# 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



November 24, 2004



**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. It turns out that it suffices to target a stable probability of delay. That procedure tends to produce time-stable performance for several other operational measures. 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: First, we develop a simulation-based iterative-staffing algorithm (ISA), and conduct experiments showing that it is effective. The ISA is appealing because it applies to very general models and is automatically validating: we directly see how well it works. Second, we extend the square-root-staffing rule, proposed by Jennings et al., which is based on the associated infinite-server model. The rule dictates that the staff level at time $t$ be $s_t = m_t + \beta\sqrt{m_t}$, where $m_t$ is the offered load (mean number of busy servers in the infinite-server model) and the constant $\beta$ reflects the service grade. We show that the service grade $\beta$ in the staffing formula can be represented as a function of the target delay probability $\alpha$ by using approximations for the steady-state delay probability in the associated stationary $M/M/s + M$ model, drawing on Garnett, Mandelbaum and Reiman (2002). Finally, for many-server queues with abandonment, we show that simply staffing at the offered load itself (i.e., letting $s_t = m_t$) is remarkably effective in typical operating regimes. Indeed, for practical examples with relatively short service times, it suffices to let $s_t = \lambda(t)/\mu$, where $\lambda(t)$ is the arrival rate and $1/\mu$ is the mean service time, as in a naive deterministic method.



**Keywords:** Contact centers; call centers; staffing; operator staffing; queues; non-stationary queues; queues with time-dependent arrival rates; multi-server queues; infinite-server queues; capacity planning; queues with abandonment.

# Contents


**1 Introduction** **1**

1.1 Background on Services and Call Centers . . . . . 1

1.2 The Staffing Problem . . . . . 1

1.3 Our Point of Departure . . . . . 2

1.4 Our Contributions to the Staffing Problem . . . . . 4

1.4.1 A Simulation-Based Iterative Staffing Algorithm. . . . . 4

1.4.2 An Extended Version of the Square-Root Staffing Formula. . . . . 4

1.4.3 Simple Deterministic Approximations . . . . . 5

1.5 Summary of The Paper . . . . . 6

**2 Examples and Motivation** **6**

2.1 The Time-Varying Erlang-A Model . . . . . 6

2.1.1 A Sinusoidal Arrival-Rate Function . . . . . 6

2.1.2 Time-Stable Performance . . . . . 8

2.1.3 Validating the Square-Root-Staffing Formula . . . . . 9

2.1.4 Relating $\beta$ to $\alpha$ . . . . . 11

2.2 Theoretical Motivation: The Case $\theta = \mu$ . . . . . 12

2.2.1 Connections to Other Models . . . . . 13

2.2.2 The Delay Probability . . . . . 14

2.2.3 Approximations for the Waiting-Time Distribution . . . . . 14

2.2.4 Asymptotic Time-Stability in the Many-Server Heavy-Traffic Limit . . . . . 15

2.3 The Time-Varying Erlang-C Model . . . . . 16

2.3.1 Time-Stable Performance . . . . . 16

2.3.2 Validating the Square-Root-Staffing Formula . . . . . 17

2.4 Benefits of Taking Account of Abandonment . . . . . 19

**3 The Simulation-Based Iterative-Staffing Algorithm** **20**

3.1 The ISA . . . . . 20

3.1.1 The Steps of ISA . . . . . 21

3.1.2 Implementation and Convergence . . . . . 21

3.2 Performance Measures . . . . . 22

**4 Additional Examples** **22**

4.1 The Challenging Example . . . . . 22

4.2 The $M_t/M/s_t + M$ Model with More and Less Patient Customers . . . . . 23

4.3 Benefits of Taking Account of Abandonment Again . . . . . 24

4.4 A Practical Example . . . . . 25

4.5 Non-Exponential Service Times . . . . . 28

**5 Algorithm Dynamics** **29**

**6 Summary and Directions for Future Research** **33**

**7 Appendix** **A-1**

7.1 A Uniform-Acceleration Perspective . . . . . A-1

7.2 Case 1: $\theta_t = \mu_t$ . . . . . A-3

7.3 Case 2: $\theta_t = 0$ . . . . . A-3

# 1 Introduction

## 1.1 Background on Services and Call Centers

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. Over-staffing is wasteful, while under-staffing leads to low service-levels and overworked agents.

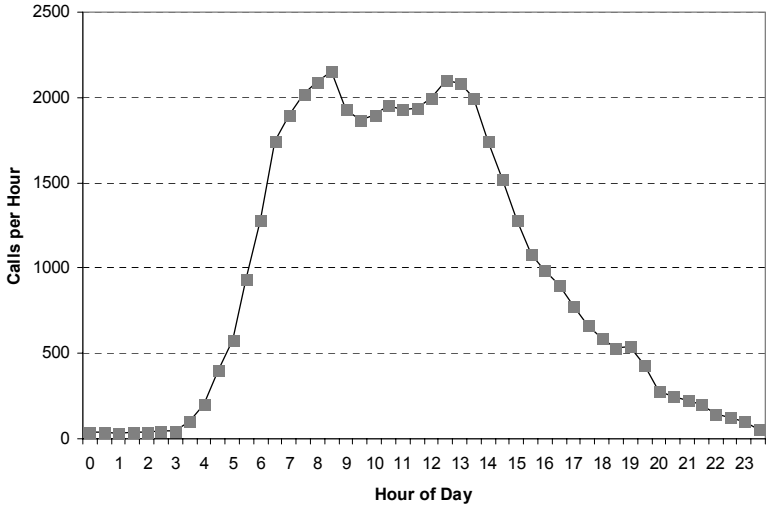
## 1.2 The Staffing Problem

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 Chapter 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) in order to benchmark our algorithm; see Section 4.)

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, 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) and Whitt (1991). The PSA is

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




the leading term in the *uniform-acceleration* (UA) approximation; see Massey and Whitt (1998) and references therein.

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 over-staffed, while others are under-staffed.

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.

## 1.3 Our Point of Departure

Our point of departure is our (with Otis B. Jennings) previous paper: Jennings, Mandelbaum, Massey and Whitt (1996). In that paper, we showed For the $M_t/M/s_t$ model that it is possible to achieve time-stable performance. That observation strongly motivates the present study.

In that paper, we proposed an **infinite-server approximation** for many-server queues with time-varying arrival rate, without customer abandonment, in particular for the $M_t/G/s$ model, having a nonhomogeneous Poisson arrival process with arrival-rate function $\lambda(t)$ and *independent and identically distributed* (IID) service times $\{S_n : n \geq 1\}$, distributed as a random variable $S$ with a general *cumulative distribution function* (cdf) $G$ having mean $E[S] = 1/\mu$. For the $M_t/G/s$ model, we suggested staffing according to the **square-root-staffing formula**:

$$s_t = m_t + \beta\sqrt{m_t}, \quad 0 \leq t \leq T, \tag{1.1}$$

where the constant $\beta$ is a measure of the **grade of service** and the deterministic function $m_t$ is the **offered load**, i.e., the mean number of busy servers in the associated $M_t/G/\infty$ infinite-server model (with same arrival process and service times).

The underlying motivation for this square root formula comes from the fact that the number of customers in the $M_t/G/\infty$ infinite-server model has a *Poisson distribution* for all times $0 < t \leq T$ whenever the number in the system at time $t = 0$ has a similar distribution (being empty is a degenerate case of a Poisson distribution). The mean and the variance are equal for a Poisson distribution. Therefore, the fact that $m_t$ equals the *mean* for the "offered load process" (infinite server model) at time $t$ implies that $\sqrt{m_t}$ equals the *standard deviation* for this offered load process. Thus we are simply setting the number of servers $s_t$ equal to the mean plus some number $\beta$ of standard deviations of the offered load.

The important insight above is that the **"right" offered load** above should be the time-dependent mean number of busy servers in the associated infinite-server model. 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 infinite-server model is a better generalization of "offered load" for most time-varying many-server models. (Indeed, it may be considered exactly the right definition for the infinite-server model itself.)

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**. Let $L_t(\infty)$ be the number of busy servers at time $t$ in the infinite-server model. Then the explicit formula for $m_t$ is

$$m_t \equiv E\left[L_t(\infty)\right] = \int_{-\infty}^{t} G^c(t-u)\lambda(u)\,du = E\left[\int_{t-S}^{t} \lambda(u)\,du\right] = E\left[\lambda(t-S_e)\right]E[S]\,, \tag{1.2}$$

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

$$G_e(t) \equiv P(S_e \leq t) \equiv \frac{1}{E[S]}\int_0^t [1-G(u)]\,du, \quad t \geq 0\,; \tag{1.3}$$

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.

¿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

$$m_t - \lambda(t)\cdot E[S] = E\left[\lambda'\left(t-(S_e)_e\right)\right]\cdot E[S_e]\cdot E[S] = \frac{1}{2}\cdot E\left[\lambda'\left(t-(S_e)_e\right)\right]\cdot E[S^2]. \tag{1.4}$$

¿From (1.4), 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 $\lambda'$). 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.

Given that we use the square-root-staffing formula in (1.1) and that we can directly calculate the offered load by (1.2), the remaining problem is to determine an appropriate grade of service $\beta$ in (1.1). Toward that end, we chose to make the **delay probability** – the probability an arrival will have to wait before beginning service – the **target performance measure**. Our goal was to have the delay probability at every time $t$ be a target $\alpha$. That choice was by no means arbitrary. As proved in Halfin and Whitt (1981) and further discussed in Whitt (1992) and Jennings et al. (1996), the delay probability is an ideal performance measure because it has a nondegenerate many-server heavy-traffic limit. That means that the delay probability tends to have a meaningful interpretation, independent of scale (the load and the number of servers). We discuss how to relate the grade of service $\beta$ to the target delay probability $\alpha$ later.

## 1.4 Our Contributions to the Staffing Problem

Our goal in this paper is to develop staffing algorithms for more complicated time-varying many-server models, such as many-server queues with abandonment. For example, we want to 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). We do not consider SBR models here, but the methods here may extend to that setting, especially when combined with the methods of Wallace and Whitt (2004), which use appropriate cross training to reduce SBR staffing to single-group staffing.

### 1.4.1 A Simulation-Based Iterative Staffing Algorithm.

Our first contribution 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 by having to have 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.)

Continuing to follow Jennings et al. (1996), we use **the delay probability** as our target performance measure, but the same method could be applied to other performance measures. 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, queue-length tail delay-probabilities and the probability of abandonment, 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 error. That statistical 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).

### 1.4.2 An Extended Version of the Square-Root Staffing Formula.

While working with ISA, we discovered that the simulation-based solutions we were finding had 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 staffing could be related to the steady-state behavior of the associated stationary $M/M/s + M$ model.

That leads us to our second contribution: We extend the square-root staffing formula. In particular, we suggest staffing according to the **square-root-staffing formula** in (1.1), where the service grade $\beta \equiv \beta(\alpha)$ is derived from a theoretical **one-to-one relation between $\alpha$ and $\beta$ for the corresponding stationary model**. In particular, we propose using $\beta(\alpha)$, for which staffing levels of $s = m + \beta\sqrt{m}$ would lead to the desired delay probability $\alpha$ in the corresponding stationary model, where $m = \lambda/\mu$ is the stationary offered load. For the $M_t/M/s_t + M$ model, we use explicit formulas relating $\alpha$ to $\beta$ obtained from the many-server heavy-traffic limits 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.

### 1.4.3 Simple Deterministic Approximations

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, Mandelbaum and Reiman (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.

Theoretical justification for this heuristic can be found in Mandelbaum, Massey and Reiman (1998). In that setting, we can apply the many-server heavy-traffic scaling to the $M_t/M/s_t + M$ model and obtain the following result for a family of number-in-system stochastic processes $\{ L^\eta \mid \eta > 0 \}$ indexed by $\eta$, associated with $M_t/M/s_t + M$ queues. If $\theta^\eta = \theta$ and $\mu^\eta = \mu$, while

$$\lambda_t^\eta = \eta \cdot \lambda_t \quad \text{and} \quad s_t^\eta = \eta \cdot m_t + \sqrt{\eta} \cdot s_t^{(1)} + o(\sqrt{\eta}). \tag{1.5}$$

where $m_t$ is the mean for the $M_t/M/\infty$ queue and so

$$\frac{d}{dt} m_t = \lambda_t - \mu_t \cdot m_t, \tag{1.6}$$

we then have

$$\lim_{\eta \to \infty} \mathsf{P}\left(L^\eta(t) \geq s_t^\eta\right) = \mathsf{P}\left(L^{(1)}(t) \geq s_t^{(1)}\right), \tag{1.7}$$

where $L^{(1)} = \left\{ L^{(1)}(t) \mid t \geq 0 \right\}$ is a one-dimensional diffusion.

Here is the implication: It says that, asymptotically, controlling the delay for this queue with abandonment is a *second order* staffing effort (selecting $s_t^{(1)}$) whereas the leading order staffing level is satisfied by using the offered load. Moreover, for the special case of the abandonment rate equaling the service rate, we can apply this argument to rigorously obtain the square-root staffing formula used in Jennings et al. (1996) for the multiserver queue without abandonment. This is also the one case where the diffusion $L^{(1)}$ is Gaussian. In the Appendix of this paper we show how these results are derived.

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 coincide with 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 (1.2).

### 1.5 Summary of The Paper

In §2 we present examples illustrating the performance of our algorithm, and provide a theoretical motivation for the derived results. In §3 we describe our algorithm and give definitions of the performance measures that we display. Then, in §4, we present additional examples: We start by revisiting the "challenging example" in Jennings et al. (1996); we follow by expanding the analysis of the Erlang-A example from §2.1 with different patience parameters, emphasizing the stationary (time-stable) performance of our staffing algorithm. Then, we analyze a realistic example (the one presented in Figure 1). In contrast to Green et. al. (2001), we also incorporate abandonment, which significantly and positively impacts staffing results. In §5, the dynamics of the iterative algorithm is discussed. In §6 we discuss directions for future research. We provide additional theoretical perspective for the square-root-staffing algorithm from a uniform-acceleration perspective in a final appendix.

## 2 Examples and Motivation

We start with two examples demonstrating the performance of our algorithm: first, the time-varying Erlang-A model (with abandonments) and, second, the corresponding time-varying Erlang-C model (without abandonments).

### 2.1 The Time-Varying Erlang-A Model

#### 2.1.1 A Sinusoidal Arrival-Rate Function

Consider a queueing system that is faced with a non-homogeneous Poisson arrival process with a **sinusoidal arrival-rate function**

$$\lambda(t) = a + b \cdot \sin(ct), \quad 0 \le t \le T\,, \tag{2.8}$$

where $a = 100$, $b = 20$ and $c = 1$; i.e.,

$$\lambda(t) = 100 + 20 \cdot \sin(t), \quad 0 \le t \le T\,. \tag{2.9}$$

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 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 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 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 (2.8) 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 will consider a practical example in Section 4.4.) From this preliminary analysis, we 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).

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 (1.2). 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 (2.8), the offered load is

$$m_t = a + \frac{b}{1+c^2}[\sin(ct) - c \cdot \cos(ct)] = 100 + 10[\sin(t) - \cos(t)] \; . \tag{2.10}$$

The second formula in (2.10) is based on the specific parameters: $a = 100$, $b = 20$ and $c = 1$ from (2.9).

In order to put our model into perspective, in Figure 2 we plot the offered load $m_t$ in (2.10) for the sinusoidal arrival-rate function in (2.8) for the parameters $a = 100$ and $b = 20$, as in our example, but with four 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 $\alpha$. In Figure 3 we present three graphs, showing the generated staffing functions for three regimes of operation: *Quality-Driven* (QD) - target $\alpha = 0.1$, *Quality-and-Efficiency-Driven* (QED) - target $\alpha = 0.5$ and *Efficiency-Driven* (ED) - target $\alpha = 0.9$. In each graph, we plot three curves: the arrival rate $\lambda(t)$ (blue), the offered load $m_t$ (green) and the staffing function $s_t$ (red).

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 $s_t = m_t$. We will see this phenomenon repeated throughout the rest of this paper. That itself is quite stunning.

Figure 2: The offered load $m_t$ for the sinusoidal arrival-rate function in (2.8) 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.

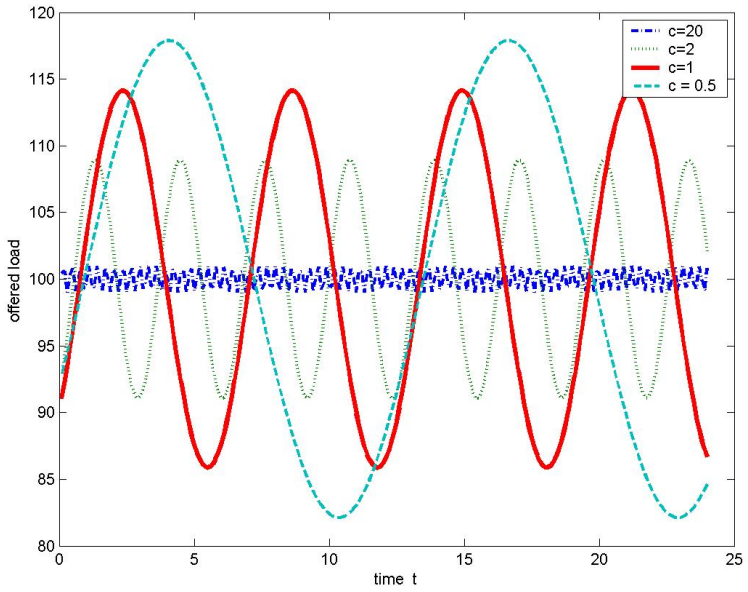
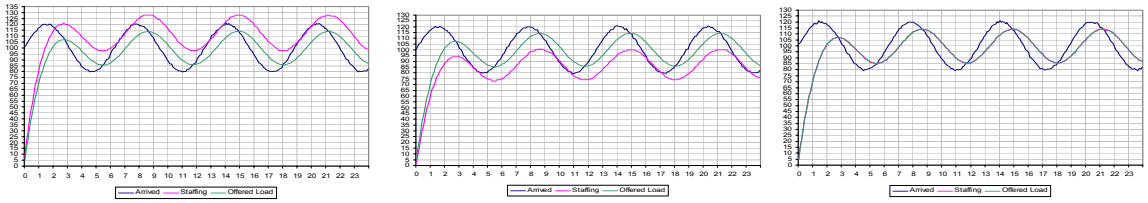



Figure 3: **Staffing function for: (1) Target $\alpha = 0.1$ (2) Target $\alpha = 0.9$ (3) Target $\alpha = 0.5$**



### 2.1.2 Time-Stable Performance

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 $\alpha$ 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.

In addition to stabilizing the delay probability, other performance measures (e.g. utilization, tail probabilities abandon probabilities, etc.) are found to be quite stable as well. Precise explanations and definitions of the performance measures are given in Section 3.2. Below are summary results graphs for all target $\alpha$'s.

However, as the target delay probability increases toward heavy loading, the abandonment probability becomes much less time-stable. We discuss this phenomenon further in Subsection 2.2 below. Other measures of congestion such as average waiting time and average queue length were also found to be relatively stable, but like the abandonment probabilities, these too become less time-stable under heavy loads.

Figure 4: **Delay probability summary for various $\alpha$'s.**

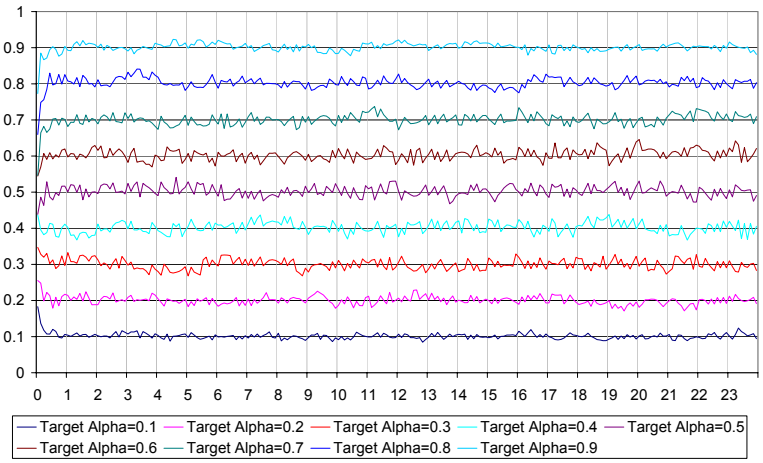
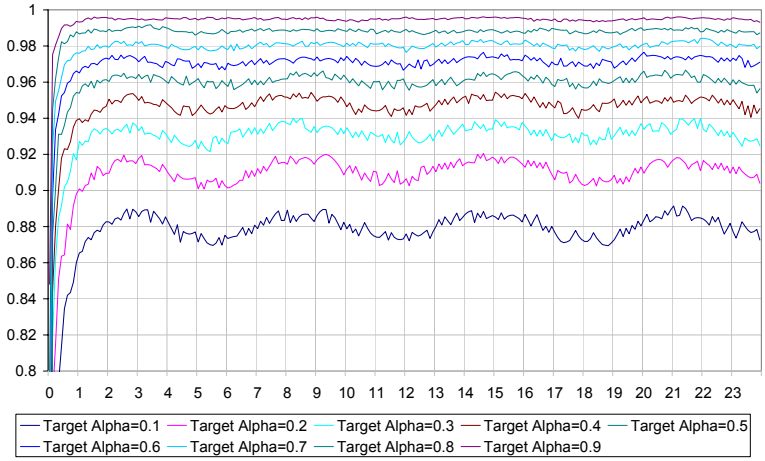



Figure 5: **Utilization summary for the Erlang-A example**




### 2.1.3 Validating the Square-Root-Staffing Formula

Next we define an **implied service grade**: A function $\{\beta_t : 0 \le t \le T\}$ is defined by setting

$$\beta_t \equiv \frac{s_t - m_t}{\sqrt{m_t}}, \quad 0 \le t \le T\ , \tag{2.11}$$

where $m_t$ is again the offered load in (1.2) and (2.10). and $s_t$ is the staffing function obtained by the ISA algorithm. Since $s_t$ is obtained from the ISA algorithm, the function $\beta_t$ is itself obtained from the ISA algorithm. It thus becomes interesting to see if the implied service grade is approximately constant as a function of time. And, indeed, it is, as shown in Figure 9.

Figure 9 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 $\alpha$ for a wide range of $\alpha$. Then Figure 9 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.

Figure 6: **Tail probability summary for the Erlang-A example**

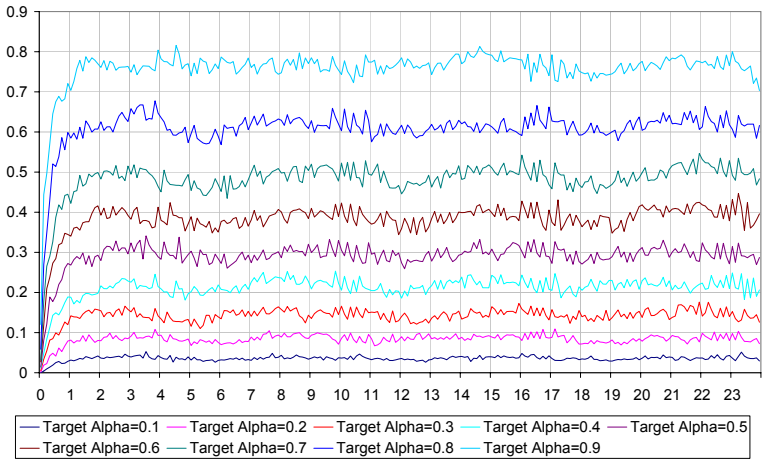



Figure 7: **Abandon probability summary for the Erlang-A example**

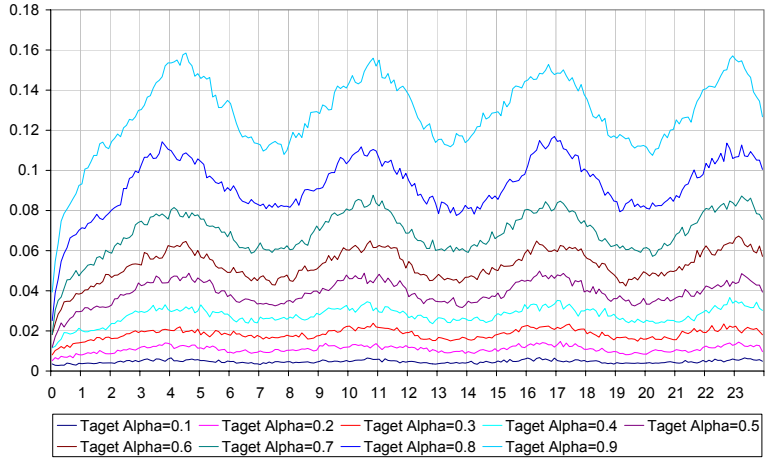
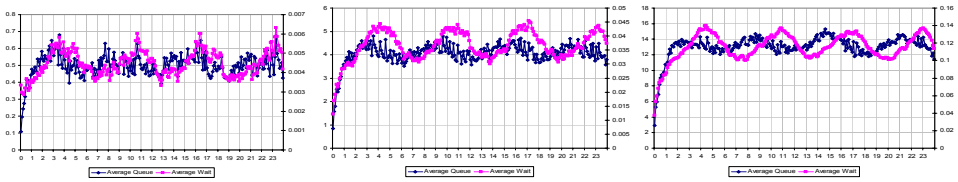



Figure 8: **Congestion: (1) Target $\alpha = 0.1$ (2) Target $\alpha = 0.5$ (3) Target $\alpha = 0.9$**



### 2.1.4 Relating $\beta$ to $\alpha$

However, one issues remains: In order to staff directly by the square-root staffing formula, we need to be able to relate the grade of service $\beta$ to the target delay probability $\alpha$. Indeed, we want a function mapping $\alpha$ into $\beta$. We propose a simple answer: For the time-varying Erlang-$A$ model, **use the associated stationary Erlang-$A$ model**, i.e., the $M/M/s + M$ model. Moreover, we suggest using simple formulas obtained from the many-

Figure 9: **Summary of Implied Service-grade $\beta$**


server heavy-traffic limit for the Erlang-$A$ model in Garnett et al. (2002). The **Garnett function** mapping $\beta$ into $\alpha$ is

$$\alpha = \left[1 + \sqrt{\frac{\theta}{\mu}} \cdot \frac{h(\hat{\beta})}{h(-\beta)}\right]^{-1}, \qquad -\infty < \beta < \infty; \tag{2.12}$$

where $\hat{\beta} = \beta\sqrt{\theta/\mu}$, with $\mu$ the individual service rate and $\theta$ 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 $\phi$ being the *probability density function* (pdf) and $\Phi$ the cdf. Of course, we want a function mapping $\alpha$ into $\beta$. Thus, we use the **inverse of the Garnett function**, which is well defined.

We now look at additional simulation output, aimed at establishing the validity of this stationary-model approach of relating $\alpha$ and $\beta$. First, we compare the empirical distribution of the customer waiting times to the theoretical distribution of those waiting times in the stationary Erlang-$A$ model. Specifically, in Figure 10 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. 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.

In Figure 10 we compare 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). Note that the approximation for the conditional waiting-time distribution in the stationary queues matches the performance of our time-varying model remarkably well.

We now go on to relate the empirical $(\alpha, \beta)$ pairs to the Garnett function in (2.12). 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 9. In Figure 11, 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 subsection by observing that other common approximations, such as the PSA or the SSA (the simple stationary approximation, using the overall time-average arrival rate; see Jennings et al. (1996)) perform poorly for this example. Demonstrations are omitted for lack of space, but such examples were already given in Jennings et al. (1996).

Figure 10: **Waiting time given wait: (1) Target $\alpha = 0.1$ (2) Target $\alpha = 0.5$ (3) Target $\alpha = 0.9$**

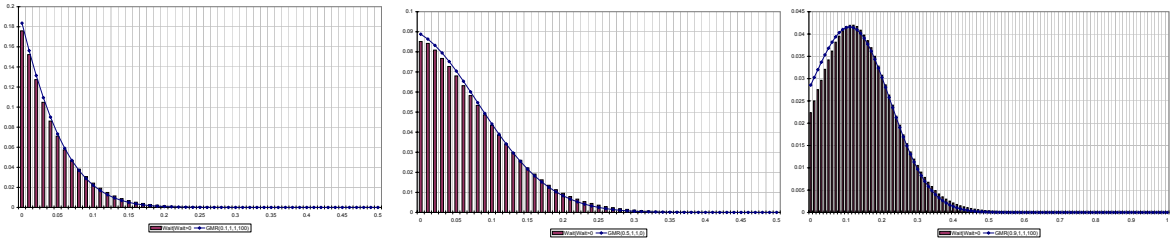
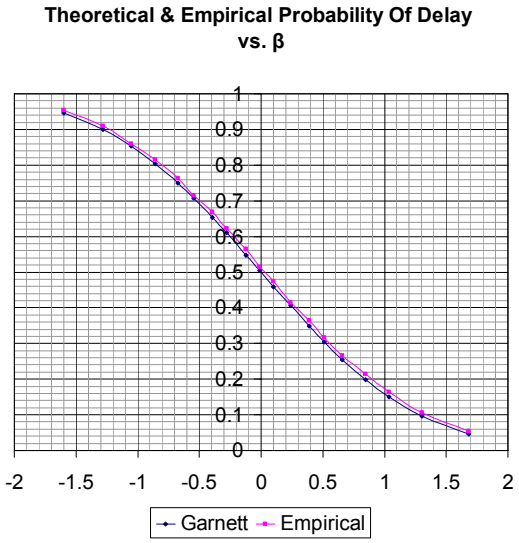


Figure 11: **Algorithm-Generated Performance vs. the Garnett Function**




## 2.2 Theoretical Motivation: The Case $\theta = \mu$

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

### 2.2.1 Connections to Other Models

With that condition, it is easy to relate the $M_t/M/s_t + M$ model to two other models that have been fully analyzed previously: 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 time-dependent family of stationary Erlang-$A$ models (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 just simulated in the previous subsection.

To express the relations, let $\{s_t : t \geq 0\}$ be an arbitrary staffing function; let $L_t \equiv L_t(M_t/M/s_t + M)$ be the **number of customers in the system** at time $t$, let $W_t^q$ 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 $t$). 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 $\{L_t : t \geq 0\}$ has the same distribution (finite-dimensional distributions) in the two models $M_t/M/s_t + M$ and $M_t/M/\infty$, i.e.,

$$\{L_t(M_t/M/s_t + M) : t \geq 0\} \stackrel{\mathrm{d}}{=} \{L_t(M_t/M/\infty) : t \geq 0\} \; . \tag{2.13}$$

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 $L_t$ have the same distribution as the steady-state number in system $L_\infty$ in the corresponding stationary model with appropriate arrival rate and number of servers.

To state the first of these results, let the service-time random variable $S$ have an exponential distribution with mean $1/\mu$. First, for each $t$,

$$L_t(M_t/M/\infty) \stackrel{\mathrm{d}}{=} L_\infty(M/M/\infty) \; , \tag{2.14}$$

where the constant arrival rate in the stationary $M/M/\infty$ model depends on $t$; in particular, the constant arrival rate $\hat{\lambda}_t$ in the $M/M/\infty$ model is chosen to be $\hat{\lambda}_t \equiv \mu m_t$, where $m_t$ is the expected number in system in the time-dependent infinite-server model in (1.2). Since $S$ has an exponential distribution, $S_e \stackrel{\mathrm{d}}{=} S$.

Theorem 1 of Eick et al. (1993a) states that, for the $M_t/M/\infty$ model with time-dependent arrival-rate function, for each $t$, $L_t$ has (exactly) a Poisson distribution with the mean $m_t$ in (1.2). On the other hand, in the stationary $M/M/\infty$ model, $L_\infty$ has a Poisson distribution with mean $m = \lambda/\mu$. Hence, by letting the fixed arrival-rate in the stationary $M/M/\infty$ model be $\hat{\lambda}_t$ above, the limiting steady-state (stationary) number in system $L_\infty$ also has a Poisson distribution with mean $m_t$.

By essentially the same reasoning, for each $t$, we can connect the distribution of $L_t$ to that in a stationary Erlang-$A$ model:

$$L_t(M_t/M/s_t + M) \stackrel{\mathrm{d}}{=} L_\infty(M/M/s + M) \; , \tag{2.15}$$

where the constant staffing level in the stationary $M/M/s + M$ model is chosen to be $\hat{s}_t \equiv s_t$ and the constant arrival rate is chosen to be $\hat{\lambda}_t$ above. Actually $L_\infty$ in the $M/M/s + M$ model is independent of $s$.

### 2.2.2 The Delay Probability

The virtual waiting time $W_t^q$ and the virtual abandonment probability $P_t(Ab)$ in the $M_t/M/s_t + M$ model are considerably more complicated. Even though it is difficult to evaluate the full distribution of $W_t^q$, 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

$$\begin{aligned} \alpha_t &\equiv P(W_t^q(M_t/M/s_t + M) > 0) = P(L_t(M_t/M/s_t + M) \geq s_t) \\ &= P(Poisson(m_t) \geq s_t) \approx P\left(N(0,1) > \frac{s_t - m_t}{\sqrt{m_t}}\right) \equiv \Phi^c\left(\frac{s_t - m_t}{\sqrt{m_t}}\right) \; , \end{aligned} \tag{2.16}$$

where $\Phi(x)$ is again the standard normal cdf, $\Phi^c(x) \equiv 1 - \Phi(x)$ is the associated *complementary cdf* and $m_t$ is the offered load in (1.2). From (2.16), we immediately obtain the square-root staffing rule in (1.1), where $\alpha$ is the target delay probability and $\beta$ is the associated target grade of service, with $\alpha$ and $\beta$ related according to

$$\alpha = P(N(0,1) > \beta) \equiv \Phi^c(\beta) \; . \tag{2.17}$$

As easily can be verified directly, the Garnett function $\alpha(\beta)$ in (2.12) reduces to simply $\Phi^c(\beta)$, as in (2.17), when $\mu = \theta$.

When aiming for a certain target delay probability $\alpha$ at all times (which is equivalent to aiming for a target grade of service $\beta$ at all times), approximation (2.16) dictates that we should choose $s_t$ according to (1.1) and (2.17). Since (2.17) agrees with (2.12) in this case (with $\mu = \theta$), we have provided theoretical support for the square-root staffing formula, using the associated stationary model to relate $\alpha$ and $\beta$.

Since the only approximation in (2.16) is the normal approximation for the Poisson distribution, we can anticipate that the approximation will perform extremely well unless $m_t$ is very small. In particular, by this argument, we have **proved** that we do indeed achieve **asymptotically time-stable delay probability** $\alpha$ in the $\mathbf{M_t/M/s_t + M}$ model with $\mu = \theta$ as $\mathbf{m_t} \to \infty$ when we staff according to (1.1) and (2.12). As a consequence, we have given a theoretical explanation for the regularity observed in Figure 4.

### 2.2.3 Approximations for the Waiting-Time Distribution

However, from Figures 7 and 10, we see that the virtual abandonment probability $P_t(Ab)$ and the expected virtual waiting time $E[W_t^q]$ fluctuate much more than the delay probability. We will explain that greater fluctuation.

We actually can mathematically analyze the time-dependent virtual waiting time $W_t^q$ and the time-dependent virtual abandonment probability $P_t(Ab)$. Here is an important initial observation: Conditional on the event that $W_t^q > 0$, whose probability we have analyzed above, $W_t^q$ 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, $W_t^q$ is distributed as the first passage time of the pure-death stochastic process with state-dependent death rate $\mu L_u$ for $u \geq t$ down from the initial value $L_t$ to the curve $\{s_u : u \geq t\}$. (Of course, $W_t^q = 0$ if $L_t < s_t$.) As a consequence, the distribution of $W_t^q$ 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 $W_t^q$ 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, $W_t^q$ 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 $W_t^q(M_t/M/s_t+M)$ and $P_t(Ab)(M_t/M/s_t+M)$, 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:

$$W_t^q \equiv W_t^q(M_t/M/s_t + M) \approx W_\infty^q(M/M/s + M) \equiv W_\infty^q \tag{2.18}$$

and

$$P_t(Ab) \equiv P_t(Ab; M_t/M/s_t + M) \approx P_\infty(Ab; M/M/s + M) \equiv P_\infty(Ab) \ , \tag{2.19}$$

where the constant staffing level in the stationary $M/M/s + M$ model on the righthand sides is chosen to be $\hat{s}_t \equiv s_t$ and the constant arrival rate is chosen to be $\hat{\lambda}_t$ above. Hence, we propose (2.18) and (2.19) as approximations.

Given approximations (2.18) and (2.19), we can use established results for the stationary $M/M/s + M$ model, e.g., as in Garnett et al. (2002) and Whitt (2004, 2005). For example, algorithms to compute the (exact) distribution of $W^q_\infty$ are given there, including the corresponding conditional distributions obtained when we condition on whether or not the customer eventually is served.

### 2.2.4 Asymptotic Time-Stability in the Many-Server Heavy-Traffic Limit

As in the literature for stationary models, e.g., Garnett et al. (2002), important insight can be gained by considering many-server heavy-traffic limits. That is achieved for our $M_t/M/s_t + M$ model, by considering a sequence of models indexed by $n$, where the arrival-rate function is allowed to depend upon $n$. We can leave the service rate and abandonment rate unchanged, independent of $n$ (and $t$). Thus, for each $n$, we have arrival-rate function $\lambda_n \equiv \{\lambda_n(t) : t \geq 0\}$. As in the stationary context, we want to let the arrival rate increase as $n \to \infty$. However, now we need to carefully specify how the entire function $\lambda_n$ increases. Since we are staffing in response to the arrival rate, we do not need to make any direct assumptions about the staffing levels $s_t$. We will assume that we staff according to the square-root-staffing formula (1.1) with a fixed target delay probability $\alpha$. We then want to determine when that yields asymptotically time-stable performance.

As an initial condition, we want to assume that $\lambda_n(t) \to \infty$ as $n \to \infty$ for each $t$, but we will need more than that. From the analysis so far, it is clear that we need $m_{t,n} \to \infty$, where $m_{t,n}$ is the time-dependent mean number in the $n^{\text{th}}$ $M_t/M/\infty$ model. However, that actually is not enough to get asymptotic time-stability for quantities such as the mean virtual waiting time $E[W^q_t]$ and the virtual abandonment probability $P_t(Ab)$.

To proceed, we exploit the approximations in (2.18) and (2.19). From approximation (2.19), we obtain the associated approximation

$$E[W^q_t] \equiv E[W^q_t(M_t/M/s_t + M)] \approx E[W^q_\infty(M/M/s + M)] \tag{2.20}$$

where the constant staffing level in the stationary $M/M/s + M$ model on the righthand sides is chosen to be $\hat{s}_t \equiv s_t$ and the constant arrival rate is chosen to be as in (**??**).

Now we observe that previous heavy-traffic limits for the Erlang-$A$ model in the QED regime, Theorems 3 and 4 of Garnett et al. (2002), imply that

$$\sqrt{m_t}P_t(Ab)(M_t/M/s_t + M) \to \eta \quad \text{and} \quad \sqrt{m_t}E[W^q_t(M_t/M/s_t + M)] \to \frac{\eta}{\theta} \tag{2.21}$$

as $m_t \to \infty$, where

$$\eta \equiv \alpha E[N(0,1) - \beta | N(0,1) > \beta] = \alpha \left( \frac{\phi(\beta)}{\Phi^c(\beta)} - 1 \right) > 0 \tag{2.22}$$

and $\theta = \mu$.

The important practical conclusion we deduce from (2.21) is that we see that $\sqrt{m_t}P_t(Ab)$ and $\sqrt{m_t}E[W^q_t]$ will be asymptotically constant (time-stable and nondegenerate) as $m_t$ increases if we are in the QED regime.

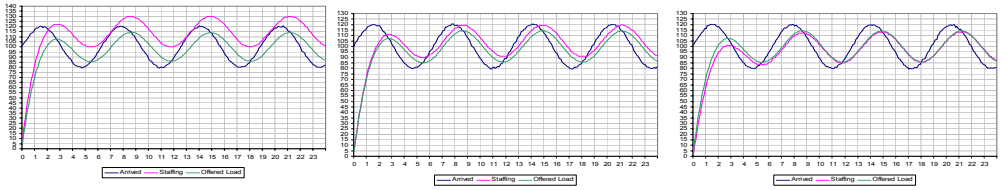
However, in general, consistent with Figures 7 and 10, the performance measures $P_t(Ab)$ and $E[W_t^q]$ themselves need not be asymptotically time-stable. In order for them to be asymptotically time-stable too, we need to ensure that the mean function $m_t$ itself is asymptotically time-stable, which requires an extra condition.

We actual see the greatest departures from time-stability of $P_t(Ab)$ and $E[W_t^q]$ for the $M_t/M/s_t + M$ model in Figures 7 and 10 when the target delay probability is large. In those cases, it is evident that the system actually should be regarded as in the ED regime, not the QED regime. From Garnett et al. (2002) and Whitt (2004), we can see the appropriate ED asymptotics, which also suggests that time-stability will not hold for the performance measures $P_t(Ab)$ and $E[W_t^q]$, staffing as we have done. Moreover, it suggests that we might consider a different staffing method designed to achieve time-stable abandonment in the ED regime. In particular, ISA extends directly by changing the target performance measure from the delay probability to the abandonment probability. The performance of such alternative iterative-staffing algorithms is a topic for future research.

## 2.3 The Time-Varying Erlang-C Model

For comparison, we now show the performance of ISA for the same system described in Section 2.1 only without abandonment (infinite patience). 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.

Figure 12: **Staffing levels: (1) Target $\alpha = 0.1$ (2) Target $\alpha = 0.5$ (3) Target $\alpha = 0.9$**



### 2.3.1 Time-Stable Performance

As before, we achieve accurate time-stable delay probabilities when we apply the ISA.

The service grade $\beta$ is stabilizing as well, only in much slower rate, as can be seen below for large $\alpha$'s.

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).

Figure 17 shows that here the time until system reaches (dynamic) steady-state is much longer compared to a system with abandonment. In fact, steady-state was not yet reached after 24 time-units in the case above.

### 2.3.2 Validating the Square-Root-Staffing Formula

Just as for the time-varying Erlang-$A$ model, we want to validate the square-root-staffing formula in (1.1). We thus repeat the various experiments we did in Section 2.1.

Figure 13: **Delay probability summary for the Erlang-C example**

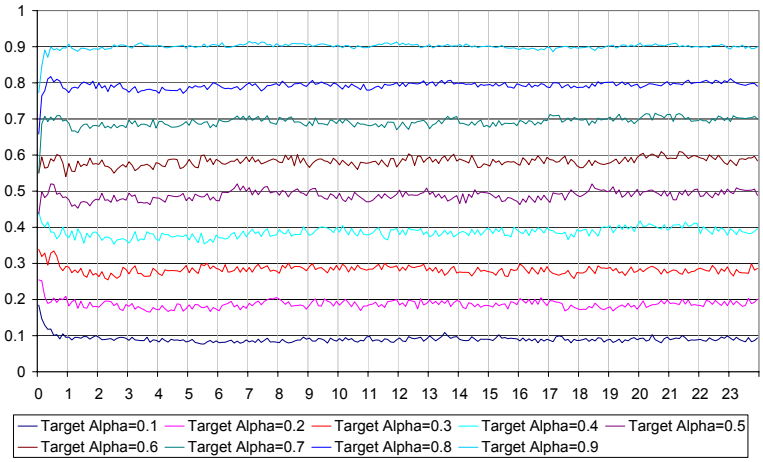
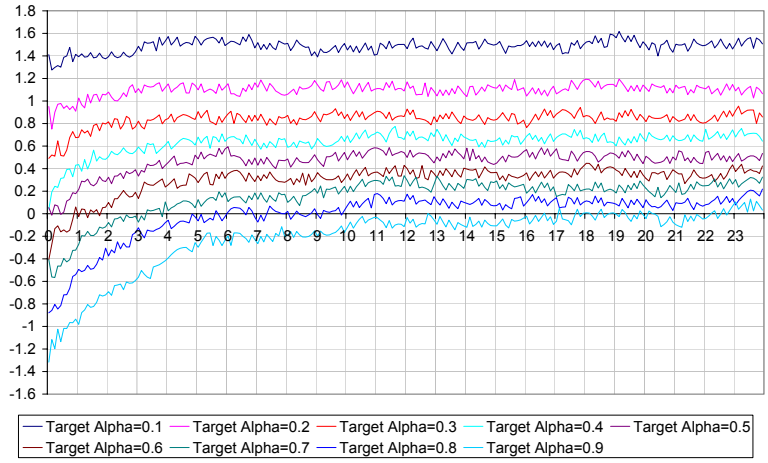



Figure 14: **Implied service grade $\beta$ summary for the Erlang-C example**




Recall that, for the *stationary* $M/M/s$ queue, the conditional waiting-time $(W_q \mid W_q > 0)$ is (exactly) exponentially distributed. As seen in Figure 17, the empirical conditional waiting-time distribution given wait, in our *time-varying* queue and over *all* customers, also fits the exponential distribution well . The mean of the plotted exponential distribution was taken to be the overall average waiting time of those who were actually delayed during $[0, T]$.

Here, the relation between $\alpha$ and $\beta$ is compared with the **Half-Whitt function** from Halfin and Whitt (1981), namely,

$$P\{delay\} \equiv \alpha \equiv \alpha(\beta) \approx \left[1 + \beta \cdot \frac{\Phi(\beta)}{\phi(\beta)}\right]^{-1}, \quad 0 < \beta < \infty\,, \tag{2.23}$$

where $\phi$ is again the pdf) associated with the standard normal cdf $\Phi$. The Half-Whitt function in (2.23) is obtained from the Garnett function in (2.12) by letting $\theta \to 0$.

Just as we use the Garnett function to relate the target delay probability $\alpha$ to the grade of service $\beta$ in the square-root-staffing formula in (1.1) for the $M_t/M/s_t + M$ model, so we use the Half-Whitt function to relate

Figure 15: **Utilization summary for the Erlang-C example**

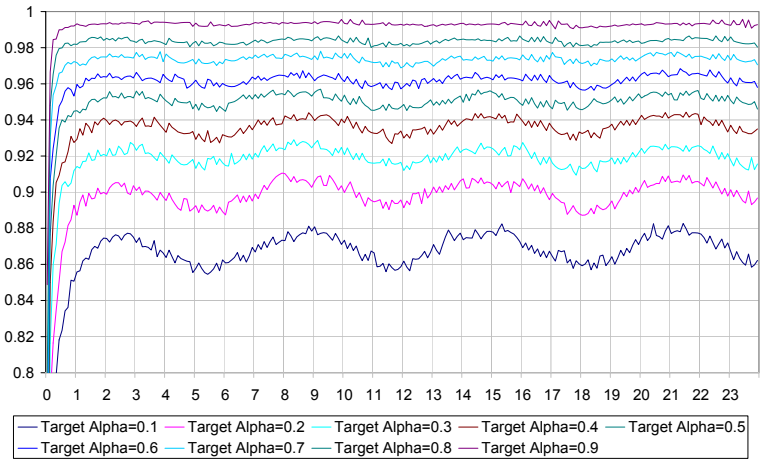



Figure 16: **Tail probability summary for the Erlang-C example**

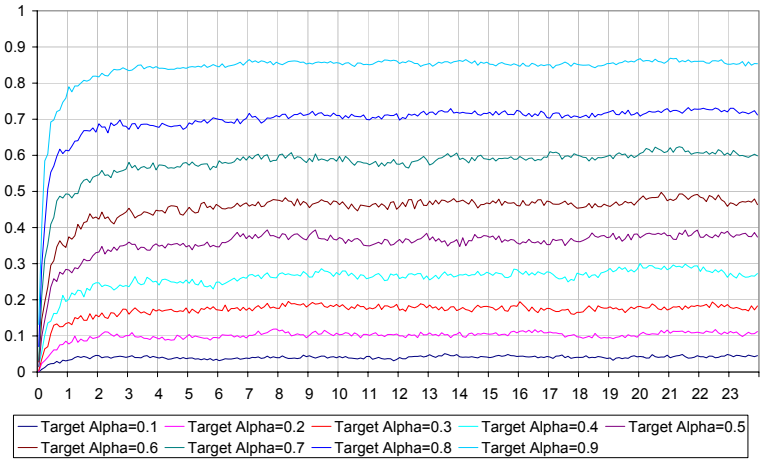
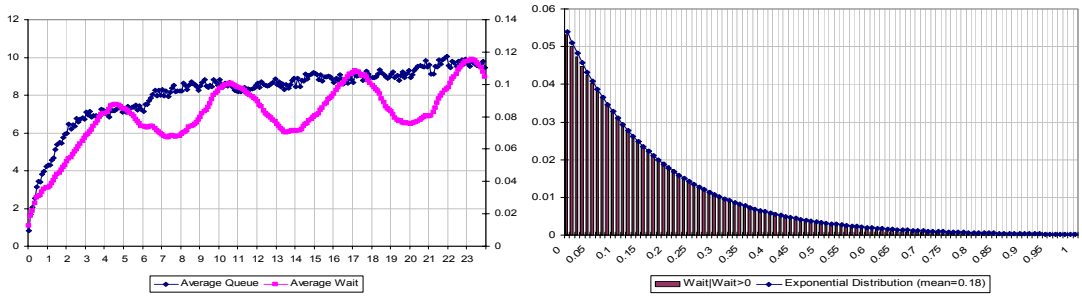
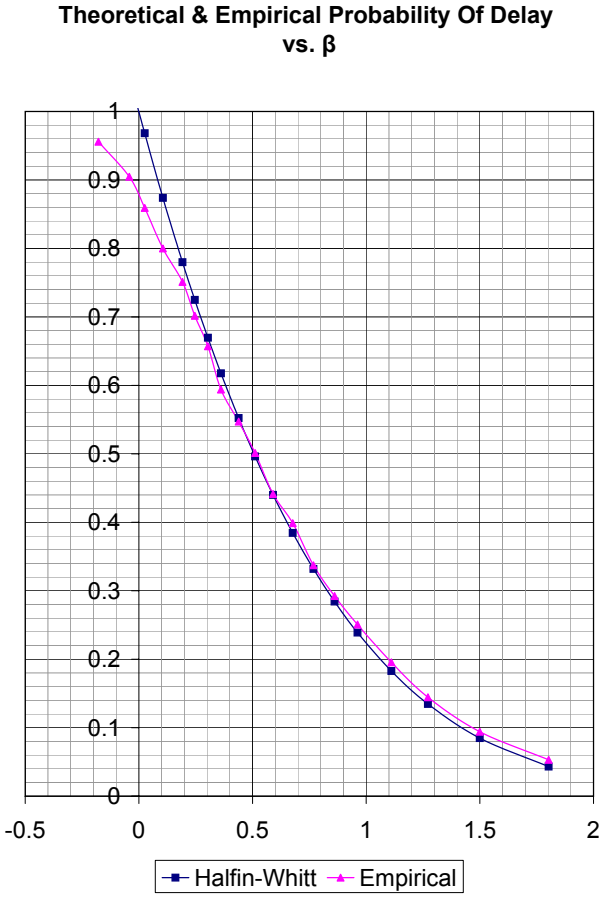



Figure 17: **Target $\alpha$=0.5: (1) Congestion (2) Waiting time given wait distribution**




$\alpha$ to $\beta$ in the square-root-staffing formula in (1.1) 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 18 are again remarkable.

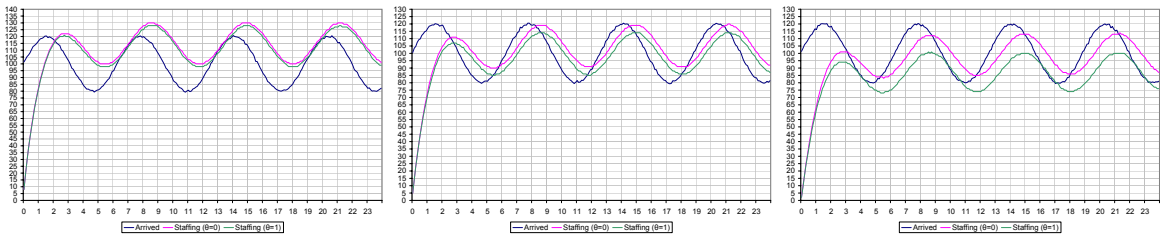
Figure 18: **Comparison of empirical results with the Halfin-Whitt approximation**




## 2.4 Benefits of Taking Account of Abandonment

The following graphs show the benefit of staffing a system taking account of abandonment (assuming that it in fact occurs). When compared to the model without abandonment, abandonment in the model reduces the required staff. We show the difference between staffing levels for the two models introduced in §2.1 and §2.3, in the three regimes of operation: *QD*, *QED* and *ED*.

Figure 19: **Staffing levels: (1) $\alpha = 0.1$ (2) $\alpha = 0.5$ (3) $\alpha = 0.9$**



It is natural to quantify the savings of labor by the area between the curves. In this case, the savings in labor, had one used $\theta = 1$, is 46.5 time units when $\alpha = 0.1$, 113.3 when $\alpha = 0.5$, and 256.4 when $\alpha = 0.9$. It may perhaps be better to quantify savings by looking at the savings of labor per shift. Dividing the saving

in time-units by the number of time-units they are taken over, we come up with savings of about 2, 5 and 12 servers per shift, for $\alpha = 0.1, 0.5, 0.9$ respectively. The labor savings increases as $\alpha$ increases.

# 3 The Simulation-Based Iterative-Staffing Algorithm

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, such as server utilization, tail probabilities, average waits and abandonment probabilities, remain stable as well.

## 3.1 The ISA

For our implementation of the algorithm, we assume that we have an $M_t/G/s_t + G \equiv M_t/GI/s_t + GI$ 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.) For application of our algorithm, assuming that we use the $M_t/G/s_t+G$ model, there are issues about model fitting. For discussion about fitting non-homogeneous Poisson arrival processes, see Massey, Parker and Whitt (1996).

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 $X_n(t)$ denote the value at time $t$ in the $n^{\text{th}}$ 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 $\Delta$. In all experiments presented in this paper, we use $\Delta = 0.1$. 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 can generate arrival times by thinning a single Poisson process with a constant rate $\lambda^*$ 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). Alternatively, the times between successive arrivals can be generated as independent events, according to probability distributions, determined by the last customer arrival time, and adjusted if necessary at ends of shifts.

In this section, let $s_n(t)$ be the staffing level at time $t$ in iteration $n$ for $0 \le t \le T$. Let $L_n(t)$ denote the random total number of customers in the system at time $t$, under this staffing function. We estimate the distribution of $L_n(t)$ 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 two examples in Section 2, it suffices to let $s_0(t) = 200$ 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 $\tau$, which we take to be 1.)

### 3.1.1 The Steps of ISA

1. Given the $i^{\text{th}}$ staffing function $\{s_i(t) : 0 \leq t \leq T\}$, evaluate the distribution of $L_i(t)$, for all $t$, using simulation.

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

$$s_{i+1}(t) = \arg\min \{c \in \mathbb{N} : P\{L_i(t) \geq c\} < \alpha\} . \tag{3.24}$$

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

$$\|s_{i+1}(\cdot) - s_i(\cdot)\|_\infty \equiv \max \{|s_{i+1}(t) - s_i(t)| : 0 \leq t \leq T\} \leq \tau , \tag{3.25}$$

then stop and let $s_{i+1}(\cdot)$ 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 $\tau = 1$.) ∎

### 3.1.2 Implementation and Convergence

For further discussion, let $\infty$ denote the index of the last iteration of ISA, so that $s_\infty(t)$ denotes the final staffing level at time $t$ and $L_\infty(t)$ denotes the number in system at time $t$ with that staffing function.. Then, if the algorithm converges, it converges to a staffing function $s_\infty(\cdot)$ for which $P\{L_\infty(t) \geq s_\infty(t)\} \approx \alpha$, $0 \leq t \leq T$.

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 Section 5. That proof shows convergence to a limit, but the limit does not necessarily meet the target delay probability; it is a best possible staffing. Experience indicates that the algorithm consistently converges and does so relatively rapidly. The number of iterations required depends on the parameters, especially the ratio $\mathbf{r} \equiv \theta/\mu$. If $\mathbf{r} = 1$, corresponding to an infinite-server queue as noted in Section 2.2, 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 Section 5.

## 3.2 Performance Measures

Throughout this paper we present several performance measures. Their method of estimation will now be described. Most measures are time-varying. We define them for each time-interval $t$, and graph their values as function over $t \in [0, T]$. Other measures are global. They are calculated either as total counts (e.g. fraction abandoning during $[0, T]$), or via time-averages. We used $T = 24$ in all our simulations.

For replication $k$, the **delay probability** *in interval t* is estimated by the fraction of customers who are not served immediately upon arrival, out of all arriving customers during the $t$ time-interval. Namely, for the $k^{\text{th}}$ replication, the estimator is:

$$\hat{\alpha}_k(t) = \frac{\sum_i 1\{\textit{customer\_i\_entered\_queue\_at\_interval\_t}\}}{\sum_i 1\{\textit{customer\_i\_entered\_system\_at\_interval\_t}\}} \equiv \frac{\hat{Q}_k(t)}{\hat{S}_k(t)} . \tag{3.26}$$

We obtain the overall estimator $\hat{\alpha}(t)$ by averaging over all replications. That was found to be essentially the same as (identical to for our purposes) the ratio of the average of $\hat{Q}_k(t)$ over all replications to the average of $\hat{S}_k(t)$.

For replication $k$, the estimator of the **average waiting time** *in interval t* is defined in an analogous way by

$$\hat{w}_k(t) = \frac{\sum_i w_i 1\{\textit{customer\_i\_entered\_system\_at\_interval\_t}\}}{\sum_i 1\{\textit{customer\_i\_entered\_system\_at\_interval\_t}\}} \tag{3.27}$$

where $w_i$ is the total waiting time of customer $i$. Again we obtain the overall estimator $\hat{w}(t)$ by averaging over all replications.

The **average queue length** *in interval t* is taken to be constant over the time-interval. The queue length is also averaged over all replications. By the **tail probability** *in interval t* we mean specifically the probability that queue size is greater than or equal to 5 (taking 5 to be illustrative). Specifically, the indicators $1\{L_\infty(t) - s_\infty(t) \geq 5\}$ are averaged over all replications.

For replication $k$, the estimator of the **server utilization** *in interval t* is the fraction of busy-servers during the time-interval, accounting for servers who are busy only a fraction of the interval:

$$\hat{\rho}_k(t) = \frac{\sum_{i=1}^{s_\infty(t)} b_i}{s_\infty(t) \cdot \Delta} \tag{3.28}$$

where $b_i$ denotes the busy time of server $i$ in interval $t$. Again, we obtain the overall estimator $\hat{\rho}(t)$ by averaging over all replications.

# 4 Additional Examples

## 4.1 The Challenging Example

In this section, we consider the "challenging example" presented in Jennings et al. (1996). It is a time-varying Erlang-$C$ model (no abandonment), with exponential service times having mean 1 and a nonhomogenous Poisson arrival process with the sinusoidal arrival-rate function $\lambda(t) = 30 + 20 \cdot \sin(5 \cdot t)$. We want to see how ISA performs on this same example. Figures 20 and 21 show that ISA also achieves time-stable performance for this example, for the full range of target delay probabilities, ranging from $0.1$ to $0.9$, just as before.

We now want to compare the empirical results, paralleling Figures 11 and 18. We do so for this example below in Figure 22. Again the results are spectacular. In Figure 22 we use the Half-Whitt function, just as in Figure 18. We also include the normal tail probability in (2.17), because that is the direct normal approximation used by Jennings et al. (1996), before they apply their refinement in their Section 4. That refinement is equivalent to working directly with the Half-Whitt function, as we propose here.

## 4.2 The $M_t/M/s_t + M$ Model with More and Less Patient Customers

We now return to the time-varying Erlang-$A$ model ($M_t/M/s_t + M$) considered in Section 2, except we change the patience parameter, i.e., the individual abandonment rate $\theta$. We consider two new cases: $\theta = 0.2$; then customers are **very patient**, since they are willing to wait, on average, five times the average service time;

Figure 20: **Delay probability summary for the Challenging example**

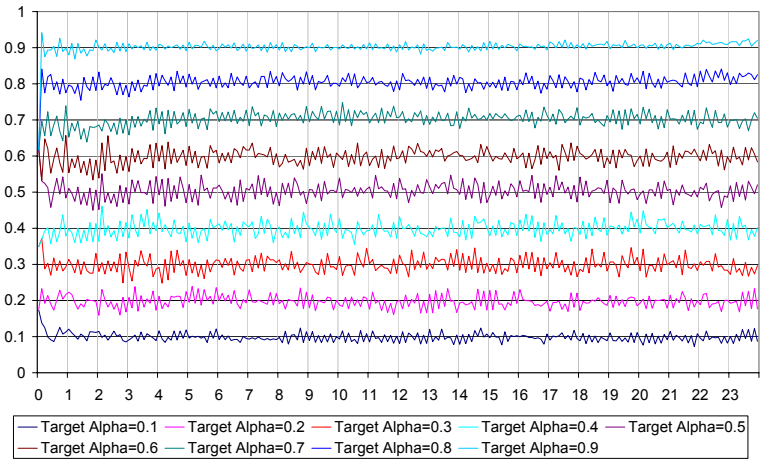
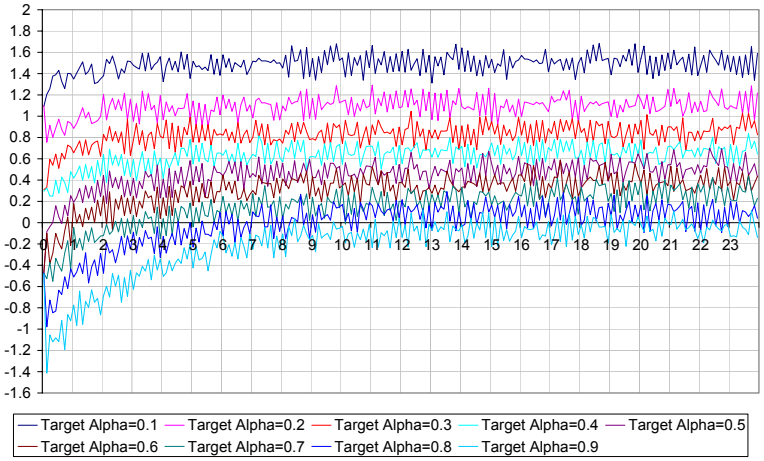



Figure 21: **Implied service grade $\beta$ summary**




and $\theta = 5.0$; then customers are **very impatient**, since they are willing to wait, on average, only one-fifth of the average service time.

In both cases, the target delay probability was achieved quite accurately and the service grade $\beta$ was stabilized, just as in the previous graphs. We compare the staffing levels for these alternative environments, for the three regimes QD ($\alpha = 0.1$), QED ($\alpha = 0.5$), and ED ($\alpha = 0.9$) in Figure 23 below. We compare the time-dependent abandonment $P_t(Ab)$ in these two scenarios in Figure 24.

We compare the empirical $(\alpha, \beta)$ pairs produced by ISA to the Garnett function in (2.12) for these two cases in Figure 25. We are no longer surprised to see that the fit is excellent.

From all our studies of ISA, we conclude that for the time-varying Erlang-$A$ model we can always use the square-root-staffing algorithm in (1.1), obtaining the required service grade $\beta$ from the target delay probability $\alpha$ by using the inverse of the Garnett function in (2.12), which reduces to the Half-Whitt function in (2.23) when $\theta = 0$. To see how the Garnett functions look, we plot the Garnett function for several values of the ratio $\mathbf{r} \equiv \theta/\mu$ in Figure 26 below.

Figure 22: **Comparison of empirical results with the Halfin-Whitt and Normal approximation**

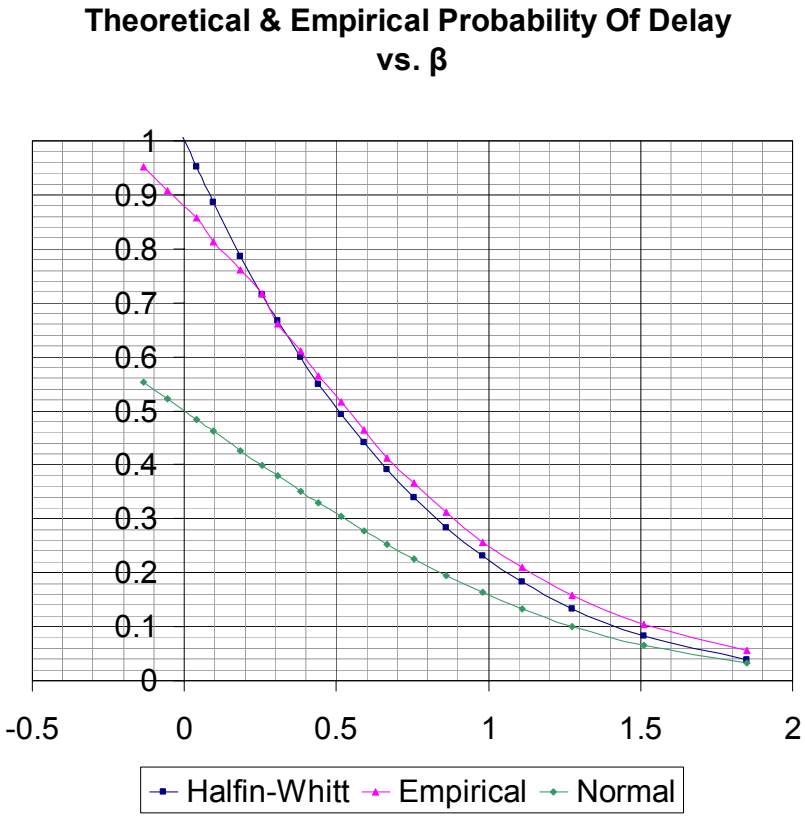
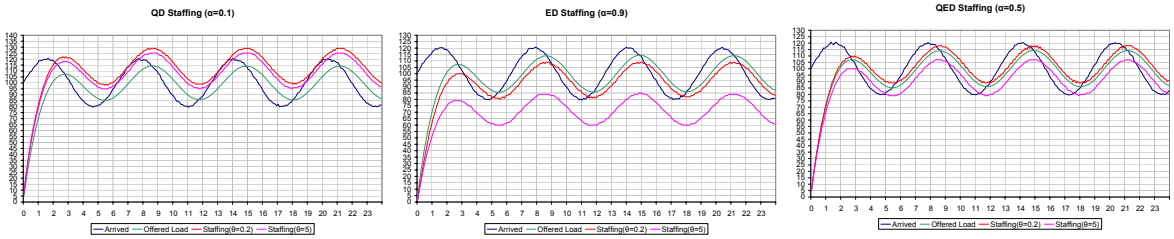



Figure 23: **Comparison of staffing levels for very patient and very impatient environments**




## 4.3 Benefits of Taking Account of Abandonment Again

Following §2.4, we now expand our comparison of staffing levels for (im)patience distribution with parameters $\theta = 0, 1, 5, 10$. Clearly, the required staffing level decreases as $\theta$ increases, bringing additional savings. In Figure 27 we show the comparison for delay probability $\alpha = 0.5$, which we consider to be a reasonable operational target.

Here, the labor savings is: 113.3 time units for $\theta = 1$, 270 time units for $\theta = 5$, and 386 time units for $\theta = 10$. The corresponding savings in workers per shift are about 5, 12 and 18 servers, for $\theta = 1, 5, 10$, respectively.

Figure 24: **Abandon probability: (1) $\theta$=5 (2) $\theta$=0.2**

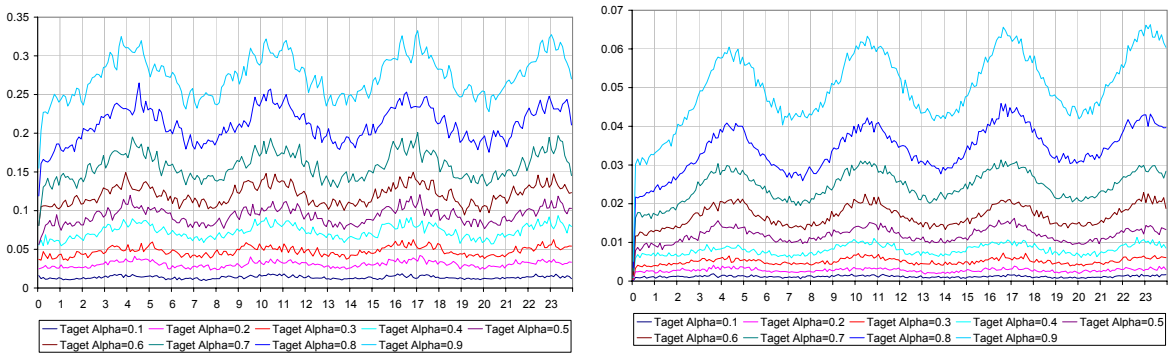
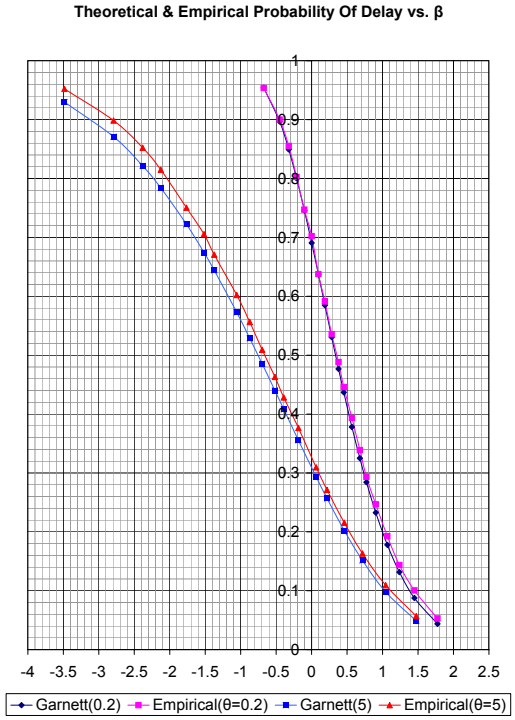



Figure 25: **Comparison of the empirical results with the Garnett approximation**




## 4.4 A Practical Example

In this section we consider the practical case that was first described in Figure 1. To make this example more realistic than previous examples, we decrease the mean service time from 1 hour to 6 minutes. That is achieved by letting $\mu = 10$. Corresponding to that, we let $\theta = 10$, so that we have $\theta = \mu$ as in Section 2.1. Results are shown below.

At first, we are struck by the observation that the algorithm 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. Moreover, not all performance measures are stable over the entire day. However, this bad behavior is quite clearly due to the extremely low arrival rates that prevail at the beginning and the end of the day. When the load is small, the

Figure 26: **The Halfin-Whitt/ Garnett functions**

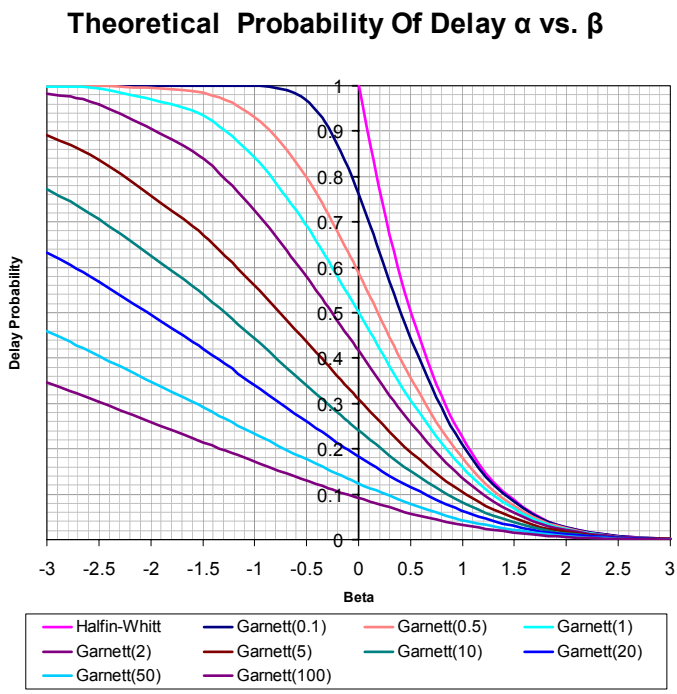
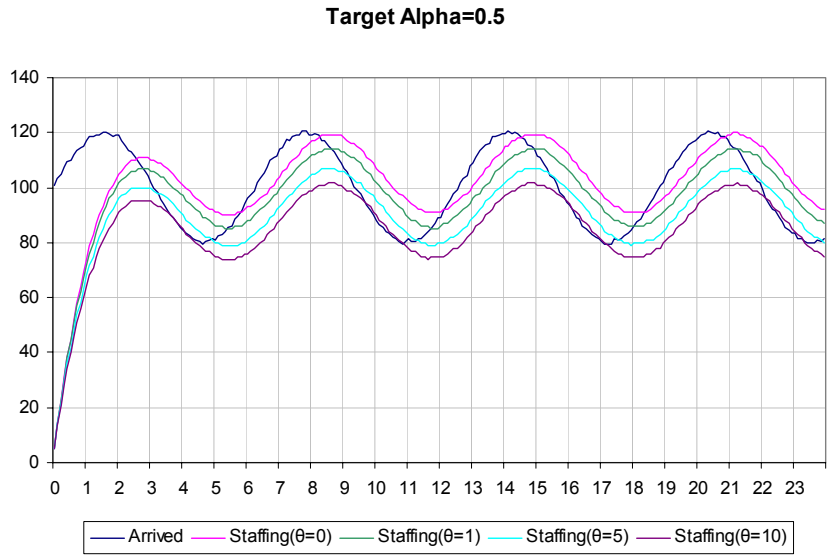



Figure 27: **Staffing under various (im)patience parameters**




addition or removal of a single server while greatly affect the delay probability. On the positive side, note that there is a clear time-interval - from 7 to 17, in which performance measures are very stable, and when operating under reasonable service grade (up to delay probability of 0.5), performance measures are varying in quite a small range, that would look appealing to most system designers.

Figure 28: **Delay probability summary for the practical example**

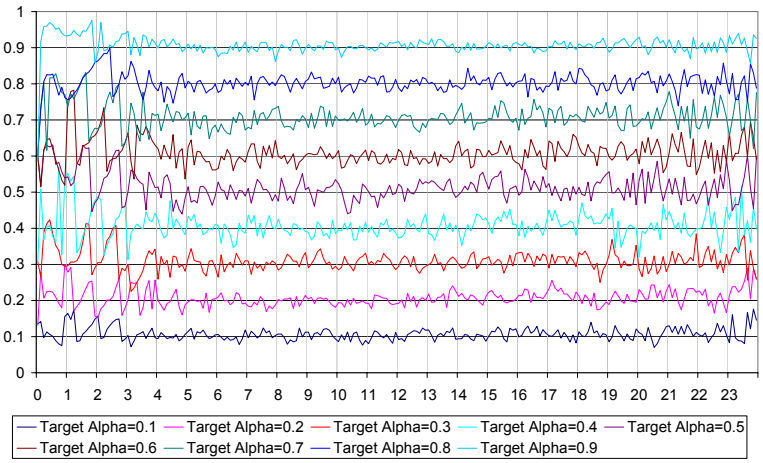
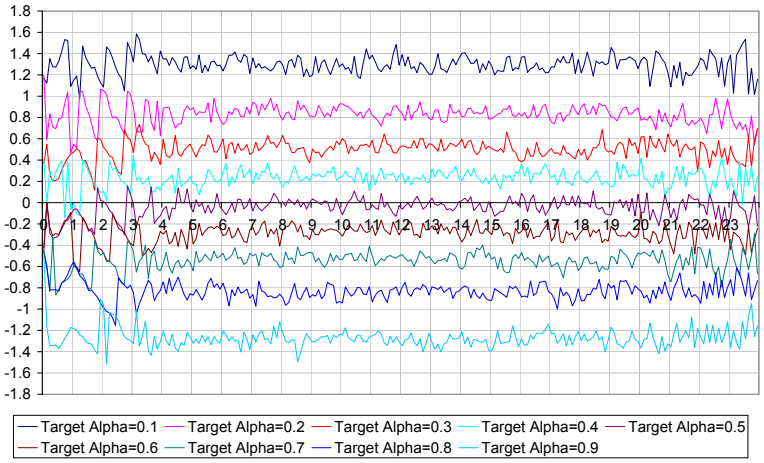



Figure 29: **Implied service grade $\beta$ summary for the practical example**




In Figures 32, 33, 34 we describe the performance of ISA in the three regions: QD ($\alpha = 0.1$), QED ($\alpha = 0.5$) and ED ($\alpha = 0.9$). There are several important observations to make here: First, note that in all cases the (infinite-server) offered load $m_t$ falls almost directly on top of the PSA offered load $\lambda(t)/\mu$, showing that in this case the square-root-staffing rule (1.1) will perform the same using the infinite-server offered load and the PSA offered load.

ISA does not differ much from PSA. However, for the time-varying Erlang-$A$ model, staffing using PSA is actually not routine.

The three regimes of operation are clearly revealed by the average waiting time: In the **QD** regime the average waiting time is relatively negligible; in the **QED** regime average waiting time is in seconds; and in the **ED** it is in minutes. Figure 33 shows, once again, that the staffing falls right on top of the offered load in the QED regime. Figure 35 shows that the excellent matching between the Garnett function and the empirical results is preserved also in this example.

Figure 30: **Abandon probability summary for the practical example**

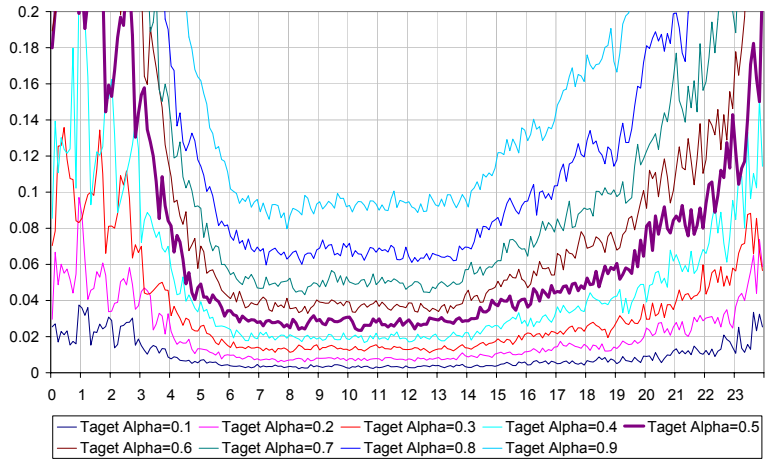
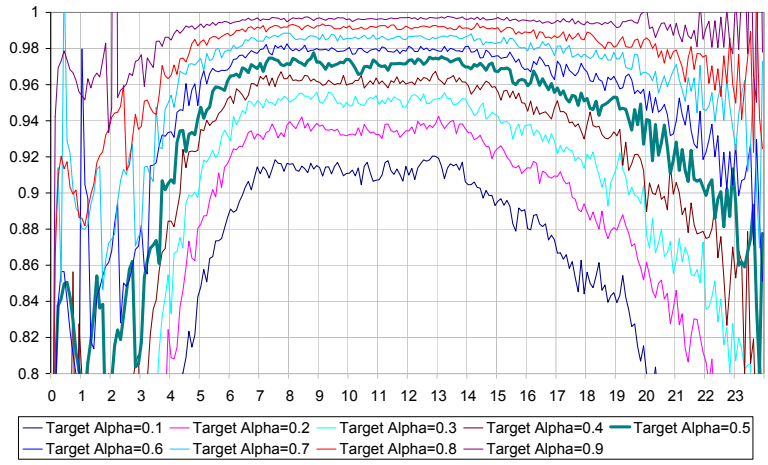



Figure 31: **Utilization summary for the practical example**




### 4.5 Non-Exponential Service Times

In addition to the time-varying Erlang-C and Erlang-A examples, we also ran experiments with different service-time distributions, such as deterministic and log-normal. The ISA was successful in achieving the desired target delay probability, and results showed time-stable performance, compatible with stationary theory, similar to here. For the case of deterministic service times, theory was taken from Jelenkovic, Mandelbaum and Momcilovic (2004).

## 5 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 $s_\infty(\cdot)$ and then afterwards we give a theoretical explanation.

Figure 32: **Target $\alpha$=0.1: (1) Staffing level, offered load and arrival function, (2) average queue and average waiting time (in average service time)**

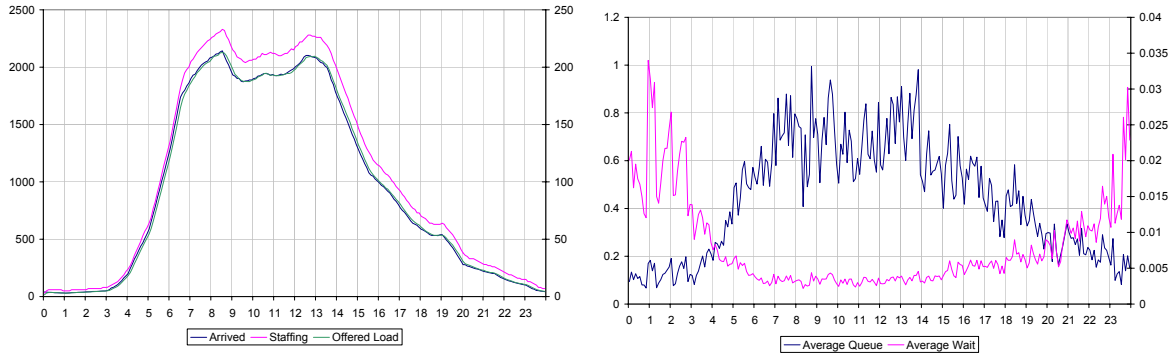



Figure 33: **Target $\alpha$=0.5: (1) Staffing level, offered load and arrival function, (2) average queue and average waiting time (in average service time)**

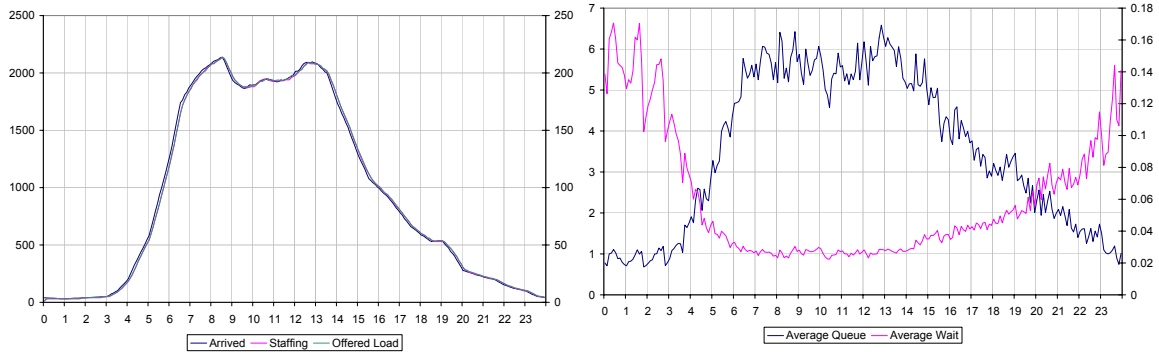
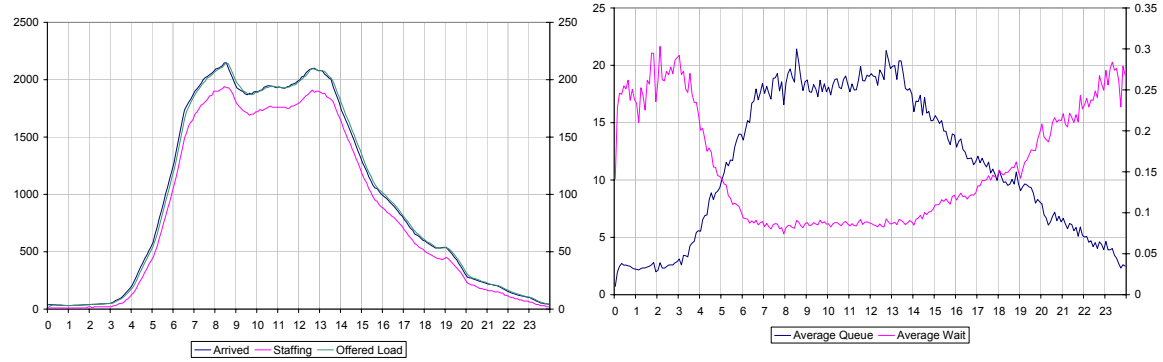
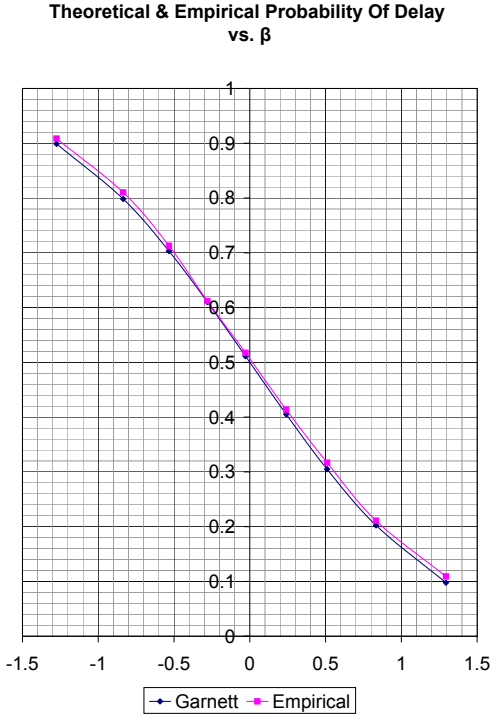



Figure 34: **Target $\alpha$=0.9: (1) Staffing level, offered load and arrival function, (2) average queue and average waiting time (in average service time)**




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

Figure 35: **Comparison of empirical results with the Garnett approximation**

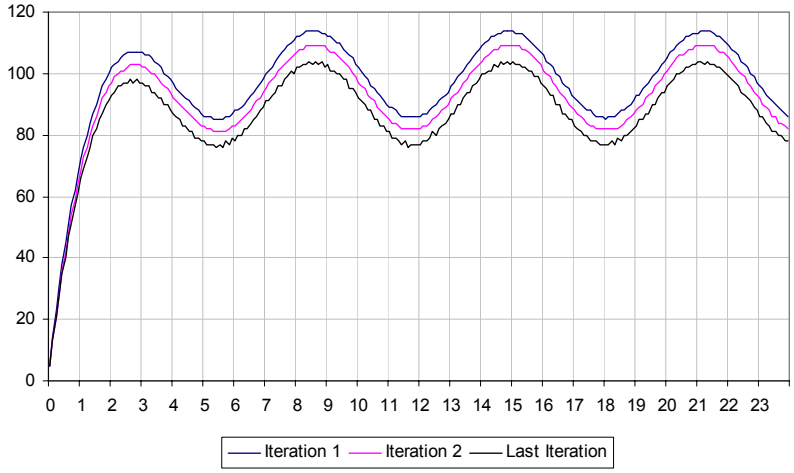



With *monotone dynamics*, when starting with $s_0(t) \equiv \infty$, $s_n(t)$ is monotone decreasing in $n$ for all $t$, i.e. the following prevails:

$$s_n(t) \leq s_m(t) \qquad \text{for all} \quad m < n\,. \tag{5.29}$$

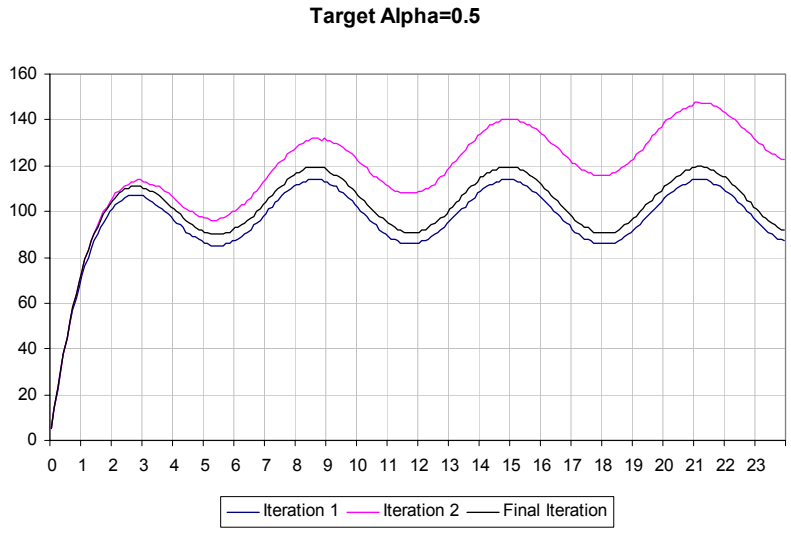
An example of the monotone dynamics is shown in Figure 36, where staffing levels are shown for the first three iterations of the algorithm for the case of arrival function $\lambda(t) = 100 + 20 \cdot \sin(t)$, service times exponential having mean 1, and impatience times that are exponential having mean 0.1 ($\mathbf{r} = 10$).

Figure 36: **Staffing levels in the $1^{st}$, $2^{nd}$ and last iterations. $\mu$=1, $\theta$=10.**



In contrast, with *oscillating dynamics*, $s_n(t)$ is oscillating for all $t$; i.e. there exist 2 subsequences $\{s_k(t)\}_{k=2n}^{\infty}$ and $\{s_l(t)\}_{l=2n+1}^{\infty}$, such that $s_{2n}(t) \downarrow s_\infty(t)$ and $s_{2n+1}(t) \uparrow s_\infty(t)$. Within the oscillating framework, there is monotonicity. An example of the oscillating dynamics can be viewed in Figure 37, where staffing levels are shown for the first three iterations for the same case except there is no abandonment ($\theta = 0$ and $\mathbf{r} = 0$).

Figure 37: **Staffing levels in the $1^{st}$, $2^{nd}$ and last iterations. $\mu$=1, $\theta$=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 $\mathbf{r} > 1$ ($\mathbf{r} < 1$), then the death rates at time $t$ decrease (increase) as $s_t$ increases. Hence, if we disregard statistical error, caused by having to estimate the 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 $s_1(t) \leq s_2(t)$ for all $t$, $0 \leq t \leq T$, and $\mathbf{r} > 1$, then

$$\{L_1(t) : 0 \leq t \leq T\} \leq_{st} \{L_2(t) : 0 \leq t \leq T\}\,, \tag{5.30}$$

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

$$E\left[f\left(\{L_1(t) : 0 \leq t \leq T\}\right)\right] \leq_{st} E\left[f\left(\{L_2(t) : 0 \leq t \leq T\}\right)\right] \tag{5.31}$$

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 (5.30) 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 (5.30).

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: $L_1(t) = L_2(t) = k$. 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

$$L_1(t) \leq_{st} L_2(t) \quad \text{for all} \quad t \ , \tag{5.32}$$

where now $\leq_{st}$ denotes conventional stochastic order for real-valued random variables, just as in Chapter 1 of Ross (1996); also see Müller and Stoyan (2002). We only need the more elementary stochastic order in (5.32), but we use the more sophisticated sample-path stochastic order in (5.30) to get it. The stochastic order is equivalent to the tail probabilities being ordered; i.e., (5.32) is equivalent to $P(L_1(t) > x) \leq P(L_2(t) > x)$ for all $x$, which implies the ordering for the staffing functions at time $t$. In particular, suppose that

$$P\left(L_2(t) \geq s_2(t)\right) \leq \alpha < P\left(L_2(t) \geq s_2(t) - 1\right) \ . \tag{5.33}$$

Since

$$P\left(L_1(t) \geq s_2(t)\right) \leq P\left(L_2(t) \geq s_2(t)\right) \leq \alpha \ , \tag{5.34}$$

necessarily $s_1(t) \leq s_2(t)$.

**Case 1: r $>$ 1.** For $s_0(t) = \infty$, we necessarily start with $s_0(t) > s_1(t)$ for all $t$, which produces first $L_1(t) \leq_{st} L_0(t)$ and then $s_2(t) \leq s_1(t)$ for all $t$. Continuing, we get $L_n(t)$ stochastically decreasing in $n$ and $s_n(t)$ 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 $s_0(t) = \infty$, we again necessarily start with $s_0(t) > s_1(t)$ for all $t$. That produces first $L_1(t) \geq_{st} L_0(t)$ and then $s_0(t) \geq s_2(t) \geq s_1(t)$ for all $t$. Afterwards, we get $L_1(t) \geq_{st} L_2(t) \geq_{st} L_0(t)$ and $s_0(t) \geq s_2(t) \geq s_3(t) \geq s_1(t)$ for all $t$. Continuing, we get $L_{2n}(t)$ stochastically increasing in $n$, while $L_{2n+1}(t)$ stochastically decreases in $n$, for all $t$. Similarly, $s_{2n}(t)$ decreases in $n$, while $s_{2n+1}(t)$ 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** $\alpha$ strongly influenced the dynamics. In particular, higher values of $\alpha$ cause larger oscillations in the oscillating case, and slower convergence to the limit in all cases. This phenomenon is illustrated in Figures 38 and 39. 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 38) and $\alpha = -0.5$ (Figure 39) for the case of arrival function $\lambda(t) = 100 + 20 \cdot \sin(t)$, service times are exponential having mean 1, and no abandonment.

Finally, we also observed a **time-dependent behavior in the convergence** of $s_n(t)$. We observed a greater gap as time increased. For example, let

$$I_t \equiv \inf\left\{j : s_i(t) = s_j(t) \quad \text{for all} \quad i \geq j\right\} . \tag{5.35}$$

We observed that $I_{t_2} \geq I_{t_1}$ for all $t_2 > t_1$. An illustration can be viewed in Figure 40. 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 left end of the time horizon. When it reaches the limit at the left, the gap will still decrease more to the right.

Figure 38: **Range of staffing level for target $\alpha$=0.1**

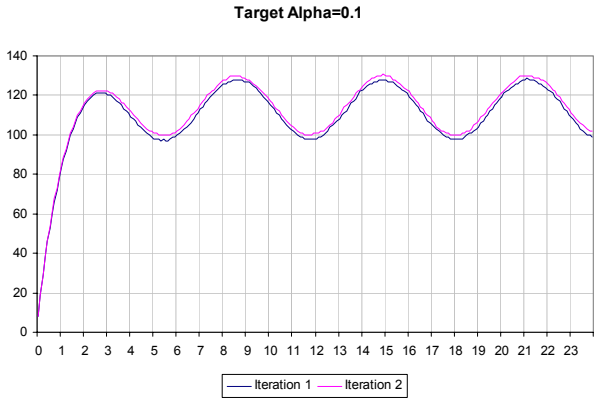



Figure 39: **Range of staffing level for target $\alpha$=0.5**

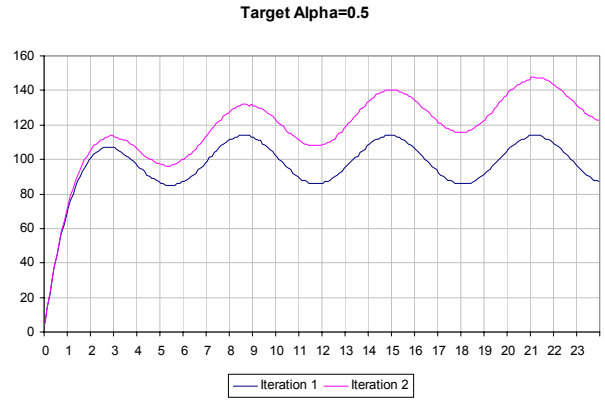
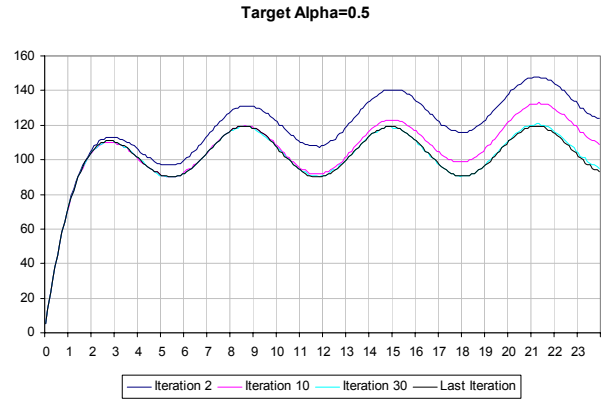



Figure 40: **Evolution of convergence during algorithm run-time**



# 6 Summary and Directions for Future Research

We have developed an algorithm (ISA) that generates staffing functions for which performance is stable in the face of time-varying loads. The results have been found to be remarkably robust, covering the ED, QD and QED operational regimes. Here are some natural "next-steps":

1. As discussed in Section 2.2, it remains to explore alternative staffing methods to achieve better time-stability of abandonment probabilities and expected waiting times, especially under heavy loads.

2. A great advantage of ISA is its generality. However, it remains to explore the ISA for additional queueing systems. We already have 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.

3. We have seen that ISA usually converges quite quickly, but it remains to analyze convergence of the algorithm more thoroughly. We have noted that the monotone and oscillating convergence, displayed in Section 7, can be explained via stochastic-ordering, but that depends strongly on the $M_t/M/s_t + M$ model structure. Even for that model, some of the phenomena have not yet been adequately explained.

4. For one special case in Section 2.2, we have shown that our staffing methods are asymptotically correct as the scale increases. It would be nice to do that much more generally. It is natural to do that within the mathematical framework of service networks, as in Mandelbaum et. al.(1998). We would like to prove much more generally that, under proper scaling, the actual time-dependent probability of delay indeed converges to the specified target as scale increases.

# References:


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

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

Gans, N., Koole, G. and Mandelbaum, A. **Telephone Call Centers: Tutorial, Review and Research Prospects**. Invited review paper by *Manufacturing and Service Operations Management (M&SOM)*, **5**(2), 2003.

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

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

Green, L. V., Kolesar, P. J. and Soares, J. **Improving the SIPP Approach For Staffing Service Systems That Have Cyclic Demand**. *Management Science*. July–August 2001.

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

Jelenkovic P., Mandelbaum A. and Momcilovic P. **Heavy Traffic Limits for Queues with Many Deterministic Servers** *Queueing Systems*, **47**, 53–69, 2004.

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

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

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

Massey, W. A., Parker, G. A. and Whitt, W. **Estimating the Parameters of a Nonhomogeneous Poisson Process with Linear Rate**. *Telecommunication Systems*, **5** 361–388, 1996.

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

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

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

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

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

Wallace, R. B. and Whitt, W. **A Staffing Algorithm for Call Centers with Skill-Based Routing**, submitted for publication, 2004.

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

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.

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

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.

Whitt, W. **Efficiency-Driven Heavy-Traffic Approximations for Many-Server Queues with Abandonments**. *Management Science*, **50** (10) 1449–1461, 2004.

Whitt, W. **Engineering Solution of a Basic Call-Center Model**. *Management Science*, **51**, 2005, to appear.

# 7 Appendix

In this appendix we present some additional material supplementing the theoretical motivation in Section 2.2.

## 7.1 A Uniform-Acceleration Perspective

We can create a rigorous framework for this square root rule by applying the asymptotic analysis of uniform acceleration to multi-server queues with abandonment. The underlying intuition for optimal staffing is that for large systems we staff exactly for the number of customers requesting service so as a first order effect, abandonment simply does not happen. Thus the associated fluid model should not be a function of any abandonment parameters. The effects of abandonment appear as second order phenomena at best and are found in the associated diffusion model. Moreover, we can show that for the special case of $\theta = \mu$, our limiting diffusion gives us exactly the square-root-staffing formula.

Let $\{ L^\eta \mid \eta > 0 \}$ by a family of multi-server queues with abandonment indexed by $\eta$, where $\theta^\eta = \theta$ and $\mu^\eta = \mu$ or the service and abandonment rates are independent of $\eta$, but

$$\lambda_t^\eta = \eta \cdot \lambda_t \quad \text{and} \quad s_t^\eta = \eta \cdot s_t^{(0)} + \sqrt{\eta} \cdot s_t^{(1)} + o(\sqrt{\eta}\,). \tag{7.36}$$

Unlike the uniform acceleration scalings that lead to the pointwise stationary approximation, this one is inspired by the scalings of Halfin and Whitt. Here we are scaling up the arrival rate (representing "demand" for our call center service) and the number of service agents (representing "supply" for our call center service) by the same parameter $\eta$. By limit theorems developed in Mandelbaum, Massey and Reiman, we know that such a family of processes have fluid and diffusion approximations as $\eta \to \infty$. We want to restrict ourselves to a special type of growth behavior for the number of servers.

**Theorem 1** *Consider the family of multiserver queues with abandonment having the growth conditions for its parameters as defined above. If we set*

$$s_t^\eta = \eta \cdot m_t + \sqrt{\eta} \cdot s_t^{(1)} + o(\sqrt{\eta}\,) \tag{7.37}$$

*where*

$$\frac{d}{dt} m_t = \lambda_t - \mu_t \cdot m_t, \tag{7.38}$$

*then*

$$\lim_{\eta \to \infty} \mathsf{P}\left(L^\eta(t) \geq s_t^\eta\right) = \mathsf{P}\left(L^{(1)}(t) \geq s_t^{(1)}\right), \tag{7.39}$$

*where the diffusion $L^{(1)} = \left\{ L^{(1)}(t) \mid t \geq 0 \right\}$ is the unique sample path solution to the integral equation*

$$\begin{aligned} L^{(1)}(t) &= L^{(1)}(0) + \int_0^t (\mu_u - \theta_u) \cdot (s_u^{(1)})^- du \\ &\quad - \int_0^t \left(\theta_u \cdot L^{(1)}(u)^+ - \mu_u \cdot L^{(1)}(u)^-\right) du + B\left(\int_0^t (\lambda_u + \mu_u \cdot m_u) du\right) \end{aligned} \tag{7.40}$$

*and the process $\{ B(t) \mid t \geq 0 \}$ is standard Brownian motion.*

Thus we can reduce the analysis of the probability of delay (approximately) to the analysis of a one-dimensional diffusion $L^{(1)}$. Notice that since $\lambda_t$ and $\mu_t$ are given, then so is $m_t$. Thus server staffing for this model can only be controlled by the selection of $s^{(1)}$. Also notice that the diffusion $L^{(1)}$ is independent of $s^{(1)}$ *as long as* $\theta_t = \mu_t$ or $s_t^{(1)} \geq 0$ for all time $t \geq 0$.

For the special case of $\mu = \theta$ we can give a complete analysis of the delay probabilities that gives the server staffing heuristic of Jennings, Mandelbaum, Massey and Whitt.

**Corollary 2** *If $\theta = \mu$ and $s_t^\eta = \eta \cdot m_t + \Phi^{-1}(1-\alpha) \cdot \sqrt{\eta \cdot m_t}$, where*

$$\frac{1}{\sqrt{2\pi}} \int_{\Phi^{-1}(1-\alpha)}^{\infty} e^{-x^2/2} dx = \alpha, \tag{7.41}$$

*then we have*

$$\lim_{\eta \to \infty} \mathsf{P}\left(L^\eta(t) \geq s_t^\eta\right) = \alpha \tag{7.42}$$

*for all $t > 0$.*

Unfortunately, $L^{(1)}$ in general is *not* a Gaussian process. This also means that the following set of differential equations are not autonomous.

**Corollary 3** *The differential equation for the mean of $L^{(1)}$ is*

$$\frac{d}{dt}\mathsf{E}\left[L^{(1)}(t)\right] = (\mu_t - \theta_t) \cdot (s_t^{(1)})^- - \theta_t \cdot \mathsf{E}\left[L^{(1)}(t)^+\right] + \mu_t \cdot \mathsf{E}\left[L^{(1)}(t)^-\right]. \tag{7.43}$$

*Since $L^{(1)}(t)^+ \cdot L^{(1)}(t)^- = 0$, the differential equation for the variance of $L^{(1)}$ equals*

$$\begin{aligned} \frac{d}{dt}\mathsf{Var}\left[L^{(1)}(t)\right] &= -2\theta_t \cdot \mathsf{Var}\left[L^{(1)}(t)^+\right] - 2\mu_t \cdot \mathsf{Var}\left[L^{(1)}(t)^-\right] \\ &\quad -2(\theta_t + \mu_t) \cdot \mathsf{E}\left[L^{(1)}(t)^+\right] \cdot \mathsf{E}\left[L^{(1)}(t)^-\right] + \lambda_t + \mu_t \cdot m_t. \end{aligned} \tag{7.44}$$

**Proof of Theorem 1:** Define the function $f_t^\eta(\cdot)$, where

$$f_t^\eta(x) = \eta \cdot \lambda_t - \theta_t \cdot (\eta \cdot x - s_t^\eta)^+ - \mu_t \cdot (\eta \cdot x \wedge s_t^\eta). \tag{7.45}$$

Now we have

$$\begin{aligned} f_t^\eta(x) &= \eta \cdot \lambda_t - \theta_t \cdot (\eta x - s_t^\eta)^+ - \mu_t \cdot ((\eta x) \wedge s_t^\eta) \\ &= \eta \cdot \lambda_t - \eta \cdot \theta_t \cdot x + (\theta_t - \mu_t) \cdot ((\eta \cdot x) \wedge s_t^\eta). \end{aligned}$$

However

$$\begin{aligned} (\eta \cdot x) \wedge s_t^\eta &= (\eta \cdot x) \wedge \left(\eta \cdot m_t + \sqrt{\eta} \cdot s_t^{(1)} + o(\sqrt{\eta})\right) \\ &= 1_{\{x < m_t\}} \cdot (\eta \cdot x + o(\sqrt{\eta})) + 1_{\{x = m_t\}} \cdot (\eta \cdot m_t - \sqrt{\eta} \cdot (s_t^{(1)})^- + o(\sqrt{\eta})) \\ &\quad + 1_{\{x > m_t\}} \cdot (\eta \cdot m_t - \sqrt{\eta} \cdot s_t^{(1)} + o(\sqrt{\eta})) \\ &= \eta \cdot (x \wedge m_t) + \sqrt{\eta} \cdot \left((s_t^{(1)})^+ 1_{\{x > m_t\}} - (s_t^{(1)})^- 1_{\{x \geq m_t\}}\right) + o(\sqrt{\eta}) \end{aligned}$$

combining these results gives us the asymptotic expansion

$$\begin{aligned} f_t^\eta(x) &= \eta \cdot \left(\lambda_t - \theta_t \cdot (x - m_t)^+ - \mu_t \cdot (x \wedge m_t)\right) \\ &\quad + \sqrt{\eta} \cdot (\theta_t - \mu_t) \left((s_t^{(1)})^+ \cdot 1_{\{x > m_t\}} - (s_t^{(1)})^- \cdot 1_{\{x \geq m_t\}}\right) + o(\sqrt{\eta}) \end{aligned}$$

as $\eta \to \infty$.

It follows that $f_t^\eta = \eta \cdot f_t^{(0)} + \sqrt{\eta} \cdot f_t^{(1)} + o(\sqrt{\eta})$, where

$$f_t^{(0)}(x) = \lambda_t - \theta_t \cdot (x - m_t)^+ - \mu_t \cdot (x \wedge m_t) \tag{7.46}$$

and

$$f_t^{(1)}(x) = (\theta_t - \mu_t) \cdot \left( (s_t^{(1)})^+ \cdot 1_{\{x > m_t\}} - (s_t^{(1)})^- \cdot 1_{\{x \geq m_t\}} \right). \tag{7.47}$$

Now

$$\Lambda f_t^{(0)}(x; y) = (\theta_t - \mu_t) \cdot \left( y \cdot 1_{\{x < m_t\}} - y^- \cdot 1_{\{x = m_t\}} \right) - \theta_t \cdot y, \tag{7.48}$$

hence we have

$$\Lambda f_t^{(0)}(m_t; y) = \mu_t \cdot y^- - \theta_t \cdot y^+ \text{ and } f_t^{(1)}(m_t) = (\mu_t - \theta_t)(s_t^{(1)})^- \tag{7.49}$$

where $\Lambda g(x; y) = g'(x+)y^+ - g'(x-)y^-$ is the *non-smooth derivative* of any function $g$ that has left and right derivatives.

Finally, we have

$$\begin{aligned} L^{(1)}(t) &= L^{(1)}(0) + \int_0^t \left( \Lambda f_t^{(0)} \left( m_u; L^{(1)}(u) \right) + f_t^{(1)}(m_u) \right) du & (7.50) \\ &\quad + B \left( \int_0^t (\lambda_u + \mu_u \cdot m_u) du \right) \\ &= L^{(1)}(0) - \int_0^t \left( \theta_u \cdot (L^{(1)}(u)^+ + (s_u^{(1)})^-) - \mu_u \cdot (L^{(1)}(u)^- + (s_u^{(1)})^-) \right) du & (7.51) \\ &\quad + B \left( \int_0^t (\lambda_u + \mu_u \cdot m_u) du \right). \end{aligned}$$

## 7.2 Case 1: $\theta_t = \mu_t$

We then have

$$L^{(1)}(t) = L^{(1)}(0) - \int_0^t \mu_u \cdot L^{(1)}(t) du + B \left( \int_0^t (\lambda_u + \mu_u \cdot m_u) du \right). \tag{7.52}$$

It follows that $L^{(1)}$ is a zero mean Gaussian process (if $L^{(1)}(0) = 0$) and

$$\frac{d}{dt} \mathsf{Var} \left[ L^{(1)}(t) \right] = -2\mu_t \cdot \mathsf{Var} \left[ L^{(1)}(t) \right] + \lambda_t + \mu_t \cdot m_t. \tag{7.53}$$

Moreover, if $m_0 = \mathsf{Var} \left[ L^{(1)}(0) \right]$, then $\mathsf{Var} \left[ L^{(1)}(t) \right] = m_t$ for all $t \geq 0$.

## 7.3 Case 2: $\theta_t = 0$

We then have

$$L^{(1)}(t) = L^{(1)}(0) + \int_0^t \mu_u \cdot \left( L^{(1)}(u)^- + (s_u^{(1)})^- \right) du + B \left( \int_0^t (\lambda_u + \mu_u \cdot m_u) du \right). \tag{7.54}$$

with

$$\frac{d}{dt} \mathsf{E} \left[ L^{(1)}(t) \right] = \mu_t \cdot \left( \mathsf{E} \left[ L^{(1)}(t)^- \right] + (s_t^{(1)})^- \right) \tag{7.55}$$

and

$$\frac{d}{dt} \mathsf{Var} \left[ L^{(1)}(t) \right] = -2\mu_t \cdot \left( \mathsf{Var} \left[ L^{(1)}(t)^- \right] + \mathsf{E} \left[ L^{(1)}(t)^+ \right] \cdot \mathsf{E} \left[ L^{(1)}(t)^- \right] \right) + \lambda_t + \mu_t \cdot m_t. \tag{7.56}$$