Staffing of Time-Varying Queues to Achieve Time-Stable Performance

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

May 12, 2005

Abstract

Continuing research by Jennings, Mandelbaum, Massey and Whitt (1996), we investigate methods to perform time-dependent staffing for many-server queues. Our aim is to achieve time-stable performance in face of general time-varying arrival rates. As before, we target a stable probability of delay. Motivated by telephone call centers, we focus on many-server models with customer abandonment, especially the Markovian $M_t/M/s_t+M$ model, having an exponential time-to-abandon distribution (the $+M$), an exponential service-time distribution and a nonhomogeneous Poisson arrival process. We develop three different methods for staffing, with decreasing generality and decreasing complexity: (i) a simulation-based iterative-staffing algorithm (ISA), (ii) the square-root-staffing rule with service grade determined by the modified-offered-load approximation, and (iii) simply staffing at the offered load itself.

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

1 Introduction

Service systems such as banks, insurance companies and hospitals play an important role in our society. Services employ about 60–80% of the work force in western economies, and their importance is sharply on the rise, both within service and manufacturing companies. In our service-driven economy, it is estimated that over 70% of the business transactions are carried out over the phone. Most of these transactions are processed by telephone call centers, which have become the preferred and prevalent means for companies to communicate with their customers. Indeed, it is estimated that more than 3% of the U.S. work force is employed in call centers—more than in agriculture! For an overview of call centers and models of them, readers are referred to the recent review by Gans, Koole and Mandelbaum (2003).

The modern call center is a highly complex operation that fuses advanced technology and human beings. But the economic and managerial significance of the latter clearly outweighs the former. More specifically, labor costs (agents' salaries, training, etc.) typically run as high as 70% of the total operating costs of a call center, and attrition rates in call centers reach anywhere from 30% per year (considered low) to over 200% at times. In such circumstances, perhaps the most important operational decision to be made is staffing: what is the appropriate number of telephone agents that are to be accessible for serving calls. Overstaffing is wasteful, while understaffing leads to low service levels and overworked agents.

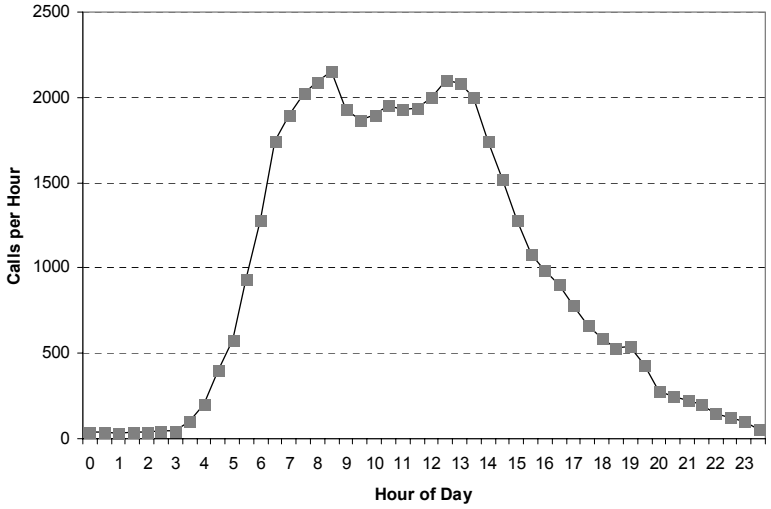
The staffing problem typically takes the following form: Under an existing operational reality, and given a desired quality of service, we seek the least number of agents at each time that is required to meet a given service-level constraint. This problem, which has received much attention over the years (see Section 4 in Gans et. al.), is challenging both theoretically and practically. The challenges are easy to understand, because the natural model for the staffing problem is a many-server queue with a time-varying arrival rate, which is notoriously difficult to analyze. The practical importance of staffing is highlighted by considering a bank employing 10,000 telephone agents and catering to millions of customers per day; even small gains in operational efficiency or service quality clearly can provide great benefit.

Figure 1 depicts a typical arrival-rate function to a telephone call center. Call volumes are low around midnight (hour 0), starting to increase in the early hours of the morning, peaking at late morning, then dropping somewhat around midday (12, lunch break), rising again afterwards, and then dropping thereafter to midnight levels. The displayed arrival-rate function is an average of several similar days; the actual number of arrivals, in a given

hour on a given day, fluctuates randomly around this average. (The functional form in Figure 1 is typical; the particular values for the arrival rates were adapted from Green, Kolesar and Soares (2001).)

Staffing planners are thus faced with two sources of variability: **predictable variability** – time-variations of the expected load – and **stochastic variability** – random fluctuations around this time-dependent average. Most available staffing algorithms are designed to cope only with stochastic variability; they avoid the predictable variability in various ways. For example, when the service times are relatively short (a few minutes), as in many call centers when service is provided by a telephone call, it is usually reasonable 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). With PSA, one performs a stationary or steady-state analysis with a stationary model having parameters that vary by the time of day; see Green and Kolesar (1991), Whitt (1991), Massey and Whitt (1998) and references therein.

Figure 1: **Hourly call volumes to a medium-size call center**



However, service times are not always short, even in call centers. If relatively lengthy interactions are not uncommon, then PSA tends to be inappropriate. When service times are not so short, significant predictable variability can cause PSA to produce poor performance. As a consequence, some parts of the day may be overstaffed, while others are understaffed; see Green et al. (2005) for additional discussion.

In this paper we address the staffing problem with *both* predictable and stochastic variability. Here is the problem we aim to solve: **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.

In particular, we aim to find an appropriate time-dependent staffing function for **any** arrival-rate function, where "appropriate" means that we achieve time-stable performance. For given service-time distribution, we allow arbitrary arrival-rate functions, i.e., arbitrary predictable variability. We aim to agree with PSA when it is appropriate and do significantly better when it is not appropriate. We emphasize the importance of achieving stable performance. Stable performance is good both for managers (easier to manage) and customers (know what to expect). With stable performance, the nearly-constant quality of service is easily adjusted up or down, as desired.

Here is how the rest of this paper is organized: We start in §2 by briefly reviewing the previous contributions by Jennings et al. (1996). We then overview our main contributions in §3. In §4 we specify our iterative-staffing algorithm in detail. In §5 we illustrate the performance of our algorithm by considering a time-varying Erlang-A-model example (with abandonment). In §6, for comparison, we consider a similar time-varying Erlang-C-model example (without abandonment). In §7 we present some supporting theory. Finally, in §8, we discuss the dynamics of the iterative algorithm, establishing monotonicity and convergence results.

We present additional material in a longer unabridged version available on line as an Internet Supplement. There we revisit the "challenging example" in Jennings et al. (1996). We expand the analysis of the time-varying Erlang-A example from §5 by considering different patience parameters. We also analyze a realistic example - the one presented in Figure 1. In contrast to Green et. al. (2001), we incorporate abandonment, which significantly impacts staffing results.

2 Our Point of Departure

Our point of departure is our (with Otis B. Jennings) previous paper: Jennings, Mandelbaum, Massey and Whitt (1996). There we considered the $M_t/G/s_t$ model (without customer abandonment), having a nonhomogeneous Poisson arrival process with arrival-rate function $\lambda(t)$ and *independent and identically distributed* (IID) service times $\{S_n : n \geq 1\}$. The service times are distributed as a random variable S with a general *cumulative distribution function* (cdf) G having mean $E[S] = 1/\mu$.

Let L_t be the number of customers in the $M_t/G/s_t$ system, either waiting or being served, at time t. We focused on the probability of delay, aiming to choose the time-dependent staffing level s_t such that

$$P(L_t \geq s_t) \leq \alpha < P(L_t \geq s_t - 1) \quad \text{for all} \quad t\,, \tag{2.1}$$

where α is the target delay probability. That problem is challenging because the time-dependent delay probability $P(L_t \geq s_t)$ depends on the staffing function before time t as well as at time t.

We proposed an **infinite-server approximation**. In particular, we proposed approximating the random variable L_t by the number L_t^∞ of busy servers in the associated $M_t/G/\infty$ model, having the same arrival process and service times. We thus choose the desired staffing function s_t so that the inequalities in (2.1) hold when L_t^∞ is substituted for L_t. That approximation provides great simplification because (i) the tail probability $P(L_t^\infty \geq s_t)$ 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 L_t^∞ is known.

In particular, as reviewed in Eick et al. (1993a), for each t, L_t^∞ has a **Poisson distribution** whenever the number in the system at time $t = 0$ 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, which depends on t. We next apply a normal approximation for the Poisson distribution, using the fact that the variance equals the mean for a Poisson distribution. We obtain the **normal approximation**

$$\begin{aligned} P(L_t \geq s_t) &\approx P(L_t^\infty \geq s_t) \\ &\approx P(N(m_t, m_t) \geq s_t) = P\left(N(0,1) \geq \frac{s_t - m_t}{\sqrt{m_t}}\right) = 1 - \Phi\left(\frac{s_t - m_t}{\sqrt{m_t}}\right)\,, \end{aligned} \tag{2.2}$$

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

An immediate consequence of the normal approximation (2.2) is the **square-root-staffing formula** for the $M_t/G/s_t$ model:

$$s_t = m_t + \beta\sqrt{m_t}, \quad 0 \leq t \leq T, \tag{2.3}$$

where the constant β is a measure of the **quality of service** and the deterministic function m_t is the mean number of busy servers in the associated $M_t/G/\infty$ infinite-server model. (We would let s_t be the least integer greater than the righthand side.) Combining the target in (2.1) and the normal approximation in (2.2), we see

that the quality of service β in (2.3) should be chosen so that

$$1 - \Phi(\beta) = \alpha \, . \tag{2.4}$$

It is also significant that, for the $M_t/G/\infty$ model, the time-dependent mean number of busy servers, m_t, has a **tractable expression**: The explicit formula for m_t is

$$m_t \equiv E\left[L_t^{\infty}\right] = \int_{-\infty}^{t} G^c(t-u)\lambda(u)\, du = E\left[\int_{t-S}^{t} \lambda(u)\, du\right] = E\left[\lambda(t-S_e)\right] E[S] \, , \tag{2.5}$$

where 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

$$G_e(t) \equiv P(S_e \le t) \equiv \frac{1}{E[S]} \int_0^t \left[1 - G(u)\right] du, \quad t \ge 0 \, ; \tag{2.6}$$

with k^{th} moment

$$E[S_e^k] = \frac{E[S^{k+1}]}{(k+1)E[S]} \, ; \tag{2.7}$$

see Theorem 1 of Eick et al. (1993a) and references therein. For more on the stationary-excess cdf G_e, see pp. 424 and 431 of Ross (2003); $G = G_e$ if and only if G is exponential. Moreover, the time-dependent mean has a convenient approximation, based on a second-order Taylor-series approximation for λ about t. In particular, the time-dependent mean can be approximated in terms of the first two moments of S_e, the mean of S and the second derivative of the arrival-rate function at t, $\lambda^{(2)}(t)$ via

$$m_t \approx \lambda(t - E[S_e])E[S] + \frac{\lambda^{(2)}(t)}{2} Var(S_e)E[S] \, ; \tag{2.8}$$

see Theorem 9 of Eick et al. (1993a).

In Section 4 of Jennings et al. we also introduced a refined approximation for the time-dependent delay probabilities that is tantamount to a **modified-offered-load** (MOL) approximation, as in Jagerman (1975) and Massey and Whitt (1994, 1997). The modified-offered-load approximation for L_t in the $M_t/G/s_t$ model at time t is the stationary number in system L_∞ in the corresponding stationary $M/G/s$ model (with the same service-time distribution and the same number of servers s_t), but using the infinite-server mean $m_t \equiv E[L_t^{\infty}]$ in (2.5) as the offered load operating at time t. Equivalently, that means letting the homogeneous Poisson arrival process in the stationary $M/G/s$ model have rate

$$\hat{\lambda}_t \equiv \frac{m_t}{E[S]} = m_t \mu \quad \text{at time} \quad t \, , \tag{2.9}$$

where m_t is the infinite-server mean in (2.5).

However, the refined modified-offered-load approximation in Jennings et al. did not involve directly applying the steady-state distribution of the $M/M/s$ model. Instead, it applied an approximation for that steady-state distribution based on a many-server heavy-traffic limit from Halfin and Whitt (1981), which produces a simple closed-from formula, namely formula (6.17) in Section 6 here, which we will discuss in more detail later.

The important insight in the modified-offered-load approximation is that the **"right" time-dependent offered load** should be the time-dependent mean number of busy servers in the associated infinite-server model - m_t. For the stationary model, the right offered load is known to be $\lambda E[S]$. The "obvious" direct time-dependent generalization is $\lambda(t)E[S]$, which is the PSA offered load. 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 time-dependent infinite-server model, m_t, is a better generalization of "offered load" than the PSA offered load for most time-varying many-server models. (Indeed, it may be considered exactly the right definition for the infinite-server model itself.)

From (a special case of) Theorem 10 in Eick et al. (1993a), we can **quantify the difference** between the infinite-server offered load m_t and the PSA offered load $\lambda(t) \cdot E[S]$. Letting $(S_e)_e$ be a random variable with the twofold stationary-excess cdf $(G_e)_e$, we have the formula

$$m_t - \lambda(t) \cdot E[S] = E\left[\lambda'\left(t - (S_e)_e\right)\right] \cdot E[S_e] \cdot E[S] = \frac{1}{2} \cdot E\left[\lambda'\left(t - (S_e)_e\right)\right] \cdot E[S^2]. \tag{2.10}$$

From (2.10), it follows that the PSA offered load will *not* be a good approximation of the infinite-server offered load when the arrival rate varies rapidly in time (large derivative λ'). For a given mean service time, they may also be far apart when the second moment of the service time, $E[S^2]$, (or variance) is large. The second condition has implications for non-exponential distributions that are heavy tailed; see Whitt (2000) for background.

In Jennings et al. we did not apply the modified-offered-load approximation directly. Instead of calculating the steady-state delay probability for the stationary Erlang-C 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 QoS parameter β. Moreover, the heavy-traffic limit provides an alternative derivation of the square-root staffing formula in (2.3), without relying on an infinite-server approximation or a normal approximation.

Jennings et al. showed that the method for setting staffing requirements in the $M_t/G/s_t$ model outlined above is remarkably effective. The performance was evaluated by doing numerical comparisons for the $M_t/M/s_t$ special case. For any given staffing function, the time-dependent distribution of L_t in that Markovian model can be derived by solving a system of time-dependent ordinary differential equations. The most important conclusion from those previous experiments 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.

3 Our Contributions Here

In this paper we develop staffing algorithms for more complicated time-varying many-server models, such as many-server queues with abandonment. For example, we treat the much more realistic $M_t/G/s + G$ model with non-exponential service times (the first G) and non-exponential abandonments (the $+G$).

For call centers, our ultimate goal is to treat realistic multi-server systems with multiple call types and skill-based routing (SBR), but we do not pursue that here. In that setting, it is natural to apply SBR methods for stationary models after using the modified-offered-load approximation in (2.9) for each call type at time t. Approaches based on that idea remain to be investigated.

Our first contribution here is a **simulation-based Iterative-Staffing Algorithm (ISA)** for many-server queues with time-varying arrival rate. 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 restricted to a model that is 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 $\{s_t : 0 \le t \le T\}$.

Following Jennings et. al. (1996), we assume that, in principle, any number of servers can be assigned at any time. In our implementation, however, time is divided into short intervals (we take 0.1 service times), and we keep the number of servers fixed over each of these small intervals. The service discipline is FCFS, and servers

follow an exhaustive service discipline: a server that finishes a shift in the middle of a service will complete the service and sign out only when finished. (Our results prevail also for preemptive service disciplines under which servers leave at end-of-shifts and their customers, if any, are moved to the front of the queue.)

In practice, staffing is required to be fixed over longer staffing intervals - typically ranging from 15 minutes to an hour. Here we ignore that constraint. An initial staffing function with such constraints is obtained from our results by using in each staffing interval the maximum required staffing level at any time point within that staffing interval. That will yield an upper bound on the required staffing. Simulation can then be used, in the manner of the ISA, to see if these initial staffing levels can be decreased, while still meeting the performance target.

Continuing to follow Jennings et al. (1996), we use the delay probability as our target performance measure. Specifically, given a target probability of delay, we identify time-varying staffing levels under which the actual probability of delay remains approximately equal to the given target at all times. Other performance measures, such as the average waiting time and queue-length tail delay-probabilities, turn out to be relatively constant over time as well.

For the main model we study, the Markovian $M_t/M/s_t + M$ model, we not only implement and evaluate ISA, but we also provide a proof of convergence. To do so, we must set aside the (important) issue of estimating the time-dependent delay probability for any given staffing function by computer simulation, which is subject to statistical sampling error. That statistical sampling error decreases as we increase the number of independent replications, so it can be made arbitrarily small at the expense of computational effort, but for any given amount of computational effort it is always present. However, if we assume that we actually know the true delay probabilities associated with each staffing function, then we obtain monotone convergence to a limiting staffing function. That is accomplished by applying sample-path stochastic-order notions, as in Whitt (1981).

While working with ISA, we discovered that **the simulation-based solutions have astonishing regularity**. In particular, we found that global performance measures coincide with the performance measures of the associated stationary model. In particular, when we used ISA to staff the time-varying $M_t/M/s_t + M$ model, we found that the resulting staffing could be related to the steady-state behavior of the associated stationary $M/M/s + M$ model. That implies that the modified-offered-load approximation will also work well for the $M_t/M/s_t + M$ model. (We also obtained similar results for $M_t/G/s_t + M$ models with non-exponential service-time distributions.)

That leads us to our second contribution: We extend the **square-root staffing formula** based on the modified-offered-load approximation to the $M_t/M/s_t + M$ model. In particular, we suggest staffing according to the square-root-staffing formula in (2.3), where the QoS parameter $\beta \equiv \beta(\alpha)$ is derived from a theoretical one-to-one relation between α and β for the corresponding stationary model. However, just as in Jennings et al., we do not actually work directly with the steady-state distribution. Instead, for the $M_t/M/s_t + M$ model, we again use explicit formulas relating α to β obtained from a many-server heavy-traffic limit - here the corresponding limit for the $M_t/M/s + M$ model in Garnett, Mandelbaum and Reiman (2002). We justify this simple analytic staffing formula by conducting experiments for the $M_t/M/s_t + M$ model, but we propose the approximation more generally. The effectiveness in any other context can be verified by applying the simulation-based ISA.

Finally, we make yet one more contribution. To describe it, we remind readers of the three heavy-traffic regimes for many-server queues: *Quality-Driven* (QD, lightly loaded), *Efficiency-Driven* (ED, heavily loaded) and *Quality-and-Efficiency-Driven* (QED, normally loaded); see Garnett et al. (2002). In our experiments for the many-server queue with abandonments we found that **simply staffing according to the offered load itself** is remarkably effective in the QED regime, i.e., staffing by letting $s_t = m_t$ for the $M_t/M/s_t + M$ model works very well in the QED regime. Needless to say, abandonments play a crucial role in this property. This is another example of the importance of including abandonments in the model, when customers actually do abandon; see Garnett et al. (2002) for more discussion.

Even though staffing according to the offered load is a remarkably simple method, there remains substantial sophistication, because we have to know that we should use the deterministic offered-load function m_t. When the service times are relatively short (compared to the fluctuations in the arrival-rate function), we can use a truly **naive deterministic approximation**: We can then simply staff according to the PSA offered load: we can set $s_t = \lambda(t)/\mu$ (which will be close to the offered load, m_t, in that scenario). When we staff according to the PSA offered load $\lambda(t)/\mu$, we are truly ignoring all stochastic variability; we are using only deterministic data about the model: the deterministic arrival-rate function $\lambda(t)$ and the deterministic mean service time $1/\mu$. Even though the infinite-server offered load m_t is a deterministic function, it depends on the service-time distribution beyond its mean, as is apparent from (2.5).

We conclude by mentioning that the naive deterministic approximation is remarkably effective in the setting of the realistic large example in Figure 1, when there is customer abandonment in the QED regime. With short service times - a mean of six minutes - as occur in practice, the naive deterministic approximation $\lambda(t)/\mu$,

the time-dependent offered load m_t and the ISA staffing level s_t all fall on top of each other when $\alpha = 0.5$, producing three curves looking just like the one in Figure 1; see the Internet Supplement. Then 50% of the customers are served without delay, stably over the day.

4 The Simulation-Based Iterative-Staffing Algorithm (ISA)

In this section we describe the simulation-based interactive-staffing algorithm (ISA). As indicated before, we determine time-dependent staffing levels aiming to achieve a given constant probability of delay at all times. In the process of applying the ISA, we directly confirm that our goal is being met. Indeed, the goal will necessarily be met, to a specified tolerance, if the algorithm converges. We then can confirm that other performance measures remain relatively stable as well.

For our implementation of the algorithm, we assume that we have an $M_t/G/s_t + G$ model with independent sequences of IID service times and IID times to abandon, which are independent of the arrival process, having general distributions, and a nonhomogeneous Poisson arrival process, which is fully specified by its arrival-rate function $\{\lambda(t); 0 \le t \le T\}$. (It will be evident that our approach extends to more general models.)

To start, we fix an arrival-rate function, a service-time distribution, a time-to-abandon (patience) distribution (when relevant) and a time-horizon $[0, T]$. For any random quantity of interest, let X_t^n denote the value at time t in the n^{th} iteration, for $t \in [0, T]$ (the given time horizon). 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.

In this section, let $s_t^{(n)}$ be the staffing level at time t in iteration n for $0 \le t \le T$. Let $L_t^{(n)}$ denote the random total number of customers in the system at time t, under this staffing function. We estimate the distribution of $L_t^{(n)}$ 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 a large finite number, ensuring that the probability of delay (i.e., of having all servers busy upon arrival) is negligible for all t. For the examples in §5 and §6, it suffices to let $s_t^{(0)} = 200$ for all t.

The algorithm iteratively performs the following steps, until convergence is obtained. (Here, convergence means that the staffing levels do not change much after an iteration. Practically, they are allowed to change by some threshold τ, 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 $L_t^{(i)}$, 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 such that the delay-probability constraint is met at time t; i.e., let

$$s_t^{(i+1)} = \arg\min \{c \in \mathbb{N} : P\left(L_t^{(i)} \ge c\right) \le \alpha\} . \tag{4.11}$$

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

$$\|s^{(i+1)} - s^{(i)}\|_\infty \equiv \max\{|s_t^{(i+1)} - s_t^{(i)}| : 0 \le t \le T\} \le \tau , \tag{4.12}$$

then stop and let $s^{(i+1)}$ be the proposed staffing function. Otherwise, advance to the next iteration, i.e., replace i by $i+1$ and go back to step 1. (We let $\tau = 1$.) ∎

For further discussion, let ∞ denote the index of the last iteration of ISA, so that $s_t^{(\infty)}$ denotes the final staffing level at time t and $L_t^{(\infty)}$ denotes the number in system at time t with that staffing function $s^{(\infty)}$. Then, if the algorithm converges, it converges to a staffing function $s^{(\infty)}$ for which $P\left(L_t^{(\infty)} \ge s_t^{(\infty)}\right) \approx \alpha$, $0 \le t \le T$.

Our implementation of ISA was written in C++. For the special case of the Markovian $M_t/M/s_t + M$ model, we can rigorously establish convergence of the algorithm, as we explain in §8. Experience indicates that the algorithm consistently converges and does so relatively rapidly. The number of iterations required depends on the parameters, especially the ratio $\mathbf{r} \equiv \theta/\mu$, where θ is the individual abandonment rate. If $\mathbf{r} = 1$, corresponding to an infinite-server queue (§7), then no more than two iterations are needed, since the distribution of the number in system does not depend upon the number of servers. As $\mathbf{r}$ departs from 1, the number of required iterations typically increases. For example, when $\mathbf{r} = 10$, the number of iterations can get as high as $6 - 12$. When $\mathbf{r}$ 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 §8.

5 An Example with the Time-Varying Erlang-A Model

We demonstrate the performance of ISA by considering a time-varying Erlang-A model ($M_t/M/s_t + M$) with a sinusoidal arrival-rate function. Let the queueing system be faced with a non-homogeneous Poisson arrival

process with a **sinusoidal arrival-rate function**

$$\lambda(t) = a + b \cdot \sin(ct), \quad 0 \le t \le T\,, \tag{5.13}$$

where $a = 100$, $b = 20$ and $c = 1$. Let the service times and the customer times to abandon (if they have not yet started service) come from independent sequences of independent and identically distributed (IID) exponential random variables, both having mean 1. As can be seen from PSA, the arrival rate is sufficiently large, that about 100 servers are required, so 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 arbitrary arrival-rate functions such as Figure 1.

An important issue, however, is the rate of fluctuation in the arrival-rate function compared to the expected service-time distribution. 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 (5.13) 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. That roughly matches the daily cycle in Figure 1 for the six-hour period around 12:00 noon.

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 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 practical example directly related to Figure 1 in the Internet Supplement.) From this preliminary analysis, we should anticipate that the service times are sufficiently long in our example that the traditional PSA method is likely to perform poorly here, just as in Jennings et al. (1996), and it does. As before, we are deliberately choosing a difficult case.

The arrival rate coincides with the PSA offered load, because the mean service time here is 1. The (infinite-server) offered load is given in (2.5). Since we have a sinusoidal arrival-rate function, we can apply Eick et al. (1993b) to give an explicit formula for the offered-load m_t, i.e., the mean number of busy servers in the

associated infinite-server system. Since the service-time distribution is exponential, we can apply formula (15) of Eick et al. (1993b). For the sinusoidal arrival-rate function in (5.13), the offered load is

$$m_t = a + \frac{b}{1+c^2}[\sin(ct) - c \cdot \cos(ct)] = 100 + 10[\sin(t) - \cos(t)] \ . \tag{5.14}$$

The second formula in (5.14) is based on the specific parameters: $a = 100$, $b = 20$ and $c = 1$.

In order to put our model into perspective, in Figure 2 we plot the offered load m_t in (5.14) for the sinusoidal arrival-rate function in (5.13) 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 20. Note that the 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 size of the fluctuations decrease. As c increases, the modified offered load approaches the average value $a = 100$. It is important to understand the offered load, because it is a primary determinant of the required staffing, as we will see.

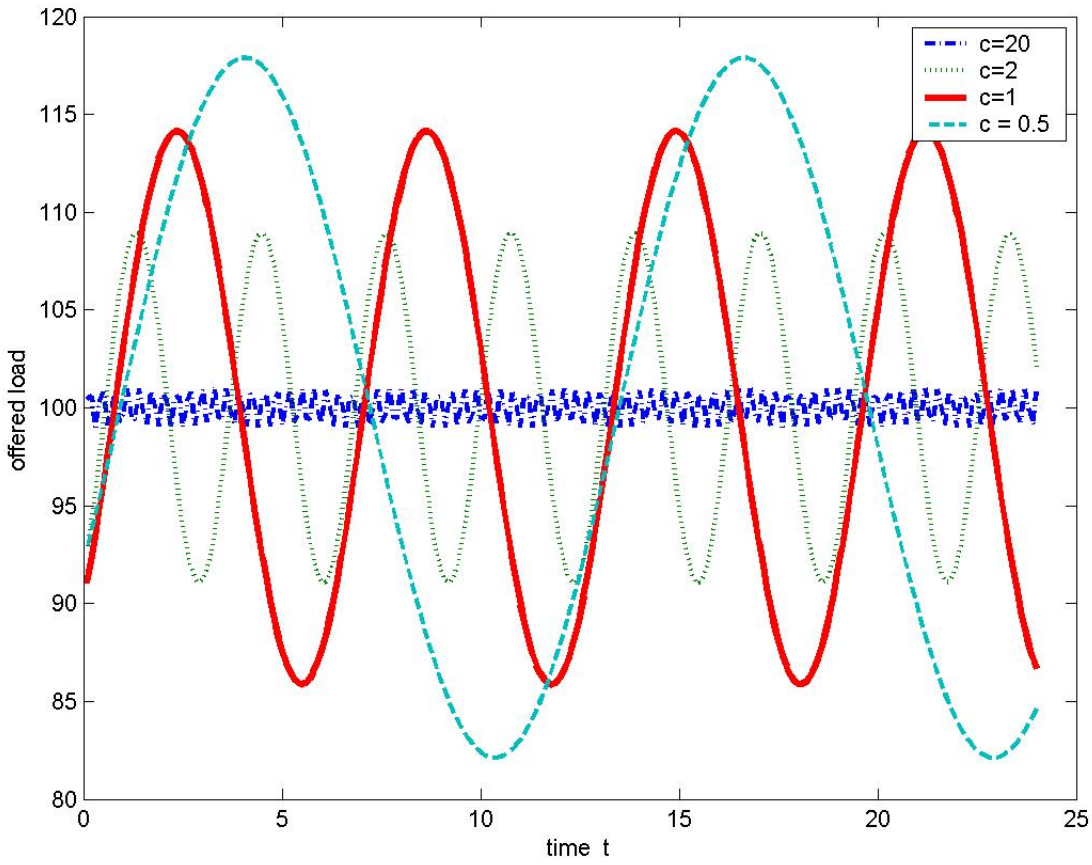
Our simulation-based iterated-staffing algorithm ISA generates staffing functions, for any given target delay probability α. In Figure 3 we present three graphs, showing the generated staffing functions for three regimes of operation: *Quality-Driven* (QD) - target $\alpha = 0.1$, *Efficiency-Driven* (ED) - target $\alpha = 0.9$, and *Quality-and-Efficiency-Driven* (QED) - target $\alpha = 0.5$. In each graph, we plot three curves: the arrival rate $\lambda(t)$ (dotted), the offered load m_t (dashed) and the staffing function s_t (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. That is consistent with Section 7 of Jennings et al. (1996).

Also note that, in the QED regime ($\alpha = 0.5$), the staffing function dictated by ISA falls right on top of the offered load: In that QED case, it would have sufficed to simply let $s_t = m_t$. Staffing to the offered load proved effective in all our experiments. That itself is quite stunning.

We now show that ISA achieves time-stable performance. In Figure 4 we show the actual probability of delay obtained by applying our algorithm 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. Under the staffing levels produced by our algorithm, the delay probabilities are remarkably accurate and stable; the observed delay probabilities fluctuate around the target in each case.

Figure 2: **The offered load m_t for the sinusoidal arrival-rate function** in (5.13) with parameters $a = 100$, $b = 20$ and four possible values of c: 0.5, 1, 2 and 20. The offered load is the mean number of busy servers in the $M_t/M/\infty$ model. The plotting is done at granularity 0.1, so the plot for $c = 20$ looks a bit strange.



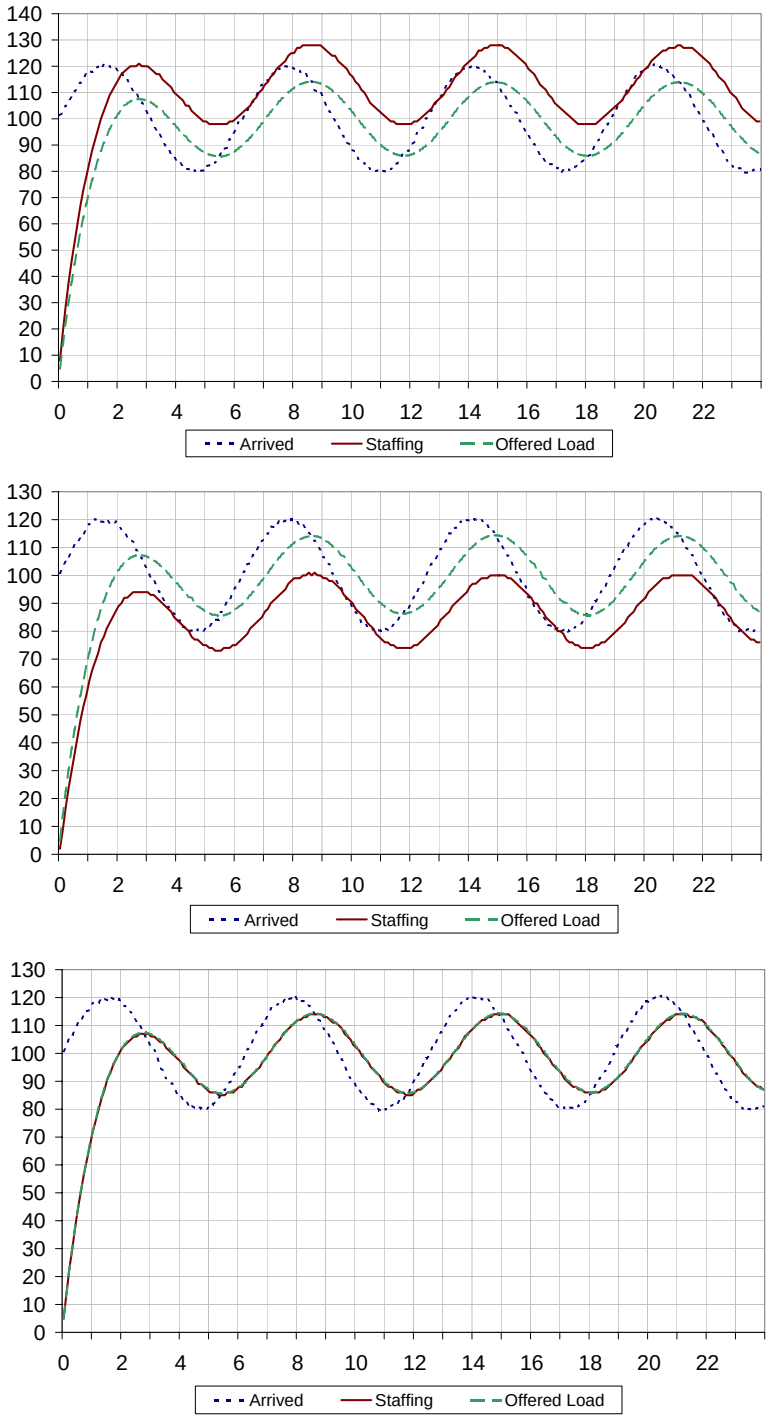
In addition to stabilizing the delay probability, other performance measures (e.g. utilization, tail probabilities, average waiting time and average queue length) are found to be quite stable as well; see the Internet Supplement. However, as the target delay probability increases toward heavy loading, the abandonment probability becomes much less time-stable, as shown in Figure 5. We discuss this phenomenon further in §7 below. But even the abandonment probability is quite stable with a lower delay-probability target (in the QD and QED regimes).

We now validate the square-root-staffing rule. For that purpose, we define an **implied empirical service quality**: A function $\{\beta_t : 0 \le t \le T\}$ is defined by setting

$$\beta_t \equiv \frac{s_t - m_t}{\sqrt{m_t}}, \quad 0 \le t \le T\,, \tag{5.15}$$

where m_t is again the offered load in (2.5) and (5.14). and s_t is the staffing function obtained by the ISA algorithm. Since s_t is obtained from the ISA algorithm, the function β_t is itself obtained from the ISA algorithm.

Figure 3: **Staffing function for: (1) Target $\alpha = 0.1$ (2) Target $\alpha = 0.9$ (3) Target $\alpha = 0.5$**



It thus becomes interesting to see if the implied service quality is approximately constant as a function of time. (That would empirically justify the square-root-staffing formula in (2.3).) And, indeed, it is, as shown in Figure 6. Again we consider 9 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 decreases. But the main point is that the empirical service quality β_t as a function of t is

Figure 4: **Delay probability summary for various α's.**

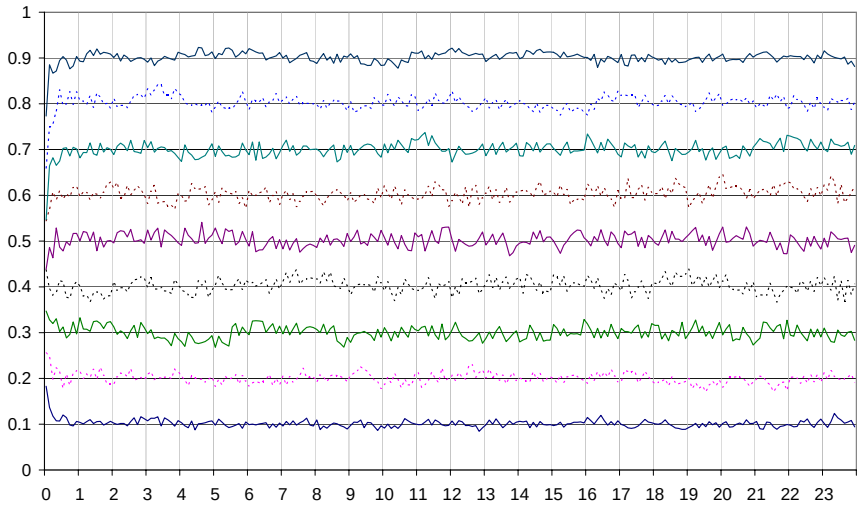
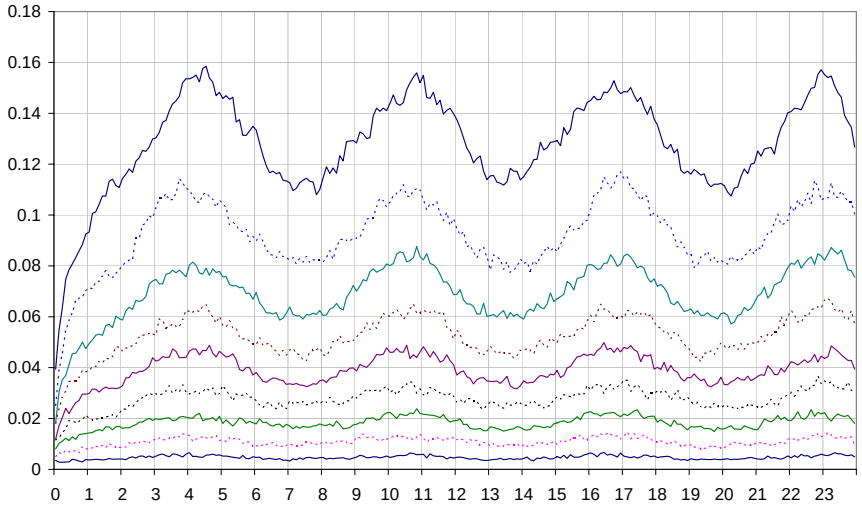


Figure 5: **Abandon probability summary for the Erlang-A example**



approximately constant as a function of t for each α over the full range from 0.1 to 0.9.

Figure 6 is extremely important because it validates the square-root-staffing formula for this example. First, Figure 4 shows that ISA is able to produce the target delay probability α for a wide range of α. Then Figure 6 shows that, 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 square-root-staffing formula instead of by the ISA. Moreover,

Figure 6: **Summary of Implied Service Quality β.** (The implied service quality decreases as α increases through the values $0.1, 0.2, \ldots, 0.9$.)

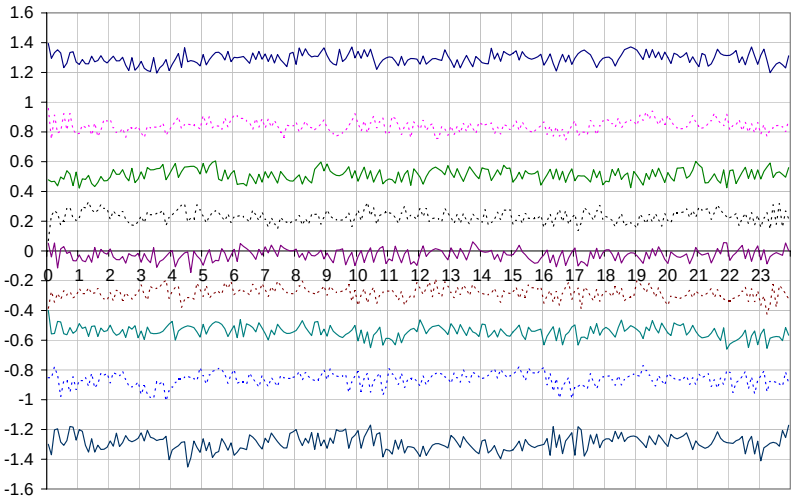


Figure 6 not only validates the square-root-staffing formula, but it also is the first step in validating the modified-offered-load approximation.

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: For the time-varying Erlang-A model, we use the associated stationary Erlang-A model, i.e., the $M/M/s+M$ model. As we observed before, that is tantamount to using the modified-offered-load approximation. Moreover, paralleling what Jennings et al. did for the Erlang-C model, we suggest using simple formulas obtained from the many-server heavy-traffic limit for the Erlang-A model in Garnett et al. (2002). The **Garnett-Mandelbaum-Reiman function, for brevity here referred to as the Garnett function**, mapping β into α is

$$\alpha = \left[1 + \sqrt{\frac{\theta}{\mu}} \cdot \frac{h(\hat{\beta})}{h(-\beta)}\right]^{-1}, \qquad -\infty < \beta < \infty; \tag{5.16}$$

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. Of course, we want a function mapping α into β. Thus, we use the **inverse of the Garnett function**, which is well defined.

We also looked at additional simulation output, aimed at establishing the validity of this stationary-model approach of relating α and β. First, we compared the empirical distribution of the customer waiting times to the theoretical distribution of those waiting times in the stationary Erlang-A model. Specifically, we plotted 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. 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 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). We found that the approximation for the conditional waiting-time distribution in the stationary queues matches the performance of our time-varying model remarkably well; see the Internet Supplement.

We next related the empirical (α, β) pairs to the Garnett function in (5.16). We define the empirical values $\bar{\alpha}$ and $\bar{\beta}$ as simply the time-averages of the observed (time-stable) values displayed in the plots in Figures 4 and 6. In Figure 7, we plot the pairs of $(\bar{\alpha}_i, \bar{\beta}_i)$ alongside the Garnett function. Needless to say, the agreement is phenomenal!

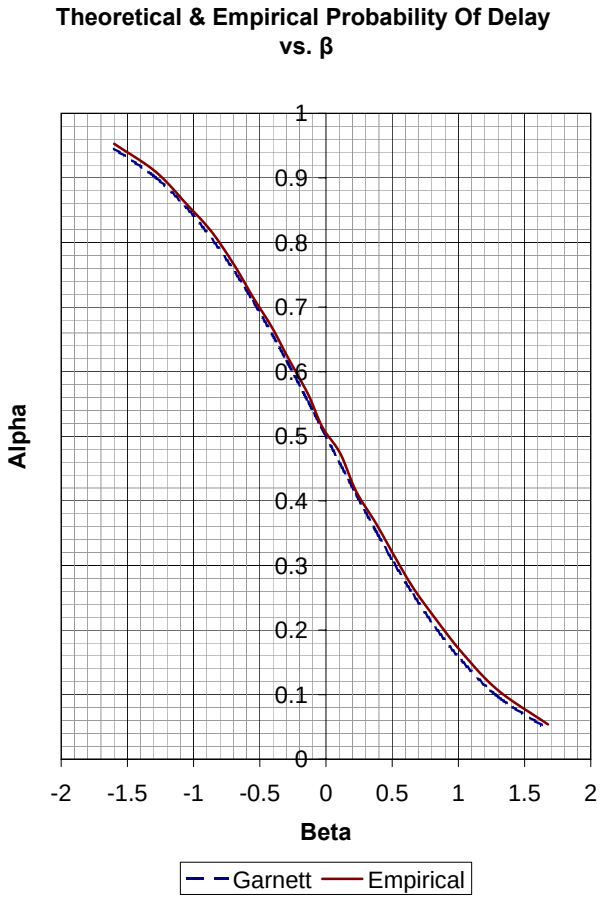
We close this section by observing that, just as in Jennings et al., other common approximations, such as the PSA or the SSA (the simple stationary approximation, using the overall time-average arrival rate) perform poorly for this example; see the Internet Supplement.

6 The Time-Varying Erlang-C Model

For comparison, we now show the performance of ISA for the same system described in §5 only without abandonment (with infinite patience) - the time-varying Erlang-C model ($M_t/M/s_t$). As expected, the required staffing levels are higher than with abandonment, for all target delay probabilities. For example, for $\alpha = 0.5$, the maximum staffing level becomes about 120 instead of 115.

As before, we achieve accurate time-stable delay probabilities when we apply the ISA; see Figure 8. The empirical service quality β_t is stabilizing as well, as can be seen from Figure 9. However, the empirical service

Figure 7: **Algorithm-Generated Performance vs. the Garnett Function**



quality β_t stabilizes at a much slower rate, especially for lower values of β (larger values of α). (The approach to steady-state is known to be slower in Erlang-C than for Erlang-A in heavy traffic.) Without abandonment the system is more congested, but still congestion measures remain relatively stable. That is just as we would expect, since the time-dependent Erlang-C model is precisely the system analyzed in Jennings et al. (1996); see the Internet Supplement for more details.

Just as for the time-varying Erlang-A model, we want to validate the square-root-staffing formula in (2.3). We thus repeat the various experiments we did in §5. Recall that, for the *stationary* $M/M/s$ queue, the conditional waiting-time $(W \mid W > 0)$ is (exactly) exponentially distributed. The empirical conditional waiting-time distribution given wait, in our *time-varying* queue and over *all* customers, also fits the exponential distribution very well (see the Internet Supplement). The mean of the plotted exponential distribution was taken to be the

Figure 8: **Delay probability summary for the Erlang-C example**

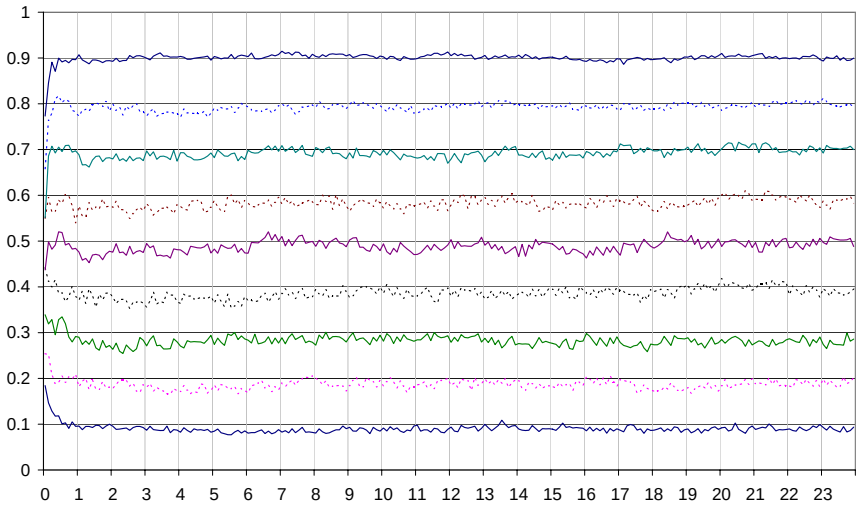
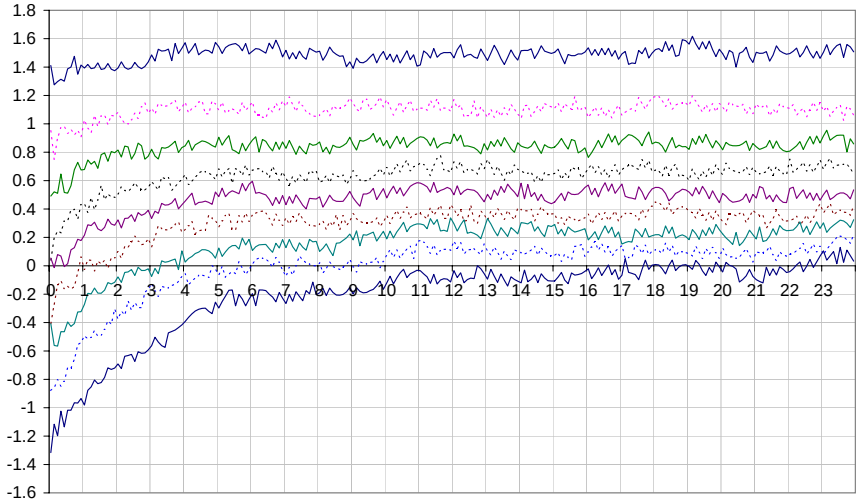


Figure 9: **Implied service quality β summary for the Erlang-C example** (The implied service quality decreases as α increases through the values 0.1, 0.2, $\ldots$, 0.9.)



overall average waiting time of those who were actually delayed during $[0, T]$.

Here, the relation between α and β is compared with the **Halfin-Whitt function** from Halfin and Whitt (1981),

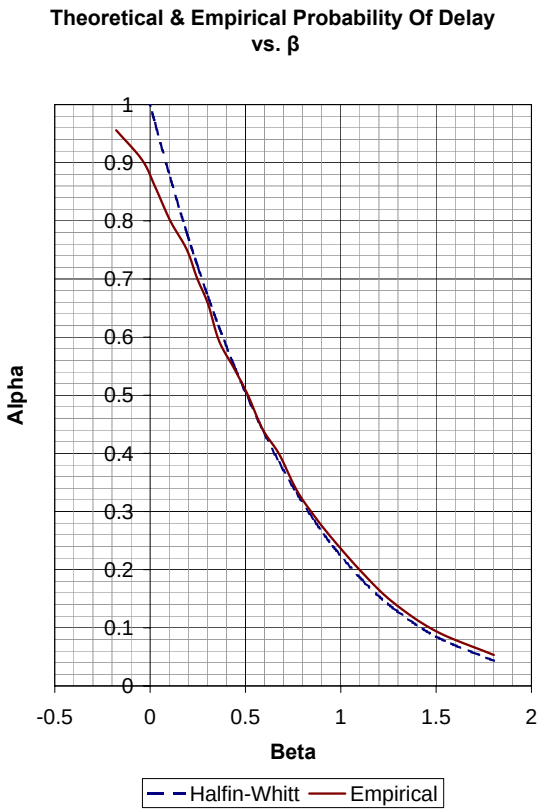
namely,

$$P(delay) \equiv \alpha \equiv \alpha(\beta) \approx \left[1 + \beta \cdot \frac{\Phi(\beta)}{\phi(\beta)}\right]^{-1}, \quad 0 < \beta < \infty\,, \tag{6.17}$$

where ϕ is again the pdf associated with the standard normal cdf Φ. The Halfin-Whitt function in (6.17) is obtained from the Garnett function in (5.16) by letting $\theta \to 0$.

Just as we use the Garnett function to relate the target delay probability α to the quality of service β in the square-root-staffing formula in (2.3) for the $M_t/M/s_t+M$ model, so we use the Halfin-Whitt function to relate α to β in the square-root-staffing formula in (2.3) for the $M_t/M/s_t$ model. And that essentially corresponds to the refinement performed in Section 4 of Jennings et al. (1996). The results in Figure 10 are again remarkable.

Figure 10: **Comparison of empirical results with the Halfin-Whitt approximation**



7 Theoretical Support in the Case $\theta = \mu$

In one special case, we can analyze the time-dependent Erlang-A model (i.e., the $M_t/M/s_t + M$ model) in considerable detail. That is the case we considered in Section 5, in which the individual service rate μ equals the individual abandonment rate θ. In this section, let θ and μ be fixed with $\theta = \mu$, but here we do not set these equal to 1.

With that condition, it is easy to relate the $M_t/M/s_t + M$ model to the corresponding time-dependent infinite-server model (the $M_t/M/\infty$ model with the same arrival-rate function and service rate) and a corresponding family of stationary Erlang-A models indexed by t (the $M/M/s + M$ model with the same service and abandonment rates, but with special arrival rate and number of servers). We can thus do some theoretical analysis for the model considered in Section 5.

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 elementary (important) 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 $\{L_t : t \geq 0\}$ in the $M_t/M/s_t + M$ model with $\theta = \mu$ has the same distribution (finite-dimensional distributions) as the corresponding process $\{L_t^\infty : t \geq 0\}$ in the $M_t/M/\infty$ model with the same arrival-rate function $\lambda(t)$ and the same individual service rate μ, i.e.,

$$\{L_t : t \geq 0\} \stackrel{\mathrm{d}}{=} \{L_t^\infty : t \geq 0\} \; . \tag{7.18}$$

If we appropriately define the two models on the same sample space, giving both processes the same arrivals, we can make the two equal with probability 1 as well.

The second elementary (important) observation is that, for both these models, the individual random variables L_t and L_t^∞ have the same distribution as the steady-state number in system L_∞ in the corresponding stationary model with appropriate arrival rate and number of servers (which are appropriate functions of t).

Letting the service-time random variable S have an exponential distribution with mean $1/\mu$, for each t, we have

$$L_t \stackrel{\mathrm{d}}{=} L_\infty^\infty \stackrel{\mathrm{d}}{=} L_\infty \; . \tag{7.19}$$

where the second random variable in (7.19), L_∞^∞ is the steady-state number of busy servers in the stationary $M/M/\infty$ with arrival rate $\hat{\lambda}_t$ in (2.9), with m_t again the expected number in system in the time-dependent

infinite-server model in (2.5), and the third random variable in (7.19), L_∞, is the steady-state number in system in the $M/M/s+M$ model with the constant number of servers equal to s_t and the arrival rate again being $\hat{\lambda}_t$ in (2.9).

7.1 The Delay Probability

Let W_t be the **virtual waiting time** at time t (until service or abandonment, whichever occurs first, i.e., the waiting time in queue that would be spent by an arrival at time t); 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) in the $M_t/M/s_t+M$ model. These quantities are considerably more complicated.

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

$$\begin{aligned} P(W_t > 0) &= P(L_t \geq s_t) = P(L_t^\infty \geq s_t) = P(Poisson(m_t) \geq s_t) \\ &\approx P\left(N(0,1) > \frac{s_t - m_t}{\sqrt{m_t}}\right) , \end{aligned} \tag{7.20}$$

where m_t is the offered load in (2.5), just as in (2.2), only here the infinite-server approximation is exact.

7.2 Approximations for the Waiting-Time Distribution

However, the virtual abandonment probability P_t^{ab} and the expected virtual waiting time $E[W_t]$ fluctuate much more than the delay probability; e.g., see Figure 5. We will explain that greater fluctuation.

We actually can mathematically analyze the time-dependent virtual waiting time W_t and the time-dependent virtual abandonment probability P_t^{ab}. Here is an important initial observation: Conditional on the event that $W_t > 0$, whose probability we have analyzed above, W_t is distributed (exactly) as the first passage time of the (Markovian) stochastic process $\{L_u : u \geq t\}$ from the initial value L_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 μL_u for $u \geq t$ down from the initial value L_t to the curve $\{s_u : u \geq t\}$. (Of course, $W_t = 0$ if $L_t < s_t$.) As a

consequence, the distribution of W_t and the value of P_t^{ab} depend on only L_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.

We can go further if we make approximations. Even though exact relations are difficult to obtain, it is not difficult to generate very good approximations for the case in which the number of servers tends to be large, e.g., as in the specific example in the previous subsection. Then, W_t tends to be very small, so that it is often reasonable to assume that the staffing level remains constant at s_t in the time shortly after t. In other words, to study W_t and P_t^{ab}, we 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. We are thinking of applications in which the time-dependent arrival-rate function is continuous, and the staffing changes relatively slowly.

If the future-staffing-level approximation held as an equality, then we would obtain the following approximations as equalities:

$$W_t \approx W_\infty \quad \text{and} \quad P_t^{ab} \approx P_\infty^{ab} \,, \tag{7.21}$$

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 $\hat{\lambda}_t$ in (2.9). Hence, we propose (7.21) as approximations.

Given approximations (7.21), we can use established results for the stationary $M/M/s + M$ model, e.g., as in Garnett et al. (2002) and Whitt (2005). For example, 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.

8 Algorithm Dynamics

In this section we discuss the dynamics of the iterative-staffing algorithm for the $M_t/M/s_t + M$ model. We first relate an empirical observation about the way the algorithm converges to the limiting staffing function $s^{(\infty)}$ and then afterwards we give a theoretical explanation.

In particular, we observed that the way the staffing functions converge to the limit depends on the ratio $\mathbf{r} \equiv \theta/\mu$. Whenever the (im)patience rate θ is less than the service rate μ ($\mathbf{r} < 1$), we encounter **oscillating dynamics**

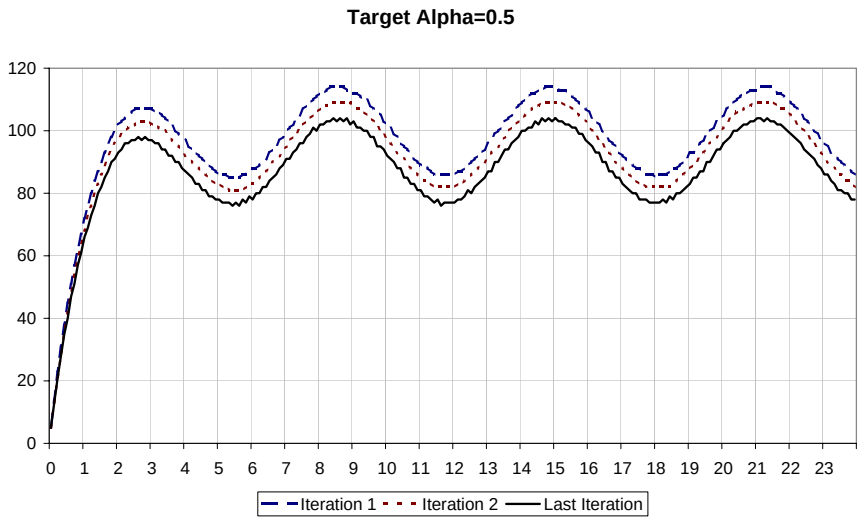
of the staffing level during the algorithm; whenever the (im)patience rate θ is greater than the service rate μ ($\mathbf{r} > 1$), we encounter **monotone dynamics** of the staffing level during the algorithm.

With *monotone dynamics*, when starting with $s_t^{(0)} \equiv \infty$, $s_t^{(n)}$ is monotone decreasing in n for all t, i.e. the following prevails:

$$s_t^{(n)} \leq s_t^{(m)} \qquad \text{for all} \quad m < n\,. \tag{8.22}$$

An example of the monotone dynamics is shown in Figure 11, where staffing levels are shown for the first three iterations of the algorithm for the case of arrival function $\lambda(t) = 100 + 20 \cdot \sin(t)$, service times exponential having mean 1, and impatience times that are exponential having mean 0.1 ($\mathbf{r} = 10$).

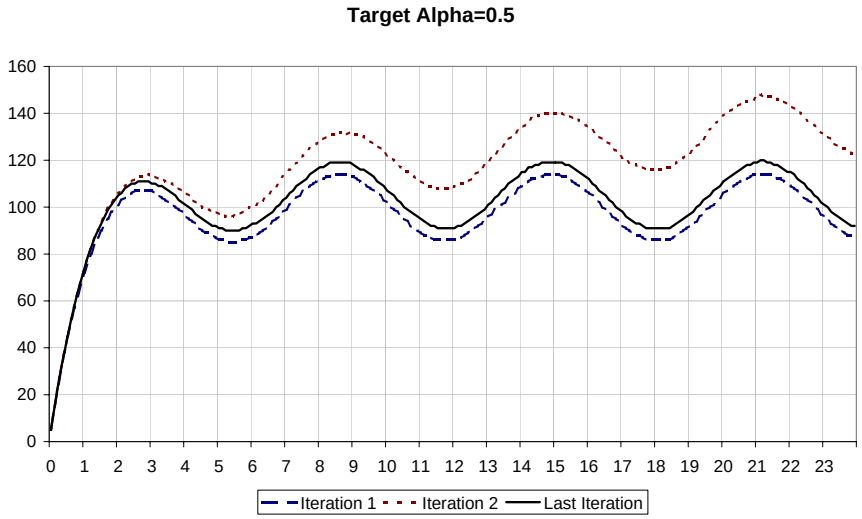
Figure 11: **Staffing levels in the 1^{st}, 2^{nd} and last iterations. μ=1, θ=10.**



In contrast, with *oscillating dynamics*, $s_t^{(n)}$ is oscillating for all t; i.e. there exist 2 subsequences $\{s_t^{(k)}\}_{k=2n}^{\infty}$ and $\{s_t^{(l)}\}_{l=2n+1}^{\infty}$, such that $s_t^{(2n)} \downarrow s_t^{(\infty)}$ and $s_t^{(2n+1)} \uparrow s_t^{(\infty)}$. Within the oscillating framework, there is monotonicity. An example of the oscillating dynamics can be viewed in Figure 12, where staffing levels are shown for the first three iterations for the same case except there is no abandonment ($\theta = 0$ and $\mathbf{r} = 0$).

For the $M_t/M/s_t + M$ model, the algorithm dynamics can be explained by stochastic-order relations for the time-varying birth-and-death process $\{L_t : t \geq 0\}$. For all systems, the arrival process is the same. However, the death rates depend systematically on the number of servers s_t. When $\mathbf{r} > 1$ ($\mathbf{r} < 1$), the death rates at time t decrease (increase) as s_t increases. Hence, if we disregard statistical error, caused by having to estimate the

Figure 12: **Staffing levels in the 1^{st}, 2^{nd} and last iterations. μ=1, θ=0**



delay probabilities associated with each staffing function, we can actually prove that the algorithm converges for the $M_t/M/s_t + M$ model. To do so, we use sample-path stochastic order, as in Whitt (1981). 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-order property** for the $M_t/M/s_t + M$ model: If $s_t^{(1)} \leq s_t^{(2)}$ for all t, $0 \leq t \leq T$, and $\mathbf{r} > 1$, then

$$\{L_t^{(1)} : 0 \leq t \leq T\} \leq_{st} \{L_t^{(2)} : 0 \leq t \leq T\} \ , \tag{8.23}$$

where $\leq_{st}$ denotes **sample-path stochastic order**, i.e.,

$$E\left[f\left(\{L_t^{(1)} : 0 \leq t \leq T\}\right)\right] \leq_{st} E\left[f\left(\{L_t^{(2)} : 0 \leq t \leq T\}\right)\right] \tag{8.24}$$

for all nondecreasing real-valued functions f on the space of sample paths. The ordering is reversed if instead $\mathbf{r} < 1$.

The ordering of the death rates in the two birth-and-death processes makes it possible to achieve the sample-path ordering. Indeed, that can be accomplished (the relation (8.23) can be rigorously justified) 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, Krengel and O'Brien (1978), that special construction is actually equivalent to the sample-path stochastic ordering in (8.23).

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. As indicated above, the two processes are 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: $L_t^{(1)} = L_t^{(2)} = k$. 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 construction is just like the thinning of a Poisson process used in the simulation of a nonhomogeneous Poisson process.

As a consequence of the sample-path stochastic order, we get ordinary stochastic order

$$L_t^{(1)} \leq_{st} L_t^{(2)} \quad \text{for all} \quad t\,, \tag{8.25}$$

where now $\leq_{st}$ denotes conventional stochastic order for real-valued random variables, just as in Chapter 9 of Ross (1996); also see Müller and Stoyan (2002). We only need the more elementary stochastic order in (8.25), but we use the more sophisticated sample-path stochastic order in (8.23) to get it. The stochastic order is equivalent to the tail probabilities being ordered; i.e., (8.25) is equivalent to $P(L_t^{(1)} > x) \leq P(L_t^{(2)} > x)$ for all x, which implies the ordering for the staffing functions at time t. In particular, suppose that

$$P\left(L_t^{(2)} \geq s_t^{(2)}\right) \leq \alpha < P\left(L_t^{(2)} \geq s_t^{(2)} - 1\right)\,. \tag{8.26}$$

Since

$$P\left(L_t^{(1)} \geq s_t^{(2)}\right) \leq P\left(L_t^{(2)} \geq s_t^{(2)}\right) \leq \alpha\,, \tag{8.27}$$

necessarily $s_t^{(1)} \leq s_t^{(2)}$.

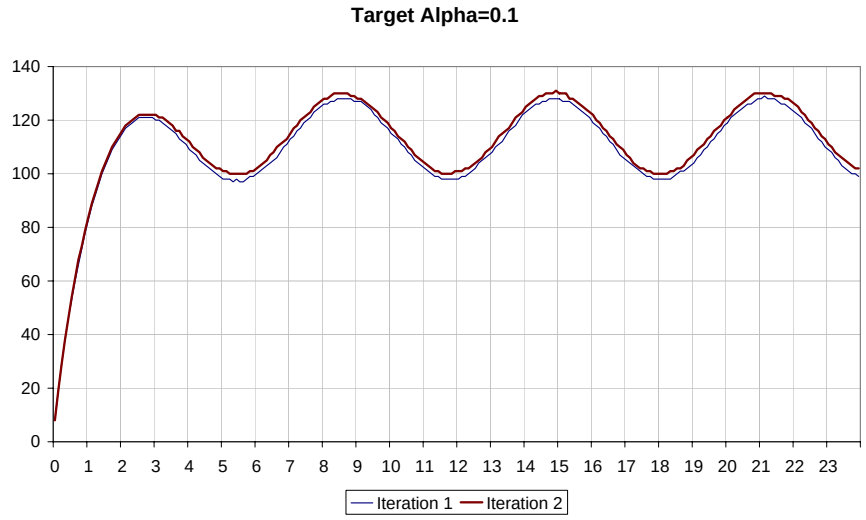
Case 1: r > 1. For $s_t^{(0)} = \infty$, we necessarily start with $s_t^{(0)} > s_t^{(1)}$ for all t, which produces first $L_t^{(1)} \leq_{st} L_t^{(0)}$ and then $s_t^{(2)} \leq s_t^{(1)}$ for all t. Continuing, we get $L_t^{(n)}$ stochastically decreasing in n and $s_t^{(n)}$ decreasing in n, again 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 $L_t^{(1)} \geq_{st} L_t^{(0)}$ and then $s_t^{(0)} \geq s_t^{(2)} \geq s_t^{(1)}$ for all t. Afterwards, we get $L_t^{(1)} \geq_{st} L_t^{(2)} \geq_{st} L_t^{(0)}$ and $s_t^{(0)} \geq$

$s_t^{(2)} \geq s_t^{(3)} \geq s_t^{(1)}$ for all t. Continuing, we get $L_t^{(2n)}$ stochastically increasing in n, while $L_t^{(2n+1)}$ 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 oscillating 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 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. This phenomenon is illustrated in Figures 13 and 14. The staffing levels in the first two iterations, which form the range of the oscillating dynamics, are plotted for both target $\alpha = 0.1$ (Figure 13) and $\alpha = 0.5$ (Figure 14) for the case of arrival function $\lambda(t) = 100 + 20 \cdot \sin(t)$, service times are exponential having mean 1, and no abandonment.

Figure 13: **Range of staffing level for target α=0.1**

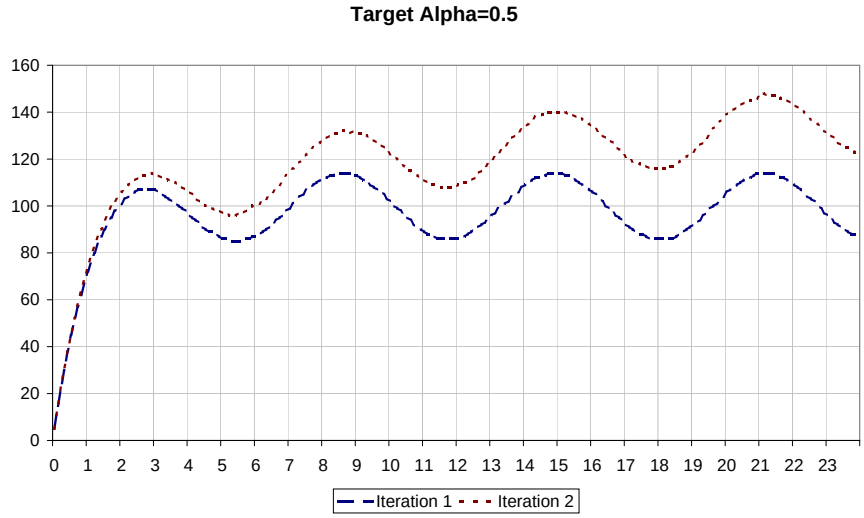


Finally, we also observed a **time-dependent behavior in the convergence** of $s_t^{(n)}$. We observed a greater gap as time increased. For example, let

$$I_t \equiv \inf \{j : s_t^{(i)} = s_t^{(j)} \quad \text{for all} \quad i \geq j\} . \tag{8.28}$$

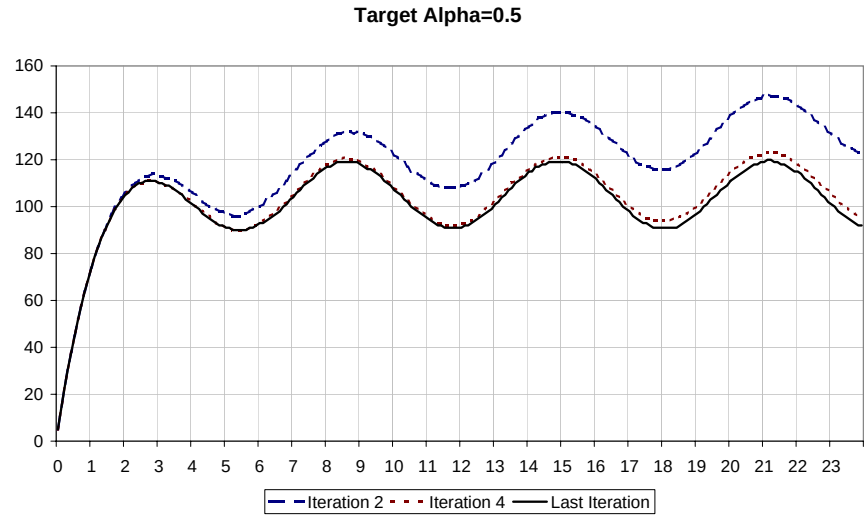
We observed that $I_{t_2} \geq I_{t_1}$ for all $t_2 > t_1$. An illustration can be viewed in Figure 15. This time-dependent behavior is understandable, because the gap between two different staffing levels persists across time, so that there is a gap in the death rates at each t. Hence, as t gets larger, the two processes can get further apart. Thus

Figure 14: **Range of staffing level for target α=0.5**



the gap can first decrease more at the left end of the time horizon. When it reaches the limit at the left, the gap will still decrease more to the right.

Figure 15: **Evolution of convergence during algorithm run-time**



References

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

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

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

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

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

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

[7] Green, L. V., Kolesar, P. J., Whitt, W. **Coping with time-varying demand when setting staffing requirements for a service system**. Columbia University. Available at: http://www.columbia.edu/∼ww2040/Coping.pdf

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

[9] Jagerman, D. L. **Nonstationary blocking in telephone traffic.** *Bell System Technical Journal*, **54**, 625–661, 1975.

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

[11] Kamae, T., Krengel, U., O'Brien, G. L. **Stochastic inequalities on partially ordered spaces.** *Annals of Probability* **5**, 899–912, 1978.

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

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

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

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

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

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

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

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

[20] Whitt, W. **The Impact of a Heavy-Tailed Service-Time Distribution upon the M/GI/s Waiting-Time Distribution**. *Queueing Systems*, **36**, 71–87, 2000.

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