$ is without loss of generality, because we are free to choose the time units. For the special case $\theta = \mu$ that we consider, we have strong supporting theory in §6, but we also found that ISA is effective with other abandonment rates. We show corresponding results for $\theta = 0.2$ and $\theta = 5.0$ in the Internet Supplement.

Since $m_t^{PSA} \equiv \lambda(t)E[S] = \lambda(t)$, this example captures the many-server spirit of a call center. However, the sinusoidal form of the arrival-rate function is clearly a mathematical abstraction, which has the essential property of producing significant fluctuations over time, i.e., significant predictable variability. This particular arrival-rate function is by no means critical for our analysis; our methods apply to an arbitrary arrival-rate function.

An important issue, however, is the rate of fluctuation in the arrival-rate function compared to the expected service time. To be concrete, we will measure time in hours, and focus on a 24-hour day, so that $T = 24$. A cycle of the sinusoidal arrival-rate function in (4.1) is $2\pi/c$; since we have set $c = 1$, a cycle is $2\pi \approx 6.3$ hours. Thus there will be about 4 cycles during the day.

Since we let the mean service time be 1 and have chosen to measure time in hours, the mean service time in this example is 1 hour. That clearly is relatively long for most call centers, where the interactions are short telephone calls. If we were to change the time units in order

Figure 1: **The offered load m_t^∞ for the sinusoidal arrival-rate function in (4.1) with parameters $a = 100$, $b = 20$ and four possible values of c: 0.5, 1, 2 and 8.**



to rectify that, making the expected service time 10 minutes, then a cycle of the arrival-rate function would become about 1 hour, making for more rapid fluctuations in the arrival rate than are normally encountered in call centers. Thus our example is more challenging than usually encountered in call centers, but may be approached in evolving contact centers if many interactions do indeed take an hour or more. We consider a more realistic example in §5.

Since we have a sinusoidal arrival-rate function, we can apply formula (15) of Eick et al. (1993b) to obtain

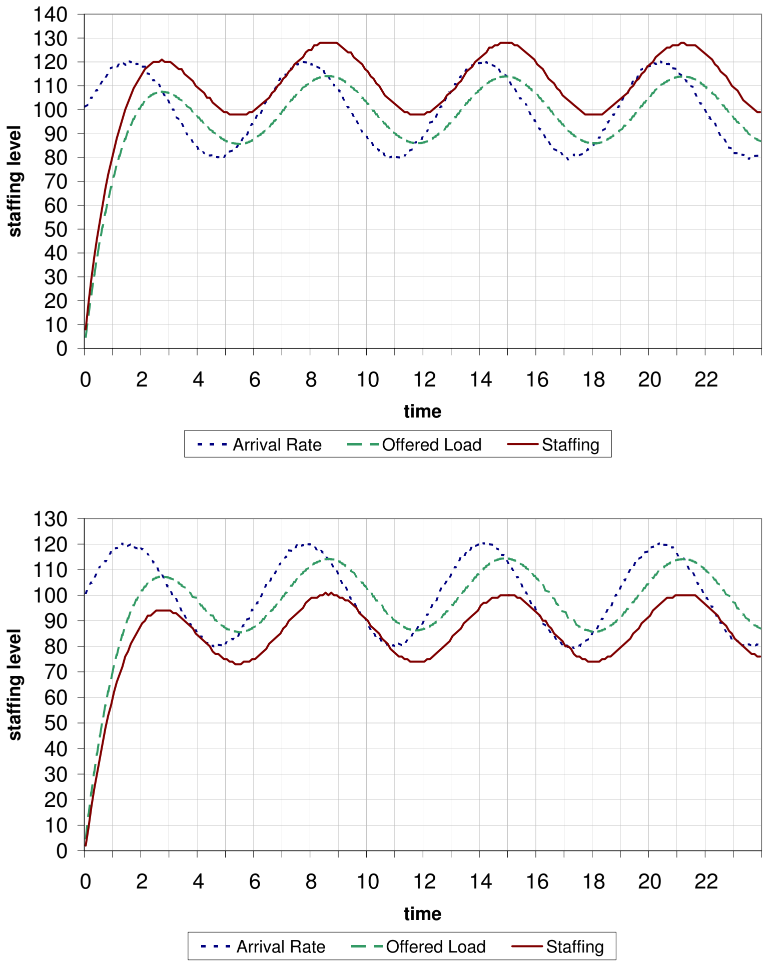
$$m_t^\infty = a + \frac{b}{1+c^2}[\sin(ct) - c \cdot \cos(ct)] \ . \tag{4.2}$$

For the specific parameters $a = 100$, $b = 20$ and $c = 1$, we get $m_t^\infty = 100 + 10[\sin(t) - \cos(t)]$.

In order to put our model into perspective, in Figure 1 we plot the time-dependent offered load m_t^∞ in (4.2) for the sinusoidal arrival-rate function in (4.1) for the parameters $a = 100$ and $b = 20$, as in our example, but with four different values of the time-scaling parameter c: 0.5, 1, 2 and 8. Note that the time-dependent offered load m_t^∞ is also a periodic function with the same period $2\pi/c$ as the arrival-rate function $\lambda(t)$, but the number of cycles increases and the amplitude (size of the fluctuations) decrease as c increases. As c increases, m_t^∞ approaches the average value $a = 100$.

In Figure 2 we present two graphs, showing the ISA staffing functions for two values of α: 0.1 and 0.9. In each graph, we plot three curves: the arrival rate $\lambda(t) \equiv m_t^{PSA}$ (dotted), the

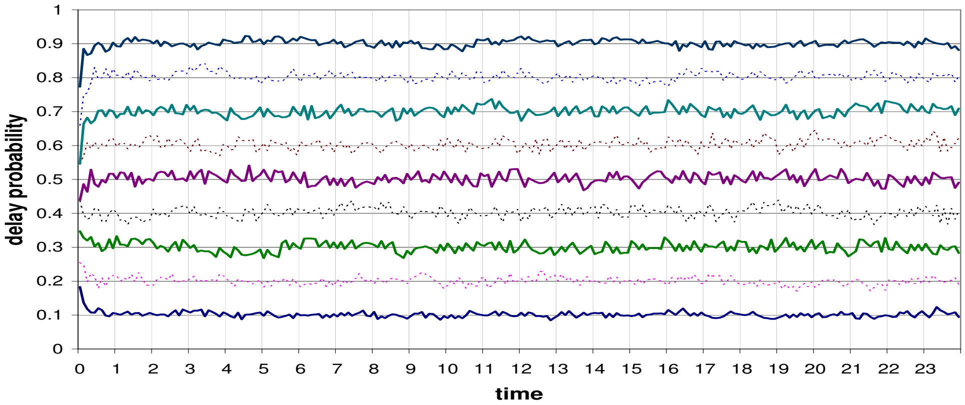
Figure 2: **Staffing - number of servers as a function of time - for the sinusoidal example: (1) $\alpha = 0.1$ (QD), (2) $\alpha = 0.9$ (ED).**



offered load m_t^{∞} (dashed) and the ISA staffing function s_t^{ISA} (solid). Note that we start our system empty. This allows us to observe the behavior of the transient stage. In particular, there is a rampup at the left side of the plot. Our methods respond appropriately to that rampup.

The two values of α used in Figure 2 plus $\alpha = 0.5$ characterize three different regimes of operation, as discussed by Garnett et al. (2002): *Quality-Driven* (QD) - target $\alpha = 0.1$, *Efficiency-Driven* (ED) - target $\alpha = 0.9$, and *Quality-and-Efficiency-Driven* (QED) - target $\alpha = 0.5$. In the QD regime, the ISA staffing function is well above the time-dependent offered load, while in the ED regime the ISA staffing function is well below the time-dependent offered load. However, in the QED regime, the ISA staffing function falls right on top of the time-dependent offered load. (For that reason, we omit the plot, since it is unnecessary.) In that

Figure 3: **Delay probabilities for the sinusoidal example with nine delay-probability targets α, ranging from 0.1 to 0.9.**

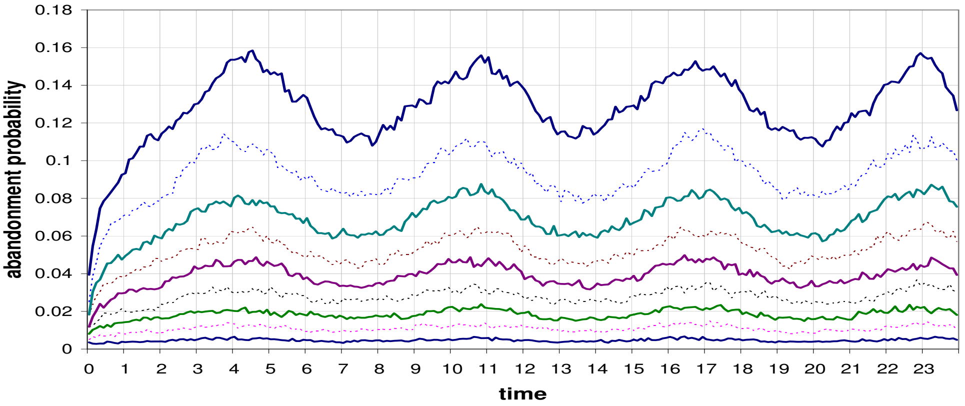


QED case ($\alpha = 0.5$), it would have sufficed to simply let $s_t = m_t$. This phenomenon held in all our experiments. That itself is quite stunning. (Note that staffing to the offered load is much easier than the full MOL approximation. Clearly, customer abandonment plays a crucial role in staffing to the offered load.)

We now show that ISA achieves the desired time-stable performance. In Figure 3 we show the ISA delay probabilities obtained with target α for $\alpha = 0.1, 0.2, \ldots, 0.9$. These delay probabilities are estimated by performing multiple (5000) independent replications with the final staffing function determined by our algorithm. (We verified that this was sufficient by repeating the experiment with independent random numbers. We saw negligible change in the plots. The observed fluctuations are largely due to the inherent discreteness: The staffing levels must be integers.) Under the ISA staffing levels, the delay probabilities are remarkably accurate and stable; the observed delay probabilities fluctuate around the target in each case.

In addition to stabilizing the delay probabilities, other performance measures (e.g. utilization, tail probabilities abandonment probabilities, etc.) are found to be quite stable as well. However, as the target delay probability increases toward heavy loading, the abandonment probabilities become much less time-stable, as shown in Figure 4. (Like the delay probability, we let the abandonment probability be for a potential arrival at time t; a precise definition is given after (6.1).) We discuss this phenomenon further in the Internet Supplement. Other measures of congestion such as average waiting time and average queue length were also found to

Figure 4: **Abandonment probabilities for the same sinusoidal example with the same nine delay-probability targets.**



be relatively stable, but like the abandonment probabilities, these too become less time-stable under heavy loads. Details are given in the Internet Supplement.

We now validate the square-root-staffing rule in (3.6). For that purpose, we define an *implied empirical quality of service* $\{\beta_t^{ISA} : 0 \le t \le T\}$ by setting

$$\beta_t^{ISA} \equiv \frac{s_t^{ISA} - m_t^{\infty}}{\sqrt{m_t^{\infty}}}, \quad 0 \le t \le T \ , \tag{4.3}$$

where m_t^{∞} is again the offered load in (3.2) and (4.2). Since s_t^{ISA} is obtained from ISA, the function β_t^{ISA} is itself obtained from ISA. It thus becomes interesting to see if the implied service grade is approximately constant as a function of time. That would empirically justify the square-root-staffing formula in (3.6).

And, indeed, it is. Again we consider nine values of α ranging from 0.1 to 0.9 in steps of 0.1. As α increases, the quality of service reflected by β_t^{ISA} decreases, from about $+1.3$ to -1.3, hitting 0 for $\alpha = 0.5$. But the main point is that β_t^{ISA} is approximately constant as a function of t for each α over the full range from 0.1 to 0.9. The oscillations in the plots are essentially the same as in Figure 3 (see the Internet Supplement).

The time-stability of β_t^{ISA} is extremely important because it validates the square-root-staffing formula in (3.6) for this example. First, Figure 3 shows that ISA is able to produce the target delay probability α for a wide range of α. When this is done, the square-root-staffing formula holds empirically. In other words, we have shown that we could have staffed directly by the infinite-server approximation and the square-root-staffing formula instead of by the ISA. The single critical non-trivial element is the offered load m_t^{∞}.

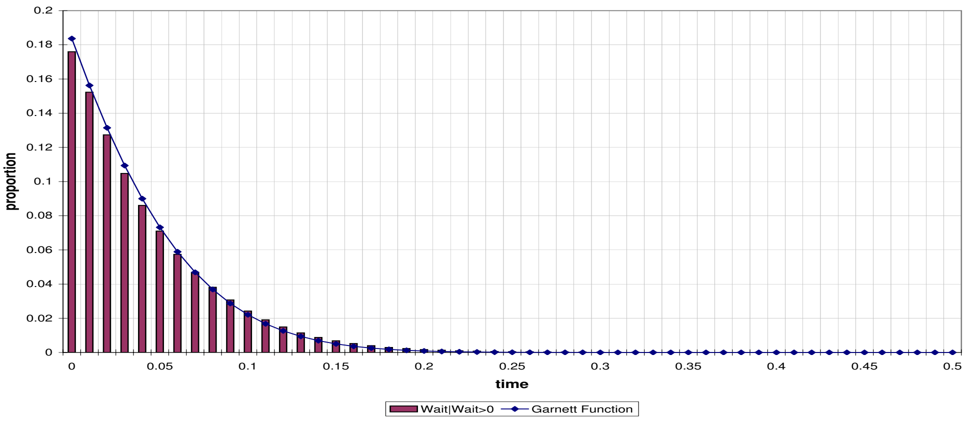
However, one issues remains: In order to staff directly by the square-root staffing formula, we need to be able to relate the quality of service β to the target delay probability α. Indeed, we want a function mapping α into β. We propose a simple answer: MOL. For the $M_t/M/s_t+M$ model with time-varying arrival-rate function $\lambda(t)$, staffing function s_t and parameters μ and θ, we use the associated stationary $M/M/s+M$ model, with the same service and abandonment rates μ and θ, and with $s = s_t$, $\lambda = \lambda_t^{MOL} = m_t^{\infty}\mu$ (as in (3.7)) for the approximation at time t. We used exact $M/M/s+M$ formulas from Garnett et al. (2002). Moreover, paralleling what was done in §4 of Jennings et al. (1996), we suggest using simple formulas obtained from the many-server heavy-traffic limit for the $M/M/s+M$ model in Garnett et al. (2002). The *Garnett function* mapping β into α is

$$\alpha = \left[1 + \sqrt{\frac{\theta}{\mu}} \cdot \frac{h(\hat{\beta})}{h(-\beta)}\right]^{-1} , \qquad -\infty < \beta < \infty , \tag{4.4}$$

where $\hat{\beta} = \beta\sqrt{\theta/\mu}$, with μ the individual service rate and θ the individual abandonment rate (both here set equal to 1 now) and $h(x) = \phi(x)/(1 - \Phi(x))$ is the *hazard rate* of the standard normal distribution, with ϕ being the *probability density function* (pdf) and Φ the cdf. To obtain the desired function mapping α into β, we can use the inverse of the Garnett function, which is well defined. For this example, the Garnett function yields essentially the same formula as the exact values for the $M/M/s+M$ model.

We also looked at additional simulation output, aimed at establishing the validity of the ISA and MOL approximations. First, we compared the empirical distribution of the customer waiting times, with ISA, to the theoretical distribution of those waiting times in the stationary $M/M/s+M$ model. To illustrate, in Figure 5 we plot the *empirical conditional waiting time pdf* given wait, i.e. the distribution of the waiting time for those who were in fact delayed, during the entire time-horizon, for the case $\alpha = 0.1$. We plot the proportions experiencing delays in intervals of length 0.01. In doing so, we are looking at all the waiting times experienced across the day. As before, we obtain statistically precise estimates by averaging over a large number of independent replications (here again 5000). In this case, the empirical conditional distribution is based on statistics gathered from the time of reaching steady state until the end of the horizon. We compared the empirical conditional waiting-time distribution to many-server heavy-traffic approximations for the conditional waiting-time distribution in the stationary $M/M/s+M$ queue, drawing on Garnett et al. (2002). Figure 5 shows that the approximation for the conditional waiting-time distribution in the stationary queues matches the performance of our time-varying model remarkably well. Plots for $\alpha = 0.5$ and $\alpha = 0.9$ in the Internet

Figure 5: **The empirical conditional waiting time distribution, given positive wait, for the $M_t/M/s_t + M$ example with delay-probability target $\boldsymbol{\alpha} = 0.1$ (QD).**



Supplement show an excellent match across the full range of delay-probability targets.

We next related the empirical (α, β) pairs to the Garnett function in (4.4). We define the empirical values $\bar{\alpha}$ and $\bar{\beta}$ as simply the time-averages of the observed (time-stable) ISA values (for α, displayed in the plot in Figure 3). In Figure 6, we plot the pairs of $(\bar{\alpha}_i, \bar{\beta}_i)$ alongside the Garnett function. Needless to say, the agreement is phenomenal!

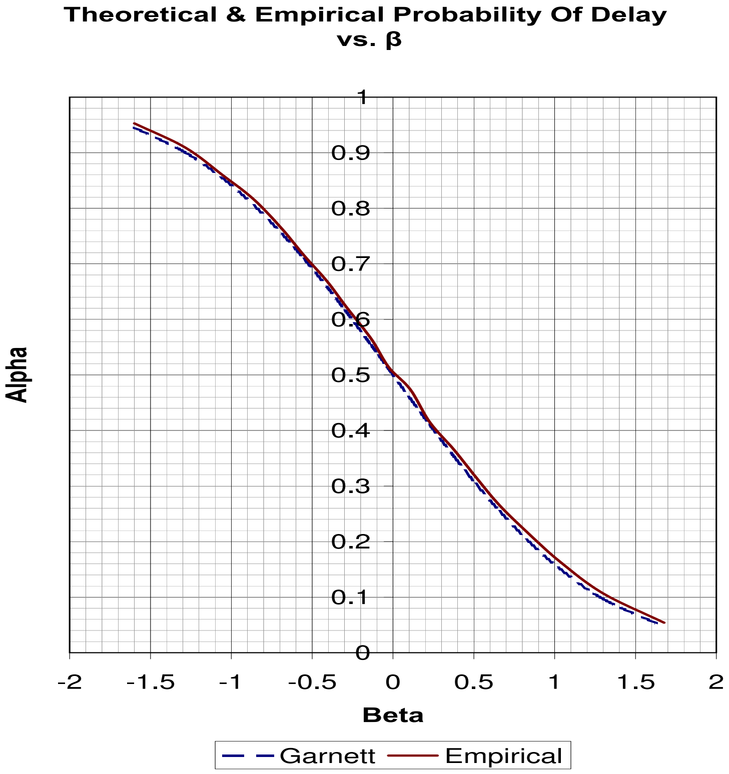
We close this section by observing that, just as in Jennings et al. (1996), other common approximations, such as the PSA or the SSA (the simple stationary approximation, using the overall time-average arrival rate) perform poorly for this example; again see the Internet Supplement.

5. A Realistic $M_t/M/s_t + M$ Example

In this section we consider a more realistic example: a medium-sized financial-services call center taken from Green et al. (2001). The hourly call volumes are shown in Figure 7. The mean service time is $E[S] = 6$ minutes. That is achieved with our hourly time scale by letting $\mu = 10$. Corresponding to that, we let $\theta = 10$, so that we have $\theta = \mu$ as in Section 4. (Green et al. (2001) did not consider customer abandonment.)

Once again, ISA is very effective. To show that, we plot the ISA delay probabilities as a function of the delay-probability target α for three values of α in Figure 8. With such short

Figure 6: **A comparison of the empirical relation between α and β with the Garnett function for the sinusoidal example.**



service times, we might think that that this should be an easy problem, for which simple PSA would also work well. Indeed, when we look at the staffing for three values of α in Figure 9, we do not see much difference, but there actually is a difference. Even though the service times are indeed short here, the arrival-rate function is changing rapidly at some times, especially in hours $4 - 6$. For this example, Figure 8 shows that simple PSA performs significantly worse than ISA.

As before, we find that ISA produces essentially the same results as MOL. Moreover, the dominant effect in MOL is captured by the time lag in (3.4); i.e., here it suffices to use *lagged PSA*, with approximate offered load $\lambda(t - E[S_e])E[S]$. Since the service-time distribution is exponential, S_e and S have a common exponential distribution, and the lagged-PSA offered load is just $\lambda(t-E[S])E[S]$. The good performance of lagged PSA is consistent with the various

Figure 7: **Hourly call volumes to a medium-size financial-services call center.**

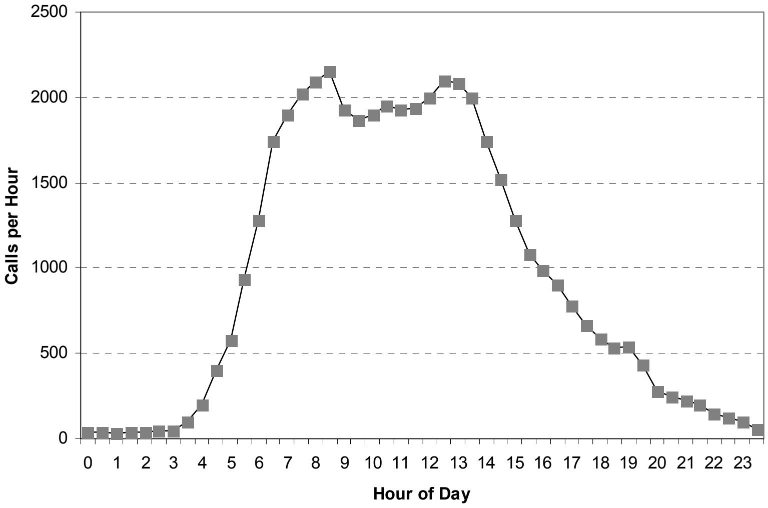
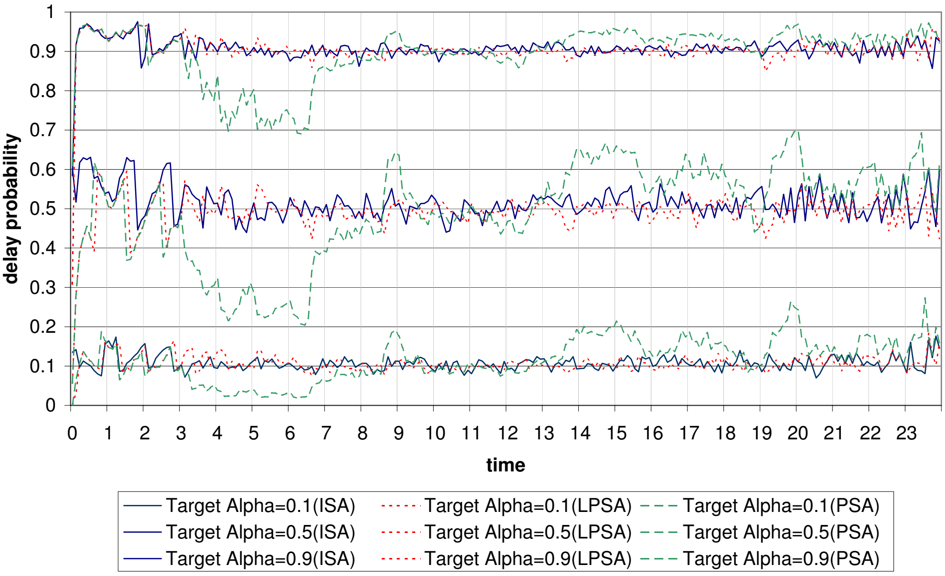
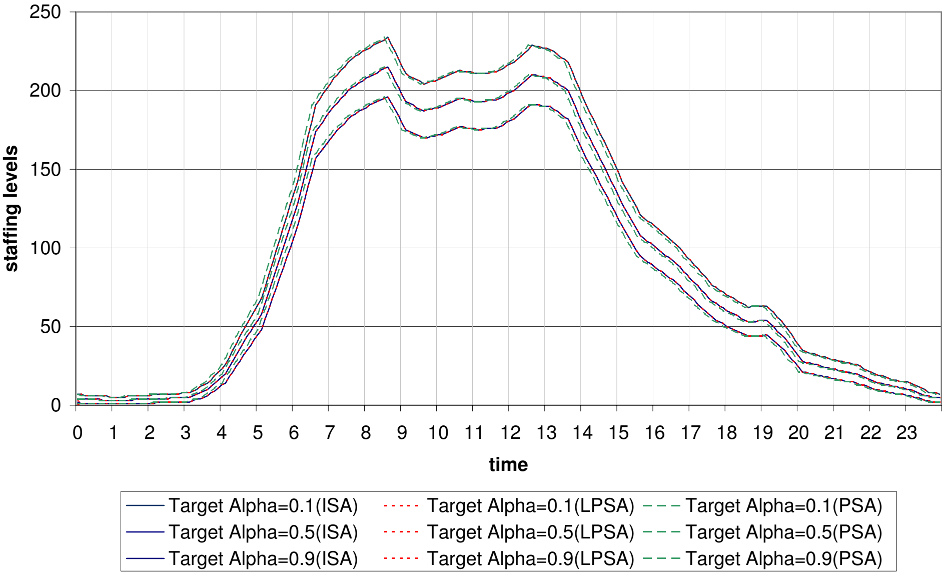


Figure 8: **A comparison of ISA, PSA and lagged PSA for the same three delay-probability targets.**



refinements proposed by Green et al. (2001). We show that simple PSA performs worse than ISA and lagged PSA by plotting the delay probabilities for these three staffing rules in Figure 8. The performance of simple PSA here is nowhere near as bad as it was in the challenging $M_t/M/s_t$ example in Jennings et al. (1996), and as it is for the example here in §4 (see the Internet Supplement), but there are clear departures from the performance targets in Figure 8. The PSA delay probabilities are significantly below the targets during the hours $4-6$ with

Figure 9: **A comparison of staffing levels based on ISA, PSA and lagged PSA for the realistic example, for three delay-probability targets: 0.1, 0.5 and 0.9.**



rapidly increasing arrival rates. The differences among the corresponding staffing functions in Figure 9 look small, but those small differences can have a significant impact, because the arrival-rate function changes rapidly.

We also observe that ISA is not as successful as before, because the target delay probability is not achieved accurately at the beginning and at the end of the day. This phenomenon is even more evident for other performance measures; see the Internet Supplement. However, this weak performance is due to the extremely low arrival rates that prevail at the beginning and the end of the day. When the load is small, the addition or removal of a single server will greatly affect the delay probability. On the positive side, there is a clear time-interval - from hours 5 to 18, in which all performance measures are stable. Finally, we remark that there is excellent matching between the Garnett function and the empirical results, just as in Figure 6; see the Internet Supplement.

6. Theoretical Support in the Case $\theta = \mu$

Relation to other models. In one special case, we can analyze the $M_t/M/s_t+M$ model in considerable detail. That is the case we considered in §4 and §5, in which $\theta = \mu$. (As in §4, we let those both be 1.) With the condition $\theta = \mu$, it is easy to relate the $M_t/M/s_t+M$ model to, first, the corresponding $M_t/M/\infty$ model with the same arrival-rate function and service rate and, second, a corresponding family of steady-state distributions of stationary $M/M/s+M$ models, indexed by t, with the same service and abandonment rates, but with special arrival rate that depends on time t.

Let $\{s_t : t \geq 0\}$ be an arbitrary staffing function. For simplicity, assume that all systems start empty in the distant past (at time $-\infty$). By having $\lambda(t) = 0$ for $t \leq t_0$, we can start arrivals at any time t_0. The first observation is that, for any arrival-rate function $\{\lambda(t) : t \geq 0\}$ and any staffing function $\{s_t : t \geq 0\}$, the stochastic process $\{N_t : t \geq 0\}$ in the $M_t/M/s_t + M$ model with $\theta = \mu$ has the same distribution (finite-dimensional distributions) as the corresponding process $\{N_t^{\infty} : t \geq 0\}$ in the $M_t/M/\infty$ model, because the birth and death rates are the same.

The second observation is that, for both these models, the individual random variables N_t and N_t^{∞} have the same Poisson distribution as the steady-state number in system $N_{\infty}^{(t)}$ in the corresponding stationary model with arrival rate m_t^{∞}.

Waiting times and abandonment probabilities. Let W_t be the *virtual waiting time* at time t (until service, i.e., the waiting time in queue that would be spent by an infinitely patient customer arriving at time t), and let P_t^{ab} be the *virtual abandonment probability* at time t (i.e., the probability of abandonment for an arrival that would occur at time t), both in the $M_t/M/s_t + M$ model. These quantities are considerably more complicated than N_t.

Even though it is difficult to evaluate the full distribution of W_t, we can immediately evaluate the virtual delay probability, because it clearly depends only on what the customer encounters upon arrival at time t. Hence, we have

$$P(W_t > 0) = P(N_t \geq s_t) = P(N_t^{\infty} \geq s_t) = P(Poisson(m_t^{\infty}) \geq s_t) \ , \qquad (6.1)$$

where m_t^{∞} is the offered load in (3.2), just as in (3.5), only here the infinite-server approximation is exact.

Next we observe that $P_t^{ab} = E[F(W_t)]$, where F is the time-to-abandon cdf, so that it suffices to determine the waiting-time distribution. Here is an important initial observation: Conditional on the event that $W_t > 0$, whose probability we have characterized above, W_t is distributed (exactly) as the first passage time of the (Markovian) stochastic process $\{N_u : u \geq t\}$ from the initial value N_t encountered at time t down to the staffing function $\{s_u : u \geq t\}$, provided that we ignore all future arrivals after time t. In other words, W_t is distributed as the first passage time of the pure-death stochastic process with state-dependent death rate N_u

for $u \geq t$ down from the initial value N_t to the curve $\{s_u : u \geq t\}$. As a consequence, the distribution of W_t and the value of P_t^{ab} depend on only N_t and the future staffing levels, i.e., $\{s_u : u \geq t\}$. The time-dependent arrival-rate function contributes nothing further.

It is easy to see that we can establish stochastic bounds on the distribution of W_t if the staffing level is monotone after time t: then setting $s_u = s_t$ for all $u \geq t$ will yield a bound. We can go further based on this observation if we make approximations. If the number of servers is large, then W_t will tend to be small, so that it is often reasonable to make the approximation $s_u \approx s_t$ for all $u > t$. We make this approximation, not because the staffing level should be nearly constant for all u after t, but because we think we only need to consider times u slightly greater than t.

If the future-staffing-level approximation held as an equality, then we would obtain the following approximations as equalities: $W_t \approx W_\infty$ and $P_t^{ab} \approx P_\infty^{ab}$, where the constant staffing level in the stationary $M/M/s+M$ model on the righthand sides is chosen to be s_t and the constant arrival rate is chosen to be λ_t^{MOL} in (3.7). Given these approximations, we can use established results for the stationary $M/M/s+M$ model, e.g., as in Garnett et al. (2002) and Whitt (2005). Algorithms to compute the (exact) distribution of W_∞ are given there, including the corresponding conditional distributions obtained when we condition on whether or not the customer eventually is served.

7. Algorithm Dynamics

In this section we establish the convergence of ISA for the $M_t/M/s_t+M$ model. In doing so, we disregard statistical error caused by having to estimate the delay probabilities associated with each staffing function in the simulation.

To prove convergence, we use sample-path stochastic order, as in Whitt (1981). We say that one stochastic process $\{N_t^{(1)} : 0 \leq t \leq T\}$ is stochastically less than or equal to another, $\{N_t^{(2)} : 0 \leq t \leq T\}$, in *sample-path stochastic order* and write

$$\{N_t^{(1)} : 0 \leq t \leq T\} \leq_{st} \{N_t^{(2)} : 0 \leq t \leq T\} \ , \tag{7.1}$$

if

$$E\left[f\left(\{N_t^{(1)} : 0 \leq t \leq T\}\right)\right] \leq_{st} E\left[f\left(\{N_t^{(2)} : 0 \leq t \leq T\}\right)\right] \tag{7.2}$$

for all nondecreasing real-valued functions f on the space of sample paths. We have ordinary stochastic order for the individual random variables $N_t^{(1)}$ and $N_t^{(2)}$ and write $N_t^{(1)} \leq_{st} N_t^{(2)}$ if $E[f(N_t^{(1)})] \leq E[f(N_t^{(2)})]$ for all nondecreasing real-valued functions on the real line; see

Chapter 9 of Ross (1996) and Müller and Stoyan (2002). Clearly, sample-path stochastic order as in (7.1) implies ordinary stochastic order for the individual random variables for all t. For the convergence, we only need ordinary stochastic-order for each time t, but in order to get that, we need to properly address what happens before time t as well.

Here is the key stochastic-comparison property for the $M_t/M/s_t + M$ model:

Theorem 7.1. (stochastic comparison) *Consider the $M_t/M/s_t+M$ model on the time interval $[0,T]$, starting empty at time* 0*. If $r \geq 1$ and $s_t^{(1)} \leq s_t^{(2)}$ for all t, $0 \leq t \leq T$, or if $r \leq 1$ and $s_t^{(1)} \geq s_t^{(2)}$ for all t, $0 \leq t \leq T$, then*

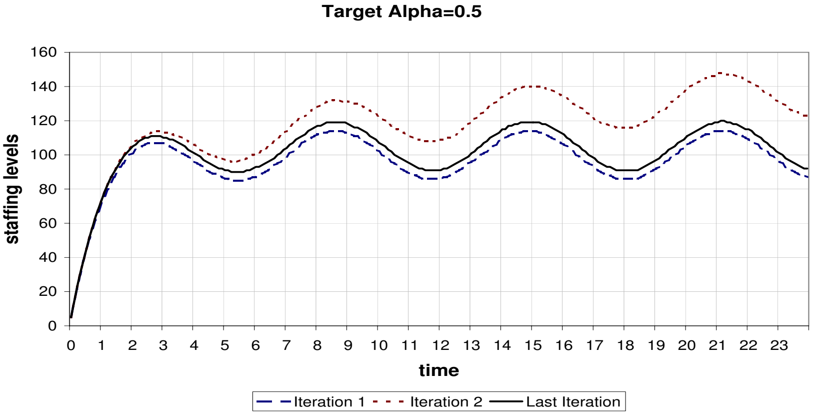
$$\{N_t^{(1)} : 0 \leq t \leq T\} \leq_{st} \{N_t^{(2)} : 0 \leq t \leq T\} \ . \tag{7.3}$$

Proof. Here is the key fact: The death rates depend systematically on the number of servers s_t. When $r > 1$ ($r < 1$), the death rates at time t decrease (increase) as s_t increases. The ordering of the death rates in the two birth-and-death processes makes it possible to achieve the sample-path ordering. Indeed, we justify the relation (7.3) by constructing special versions of the two stochastic processes on the same underlying probability space so that the sample paths are ordered with probability 1. As discussed in Whitt (1981), and proved by Kamae et al. (1978), that special construction is actually equivalent to the sample-path stochastic ordering in (7.3). The sample-path ordering obtained ensures that a departure occurs in the lower process whenever it occurs in the upper process and the two sample paths are equal. To start the construction, we let the two processes be given identical arrival streams. Then we construct all departures (service completions or abandonments) from those of the lower process at epochs when the two sample paths are equal. Suppose that at time t the sample paths are equal: $N_t^{(1)} = N_t^{(2)} = k$. Then, at that t, the death rates in the two birth and death processes are necessarily ordered by $\delta_1(k) \geq \delta_2(k)$. We only let departures occur in process 2 when they occur in process 1, so the two sample paths can never cross over. When a departure occurs in process 1 with both sample paths in state k, we let a departure also occur in process 2 with probability $\delta_2(k)/\delta_1(k)$, with no departure occurring in process 2 otherwise. This keeps the sample paths ordered w.p. 1 for all t. At the same time, the two stochastic processes individually have the correct finite-dimensional distributions. ■

The simulation experiments show that the way the staffing functions converge to the limit depends on the ratio $r \equiv \theta/\mu$: Whenever $r > 1$, we encounter monotone dynamics. Whenever $r < 1$, we encounter oscillating dynamics; and whenever $r = 1$, we encounter instantaneous convergence. As shown in §6, when $r = 1$, the number in system is independent of the staffing function, so we obtain convergence in one step.

An example of the oscillating dynamics is shown in Figure 10, where staffing levels are shown for the first two and final iterations for the model in §4 with $\mu = 1$ and $r = \theta = 0$ (no abandonment).

Figure 10: **Oscillating algorithm dynamics for the model in §4 when $r = \theta = 0$: staffing levels in the 1^{st}, 2^{nd} and final iterations.**



Theorem 7.2. (convergence) *Consider the $M_t/M/s_t + M$ model on the time interval $[0, T]$, starting empty at time 0. Suppose that we consider piecewise-constant staffing functions that only can change at multiples of $\Delta > 0$. Suppose that in each iteration n we can obtain the actual stochastic process $\{N_t^{(n)} : 0 \le t \le T\}$ associated with the staffing function $\{s_t^{(n)} : 0 \le t \le T\}$ (without statistical error). Suppose that $s_t^{(0)} = \infty$ for all t, $0 \le t \le T$.*

(a) If $r > 1$, then $s_t^{(n)} \le s_t^{(m)}$ for all $n > m \ge 0$ and there exists a positive integer n_0 such that

$$s_t^{ISA} = s_t^{(n_0)} = s_t^{(n)} \quad \text{for all} \quad t \quad \text{and} \quad n \ge n_0 \; . \tag{7.4}$$

(b) If $r < 1$, then there exist 2 subsequences $\{s_t^{(2n)}\}$ and $\{s_t^{(2n+1)}\}$, such that $s_t^{(2n)} \downarrow s_t^{(even)}$ and $s_t^{(2n+1)} \uparrow s_t^{(odd)}$.

$$s_t^{(0)} \ge s_t^{(2n)} \ge s_t^{(2n+2)} \ge s_t^{(2n+3)} \ge s_t^{(2n+1)} \ge s_t^{(1)} \tag{7.5}$$

for all t, $0 \le t \le T$, and for all $n \ge n_0$. Moreover, there exists a positive integer n_0 such that

$$s_t^{(2n)} = s_t^{(2n_0)} = s_t^{even} \ge s_t^{odd} = s_t^{(2n_0+1)} = s_t^{(2n+1)} \tag{7.6}$$

for all t*,* $0 \leq t \leq T$*, and for all* $n \geq n_0$.

Proof. Given that $s_t^{(0)} = \infty$, we necessarily have $s^(0)_t > s_t^{(1)}$ for all t, $0 \leq t \leq T$. Hence we have the ordering of the initial ordering of the staffing functions that lets us apply the stochastic order. We then proceed recursively. As a consequence of the sample-path stochastic order, we get ordinary stochastic order in (7.3), we get ordinary stochastic order $N_t^{(1)} \leq_{st} N_t^{(2)}$ for all t. Ordinary stochastic order is equivalent to the tail probabilities being ordered: $P(N_t^{(1)} > x) \leq P(N_t^{(2)} > x)$ for all x, which implies the ordering for the staffing functions at time t. In particular, suppose that

$$P\left(N_t^{(2)} \geq s_t^{(2)}\right) \leq \alpha < P\left(N_t^{(2)} \geq s_t^{(2)} - 1\right) .$$

Since $P\left(N_t^{(1)} \geq s_t^{(2)}\right) \leq P\left(N_t^{(2)} \geq s_t^{(2)}\right) \leq \alpha$, necessarily $s_t^{(1)} \leq s_t^{(2)}$.

Case 1: $r > 1$**.** For $s_t^{(0)} = \infty$, we necessarily start with $s_t^{(0)} > s_t^{(1)}$ for all t, which produces first $N_t^{(1)} \leq_{st} N_t^{(0)}$ and then $s_t^{(2)} \leq s_t^{(1)}$ for all t. Continuing, we get $N_t^{(n)}$ stochastically decreasing in n and $s_t^{(n)}$ decreasing in n, again for all t. Since the staffing levels are integers, if we use only finitely many values of t, as in our implementation, then we necessarily get convergence in finitely many steps.

Case 2: $r < 1$. For $s_t^{(0)} = \infty$, we again necessarily start with $s_t^{(0)} > s_t^{(1)}$ for all t. That produces first $N_t^{(1)} \geq_{st} N_t^{(0)}$ and then $s_t^{(0)} \geq s_t^{(2)} \geq s_t^{(1)}$ for all t. Afterwards, we get $N_t^{(1)} \geq_{st} N_t^{(2)} \geq_{st} N_t^{(0)}$ and $s_t^{(0)} \geq s_t^{(2)} \geq s_t^{(3)} \geq s_t^{(1)}$ for all t. Continuing, we get $N_t^{(2n)}$ stochastically increasing in n, while $N_t^{(2n+1)}$ stochastically decreases in n, for all t. Similarly, $s_t^{(2n)}$ decreases in n, while $s_t^{(2n+1)}$ increases in n for all t. We thus have convergence, to possibly different limits. Since the staffing levels are integers, if we use only finitely many values of t, as in our implementation, then we necessarily get convergence in finitely many steps. ■

We remark that we also obtain the convergence in Theorem 7.2 with other initial conditions. In particular, it suffices to let $s_t^{(0)}$ be sufficiently large for all t. For $r > 1$, it suffices to have $s_t^{(0)} \geq s_t^{ISA}$ for all t. For $r < 1$, it suffices to have $s_t^{(0)} \geq s_t^{even}$ for all t.

We conclude this section by making some empirical observations, for which we have yet to develop supporting theory. We also observed that the target delay probability α strongly influenced the dynamics. In particular, higher values of α cause larger oscillations in the oscillating case, and slower convergence to the limit in all cases. Finally, we also observed a time-dependent behavior in the convergence of $s_t^{(n)}$. We observed a greater gap as time increased. For example, let $I_t \equiv \inf\{j : s_t^{(i)} = s_t^{(j)} \quad \text{for all} \quad i \geq j\}$. We observed that $I_{t_2} \geq I_{t_1}$

for all $t_2 > t_1$. An illustration can be viewed in Figure 10. This time-dependent behavior is understandable, because the gap between two different staffing levels persists across time, so that there is a gap in the death rates at each t. Hence, as t gets larger, the two processes can get further apart. Thus the gap can first decrease more at the initial times. When it reaches the limit at earlier times, the gap will still have to decrease more at later times.

8. Conclusions

We have developed a simulation-based algorithm - ISA - that generates staffing functions for which performance has been shown to be stable in the face of time-varying arrival rates for the $M_t/M/s_t + M$ model. The results have been found to be remarkably robust, applying to all forms of time variation in the arrival-rate function, with or without abandonment, covering the ED, QD and QED operational regimes. All experiments were done with nine target delay probabilities, ranging from $\alpha = 0.1$ (QD) to $\alpha = 0.9$ (ED). In §7 we proved that the ISA converges for the $M_t/M/s_t + M$ model.

In our simulation experiments, we found that ISA performs essentially the same as the modified-offered-load (MOL) approximation (reviewed in §3) with and without customer abandonment. Thus we provided additional support for MOL and the square-root-staffing formula in (3.6) based on it (using arrival rate λ_t^{MOL} in (3.7)). As we saw in §5, in many applications the MOL approximation is well approximated itself by lagged PSA and, in easy cases, by PSA itself. To implement the MOL approximation with abandonments, we applied many-server heavy-traffic limits from Garnett et al. (2002), which yield the Garnett function in (4.4); just as Jennings et al. (1996) applied applied many-server heavy-traffic limits from Halfin and Whitt (1981) without customer abandonment.

Finally, we found that the simple approach of *staffing to the offered load* is remarkably effective in the QED regime (when $\alpha = 0.5$). That was substantiated time and again by having the ISA staffing function s_t^{ISA} fall on top of the offered load m_t^{∞}, as in case 3 in Figure 2. Of course, abandonment plays an important role; the staffing is always above the offered load without abandonment. When the service times are short, the offered load m_t^{∞} may agree closely with the PSA offered load $m_t^{PSA} \equiv \lambda(t)E[S]$; then staffing to the offered load reduces to the *naive deterministic approximation*: staffing to the PSA offered load m_t^{PSA}. However, it is good to be careful, because even for the realistic example in §5, PSA performed significantly worse than ISA, MOL and lagged PSA.

There is much yet to be done. Here are some natural next-steps:

1. As discussed in Section 4, for the $M_t/M/s_t + M$ model, it remains to explore alternative staffing methods to achieve better time-stability of abandonment probabilities and expected waiting times, especially under heavy loads, but experience indicates that the delay probability is a good performance target.

2. A great advantage of ISA is its generality. However, it remains to explore the ISA for additional queueing systems. We already have had partial (successful) results for deterministic and log-normal service-time distributions. It remains to consider other service-time distributions for the same models; it remains to consider other models. Some other models to analyze appear in Mandelbaum et al. (1998), e.g., queues with retrials and priority classes. Of special interest for actual call centers are multi-class models with skill-based routing. For call centers, our ultimate goal is to treat realistic multi-server systems with multiple call types and skill-based routing (SBR), but that remains to be done. In that setting, it is natural to apply SBR methods for stationary models after using the MOL approximation in (3.7) for each call type at time t. Once we have reduced the problem to a stationary SBR model, we may be able to apply the staffing method in Wallace and Whitt (2005). Approaches based on these ideas remain to be investigated. With networks of queues, the MOL approach can be applied together with results for networks of infinite-server queues; see Massey and Whitt (1993).

3. We proved that ISA converges for the $M_t/M/s_t + M$ model and we observed that it usually does so quite quickly, but it remains to analyze convergence of the algorithm more generally. Even for the $M_t/M/s_t + M$ model, some of the phenomena have not yet been adequately explained.

4. For one special case - the one with $\theta = \mu$ - we have provided strong theoretical support for our methods in §6 and the Internet Supplement. In the Internet Supplement we exploited the mathematical framework of service networks in Mandelbaum et. al.(1998). It would be nice to prove much more generally that, under proper scaling, the actual time-dependent probability of delay under ISA indeed converges to the specified target as scale increases.

9. Acknowledgments

The reported research was supported by Grant No. 2002112 from the United States–Israel Binational Science Foundation (BSF). William A. Massey and Avishai Mandelbaum were also supported by NSF grant DMI-0323668, while Ward Whitt was also supported by NSF grant DMI-0457095.

References

[1] Bolotin, V. 1994. Telephone circuit holding-time distributions. In *Proceedings of the International Teletraffic Congress, ITC* 14, J. Labetoulle and J. W. Roberts (eds.), North-Holland, Amsterdam, 125-134.

[2] Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn and L. Zhao. 2005. Statistical analysis of a telephone call center: a queueing-science perspective. *J. Amer. Statist. Assoc.* 100, 36–50.

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

[4] Eick, S., Massey, W. A., Whitt, W. 1993b. $M_t/G/\infty$ Queues with Sinusoidal Arrival Rates. *Management Science*, **39**(2), 241-252.

[5] Gans, N., Koole, G., Mandelbaum, A. 2003. Telephone Call Centers: Tutorial, Review and Research Prospects. *Manufacturing and Service Operations Management*, **5**(2), 79–141.

[6] Garnett, O., Mandelbaum, A., Reiman, M. I. 2002. Designing a Call Center with Impatient Customers. *Manufacturing and Service Operations Management*, **4**(3), 208–227.

[7] Green, L. V., Kolesar, P. J. 1991. The Pointwise Stationary Approximation for Queues with Nonstationary Arrivals. *Management Science*, **37**(1), 84–97.

[8] Green, L. V., Kolesar, P. J., Soares, J. 2001. Improving the SIPP Approach For Staffing Service Systems That Have Cyclic Demand. *Operations Research*, **49**, 549–564.

[9] Green, L. V., Kolesar, P. J., Whitt, W. 2005. Coping with Time-Varying Demand when Setting Staffing Requirements for a Service System. *Production and Operations Management*, forthcoming. Available at: http://www.columbia.edu/∼ww2040/Coping.pdf

[10] Halfin, S., Whitt, W. 1981. Heavy-Traffic Limits for Queues with Many Exponential Servers. *Operations Research*, **29**, 567–587.

[11] Jagerman, D. L. 1975. Nonstationary Blocking in Telephone Traffic. *Bell System Technical Journal*, **54**, 625–661.

[12] Jennings, O. B., Mandelbaum, A., Massey, W. A., Whitt, W. 1996. Server Staffing to Meet Time-Varying Demand. *Management Science*, **42**(10), 1383–1394.

[13] Kamae, T., Krengel, U., O'Brien, G. L. 1978. Stochastic Inequalities on Partially Ordered Spaces. *Annals of Probability* **5**, 899–912.

[14] Mandelbaum, A., Massey, W.A., Reiman, M. I. 1998. Strong Approximations for Markovian Service Networks. *Queueing Systems: Theory and Applications (QUESTA)*, **30**, 149–201.

[15] Massey, W. A., Whitt, W. 1993. Networks of Infinite-Server Queues with Nonstationary Poisson Input. *Queueing Systems* **13** (1), 183–250.

[16] Massey, W. A., Whitt, W. 1994. An Analysis of the Modified Offered Load Approximation for the Erlang Loss Model. *Annals of Applied Probability*, **4**, 1145–1160.

[17] Massey, W. A., Whitt, W. 1997. Peak Congestion in Multi-Server Service Systems with Slowly Varying Arrival Rates. *Queueing Systems*, **25**, 157–172.

[18] Massey, W. A., Whitt, W. 1998. Uniform Acceleration Expansions for Markov Chains with Time-Varying Rates. *Annals of Applied Probability*, **9** (4), 1130–1155.

[19] Müller, A., Stoyan, D. 2002. *Comparison Methods for Stochastic Models and Risks*, Wiley.

[20] Ross, S. M. 1990. *A Course in Simulation*, Macmillan.

[21] Ross, S. M. 1996. *Stochastic Processes*, second edition, Wiley.

[22] Wallace, R. B., Whitt, W. 2005. A Staffing Algorithm for Call Centers with Skill-Based Routing. *Manufacturing and Service Operations Management*, forthcoming.
Available at: http://www.columbia.edu/∼ww2040/recent.html

[23] Whitt, W. Comparing Counting Processes and Queues. 1981. *Advances in Applied Probability* **13** 207–220.

[24] Whitt, W. 1991. The Pointwise Stationary Approximation for $M_t/M_t/s$ Queues Is Asymptotically Correct as the rate Increases. *Management Science*, **37**(2), 307–314.

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

[26] Whitt, W. 2005. Engineering Solution of a Basic Call-Center Model. *Management Science*, **51**, 221–235.