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

UNABRIDGED VERSION: INTERNET SUPPLEMENT

by

Z. Feldman

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

A. Mandelbaum

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

W.A. Massey

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

W. Whitt

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

May 15, 2005

Abstract

This is a longer version of a paper with the same title, which has been submitted to a journal. Much of the work consists of simulation experiments; this longer version presents more examples; e.g., there are 44 figures here, but only 15 in the journal version. This longer version also provides additional theoretical support.

Abstract from the Journal Version

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

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

Contents

1. Introduction

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

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

1.1. The Staffing Problem

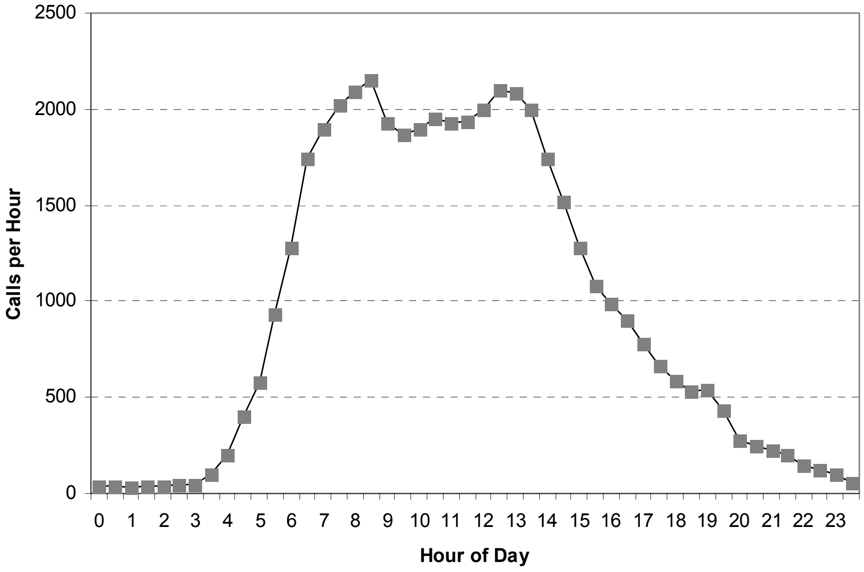
The staffing problem typically takes the following form: Under an existing operational reality, and given a desired quality of service, we seek the least number of agents at each time that is required to meet a given service-level constraint. This problem, which has received much attention over the years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

Figure 1 depicts a typical arrival-rate function to a telephone call center. Call volumes are low around midnight (hour 0), starting to increase in the early hours of the morning, peaking at late morning, then dropping somewhat around midday (12, lunch break), rising again afterwards, and then dropping thereafter to midnight levels. The displayed arrival-rate function is an average of several similar days; the actual number of arrivals, in a given hour on a given day, fluctuates randomly around this average. (The functional form in Figure 1 is typical; the particular values for the arrival rates were adapted from Green, Kolesar and Soares (2001).)

Staffing planners are thus faced with two sources of variability: **predictable variability** – time-variations of the expected load – and **stochastic variability** – random fluctuations around this time-dependent average. Most available staffing algorithms are designed to cope only with stochastic variability; they avoid the predictable variability in various ways. For example, when the service times are relatively short, 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 the leading term in the more sophisticated *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

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



not so short, significant predictable variability can cause PSA to produce poor performance. As a consequence, some parts of the day may be overstaffed, while others are understaffed.

In this paper we address the staffing problem with *both* predictable and stochastic variability. Here is the problem we aim to solve: **Given a daily performance goal, and faced with both predictable and stochastic variability, we seek to find the minimal staffing levels that meet this performance goal stably over the day.**

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

1.2. Organization of this Paper

More than half this paper repeats a shorter journal version. We have added subsections and a Table of Contents to better communicate the organization.

The material in §§2-6 mostly repeats what is in the shorter main paper; there are only a few additions. We start in §2 by briefly reviewing the previous contributions by Jennings et al. (1996). We then overview our main contributions in this paper in §3. In §4 we specify our iterative-staffing algorithm in detail. In §5 we illustrate the performance of our algorithm by considering an Erlang-A-model example (with abandonment). In §6, for comparison, we consider a similar Erlang-C-model example (without abandonment).

We present some additional examples, not included in the main paper, in §§7-9. In §7 we revisit the "challenging example" in Jennings et al. (1996), now applying our iterative-staffing algorithm. In §8 we expand the analysis of the Erlang-A example from §5 by considering different patience parameters. In §5 we let the abandonment rate equal the service rate, which is often realistic in practice (approximately). However, in §8 we let the abandonment rate be 5 times and 0.2 times the service rate, representing the cases of very impatient customers and very patient customers, respectively. We show that the staffing methods continue to perform well in those alternative cases.

Then, in §9 we analyze a realistic example, related to Figure 1, including short service times (in particular 6 minutes). In contrast to Green et. al. (2001), which is the source of Figure 1, we incorporate abandonment, which significantly impacts staffing results. When abandonment is present, as it often is in practice, it is possible to achieve good performance with significantly fewer agents than when abandonment is present. We show that conservative rules of thumb without abandonment tend to overstaff substantially.

In §§10 and 11 we return to material in the journal version. In §10 we present some supporting theory. That mostly repeats what is in Section 7 of the journal version, but the final subsection does not appear there. In §11, we discuss the dynamics of the iterative algorithm, establishing monotonicity and convergence results. That too mostly repeats what is in the main paper (§8), but includes some additional figures and discussion.

The remaining material is not in the main paper: In §12, we explain and define the performance measures used in our simulations. In §13 we provide additional insight into the square-root-staffing formula from the perspective of many-server heavy-traffic limits, using the Markovian-service-network framework from Mandelbaum, Massey and Reiman (1998).

Finally, in §14 we summarize our main contributions and discuss directions for future research.

2. Our Point of Departure

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

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

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

where α is the target delay probability.

In (2.1) above, we choose a constant target delay probability α for all times t. To achieve that for time varying arrival rate $\lambda(t)$, we aim to find an appropriate staffing function s_t. This problem is challenging because the time-dependent delay probability $P(L_t \geq s_t)$ in (2.1) depends on the staffing function before time t as well as at time t.

2.1. An Infinite-Server Approximation

In Jennings et al. we proposed an **infinite-server approximation**. In particular, we proposed approximating the random variable L_t by the number L_t^∞ of busy servers in the associated $M_t/G/\infty$ model, having the same arrival process and service times. We thus choose the desired staffing function s_t so that the inequalities in (2.1) hold when L_t^∞ is substituted for L_t. That approximation provides great simplification because (i) the tail probability $P(L_t^\infty \geq s_t)$ at time t depends on the staffing function $\{s_t : t \geq 0\}$ only through its value at the single time t and (ii) the exact time-dependent distribution of L_t^∞ is known.

The first simplification follows from the fact that the distribution of the stochastic process $\{L_t^\infty : t \geq 0\}$ is totally independent of the "staffing function" $\{s_t : t \geq 0\}$. Thus, the distribution of the individual random variable L_t^∞ is independent of $\{s_t : t \geq 0\}$ too. When we calculate $P(L_t^\infty \geq s_t)$, s_t just serves as the argument of the tail-probability function.

The second simplification stems from basic properties of time-varying infinite-server queues. In particular, as reviewed in Eick et al. (1993a), for each t, L_t^∞ has a **Poisson distribution** whenever the number in the system at time $t = 0$ has a Poisson distribution. (Being empty is a degenerate case of a Poisson distribution.) That Poisson distribution is fully characterized by its mean m_t, which depends on t. We next apply a normal approximation for the Poisson distribution, using the fact that the variance equals the mean for a Poisson distribution. We obtain the **normal approximation**

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

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

2.2. The Square-Root-Staffing Formula

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

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

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

(2.1) and the normal approximation in (2.2), we see that the quality-of-service parameter β in (2.4) should be chosen so that

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

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

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

where S_e is a random variable with the associated **stationary-excess cdf** (or equilibrium-residual-lifetime cdf) G_e associated with the service-time cdf G, defined by

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

with k^{th} moment

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

see Theorem 1 of Eick et al. (1993a) and references therein. For more on the stationary-excess cdf G_e, see pp. 424 and 431 of Ross (2003); $G = G_e$ if and only if G is exponential. The different expressions in (2.5) provide useful insight; see Eick et al. (1993a) and Section 4.2 of Green et al. (2005).

Moreover, the time-dependent mean has a convenient approximation, based on a second-order Taylor-series approximation for λ about t. In particular, the time-dependent mean can be approximated in terms of the first two moments of S_e, the mean of S and the second derivative of the arrival-rate function at t, $\lambda^{(2)}(t)$, by

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

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

2.3. The Modified-Offered-Load Approximation as a Refinement

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

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

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

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

Since the infinite-server approximation suggests shifting from the PSA offered load $\lambda(t)E[S]$ to the "infinite-server" offered load m_t, it is useful to **quantify the difference** between these quantities. For that purpose, the Taylor-series approximation in (2.8) is very useful. It says that the infinite-server offered load m_t coincides with the PSA offered load $\lambda(t)E[S]$, approximately, except for a deterministic **time shift** by $E[S_e]$ and a deterministic **space shift** by $(\lambda^{(2)}(t)/2)Var(S_e)E[S]$. Since $\lambda^{(2)}(t)$ will be negative at a peak, we see that the actual requirements at times of peak demand are less than predicted by PSA. The mapping of the arrival-rate function into the infinite-server mean m_t acts as a smoothing operator, making the results less extreme. Of course, that is convenient for meeting practical constraints on staffing schedules (tours of duty).

We can go further: 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 $\lambda(t) \cdot E[S]$ in another way. Letting $(S_e)_e$ be a random variable with the twofold stationary-excess cdf $(G_e)_e$, we have the formula

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

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

When we applied the MOL approximation in §4 of Jennings et al., we did not apply the MOL approximation directly. Instead of calculating the steady-state delay probability for the stationary Erlang-C model, we exploited an approximation for the delay probability based on a many-server heavy-traffic limit in Halfin and Whitt (1981). That produces a simple formula relating the delay probability α and the service quality β. Moreover, the heavy-traffic limit provides an alternative derivation of the square-root staffing formula in (2.3), without relying on an infinite-server approximation or a normal approximation.

However, the limit in Halfin and Whitt (1981) was restricted to the stationary model, exploiting the MOL approximation. See Mandelbaum, Massey and Reiman (1998) and §13 here for corresponding limits directly in the framework of time-varying models.

2.4. It is possible to Achieve Time-Stable Performance!

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

The most important conclusion from those previous experiments is that it is indeed possible to achieve time-stable performance for the $M_t/M/s_t$ model by an appropriate choice of a staffing function s_t, even in the face of a strongly time-varying arrival-rate function.

3. Our Contributions Here

We develop staffing algorithms for more complicated time-varying many-server models, such as many-server queues with abandonment. For example, we treat the much more realistic $M_t/G/s+G$ model with non-exponential service times (the first G) and non-exponential abandonments (the $+G$). We emphasize that models with customer abandonment were not considered by Jennings et al. or anybody else.

For call centers, our ultimate goal is to treat realistic multi-server systems with multiple call types and skill-based routing (SBR), but we do not pursue that here. In that setting, it is natural to apply SBR methods for stationary models after using the MOL approximation in (2.9) for each call type at time t. Once we have reduced the problem to a stationary SBR model, we may be able to apply the staffing method in Wallace and Whitt (2004). Approaches based on these ideas remain to be investigated, however.

3.1. A Simulation-Based Iterative Staffing Algorithm (ISA)

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 $\{s_t : 0 \leq t \leq T\}$.

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; e.g., see Ingolfsson (2005).)

In practice, staffing is required to be fixed over longer staffing intervals - typically ranging from 15 minutes to an hour. Here we ignore that constraint. An initial staffing function with such constraints is obtained from our results by using in each staffing interval the maximum required staffing level at any time point within that staffing interval. That will yield an upper bound on the required staffing. Simulation can then be used, in the manner of the ISA, to see if these initial staffing levels can be decreased, while still meeting the performance target at every time. See Green et al. (2005) for additional discussion on this point, and references studies of the impact of staffing intervals.

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 and the queue-length tail delay-probabilities, turn out to be relatively constant over time as well.

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

While working with ISA, we discovered that the simulation-based solutions have astonishing regularity. In particular, we found that global performance measures coincide with the performance measures of the associated stationary model. In particular, when we used ISA to staff the time-varying $M_t/M/s_t+M$ model, we found that the staffing could be related to the steady-state behavior of the associated stationary $M/M/s+M$ model. That implies that the modified-offered-load approximation will work well for the $M_t/M/s_t+M$ model.

3.2. An Extension of the Square-Root-Staffing Formula

That leads us to our second contribution: We extend the **square-root staffing formula** based on the **modified-offered-load approximation** in Jennings et al. to the $M_t/M/s_t{+}M$ model. In particular, we suggest staffing according to the square-root-staffing formula in (2.3), where the service quality $\beta \equiv \beta(\alpha)$ is derived from a theoretical **one-to-one relation between α and β 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 α in the corresponding stationary model. For that purpose, we could calculate the steady-state probability of delay in the associated stationary $m/M/s+M$ model, which is routine because the number in system L_t is a birth-and-death stochastic process. We can also calculate all other performance measures in the $M/M/s+M$ model; e.g., see Garnett et al. (2002) and Whitt (2005).

However, instead of calculating the exact steady-state delay probability in the stationary model, we propose using an approximation for the steady-state delay probability - a simple formula based on a heavy-traffic limit - just as Jennings et al. applied the many-server heavy-traffic limit from Halfin and Whitt (1981). For the $M_t/M/s_t + M$ model, we use explicit formulas relating α to β 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.

3.3. Staffing at the Offered Load

Finally, we make yet one more contribution. To describe it, we remind readers of the three heavy-traffic regimes for many-server queues: *Quality-Driven* (QD, lightly loaded), *Efficiency-Driven* (ED, heavily loaded) and *Quality-and-Efficiency-Driven* (QED, normally loaded); see Garnett et al. (2002). In our experiments for the many-server queue with abandonments we found that **simply staffing according to the offered load itself** is remarkably effective in the QED regime, i.e., staffing by letting $s_t = m_t$ for the $M_t/M/s_t + M$ model works very well in the QED regime. Needless to say, abandonments play a crucial role in this property. This is another example of the importance of including abandonments in the model, when customers actually do abandon; see Garnett et al. (2002) for more discussion. See Section 13 for additional theoretical support, based on the Markovian-service-network framework of Mandelbaum, Massey and Reiman (1998).

3.4. The Naive Deterministic Approximation

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

(2.5).

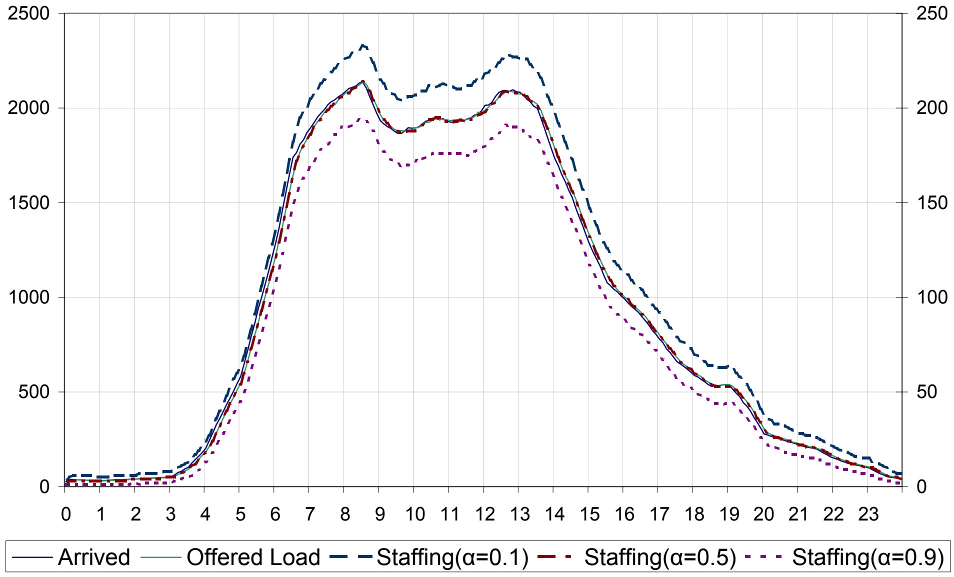
We conclude by mentioning that the naive deterministic approximation is remarkably effective in the setting of the realistic large example in Figure 1, when there is customer abandonment in the QED regime. We show that right now by plotting the staffing levels in the QD ($\alpha = 0.1$), QED ($\alpha = 0.5$), ED ($\alpha = 0.9$) regimes in Figure 2. For this realistic example, we assume that the service times are short; in particular, the mean service time is 6 minutes.

From Figure 2, we see that in the QED regime, with $\alpha = 0.5$, the staffing falls right on top of the offered load, which in turn agrees with the PSA offered load (or arrival rate with the right time units). For more discussion about this realistic example, see Section 9.

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

In this section we describe the simulation-based interactive-staffing algorithm (ISA). As indicated before, we determine time-dependent staffing levels aiming to achieve a given constant probability of delay at all times. In the process of applying the ISA, we directly confirm that

Figure 2: **Staffing for the medium-size call center in Figure 1 with short service times (6 minutes) in three regimes: QD ($\alpha = 0.1$), QED ($\alpha = 0.5$) and QED ($\alpha = 0.9$).**



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 relatively stable as well.

4.1. The Simulation Framework

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 $\{\lambda(t); 0 \leq t \leq T\}$. (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^{th} iteration, for $t \in [0, T]$ (the given time horizon). Although our algorithm is time-continuous, we make staffing changes only at discrete times. That is achieved by dividing the time-horizon into small intervals of length Δ. In all experiments presented in this paper, we use $\Delta = 0.1/\mu$, where $1/\mu$ is the mean service time. We then let the number of servers be constant within each of these intervals.

For any specified staffing function, the system simulation can be performed in a conventional manner. 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 λ^* exceeding the maximum of the arrival-rate function $\lambda(t)$ for all t, $0 \leq t \leq T$. 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 $s_t^{(n)}$ be the staffing level at time t in iteration n for $0 \leq t \leq T$. Let $L_t^{(n)}$ denote the random total number of customers in the system at time t, under this staffing function. We estimate the distribution of $L_t^{(n)}$ for each n and t by performing multiple (5000) independent replications. We think of starting off with infinitely many servers. Since this is a simulation, we choose a large finite number, ensuring that the probability of delay (i.e., of having all servers busy upon arrival) is negligible for all t. For the examples in §5 and §6, it suffices to let $s_t^{(0)} = 200$ for all t.

4.2. The Algorithm

The algorithm iteratively performs the following steps, until convergence is obtained. Here, convergence means that the staffing levels do not change much after an iteration. (Practically, they are allowed to change by some threshold τ, which we take to be 1.)

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

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

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

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

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

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

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

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 §11. 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 $\mathbf{r} \equiv \theta/\mu$, where θ is the individual abandonment rate. If $\mathbf{r} = 1$, corresponding to an infinite-server queue (§10), 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 §11.

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

We demonstrate the performance of ISA by considering a time-varying Erlang-A model ($M_t/M/s_t+M$) with a special structured arrival-rate function.

5.1. A Sinusoidal Arrival-Rate Function

We consider a sinusoidal arrival-rate function. In particular, let the queueing system be faced with a non-homogeneous Poisson arrival process with a **sinusoidal arrival-rate function**

$$\lambda(t) = a + b \cdot \sin(ct), \quad 0 \le t \le T \ , \tag{5.1}$$

where $a = 100$, $b = 20$ and $c = 1$. Let the service times and the customer times to abandon (if they have not yet started service) come from independent sequences of independent and identically distributed (IID) exponential random variables, both having mean 1. As can be seen from PSA, the arrival rate is sufficiently large, that about 100 servers are required, so this example captures the many-server spirit of a call center. However, the sinusoidal form of the arrival-rate function is clearly a mathematical abstraction, which has the essential property of producing significant fluctuations over time, i.e., significant predictable variability. This particular arrival-rate function is by no means critical for our analysis; our methods apply to arbitrary arrival-rate functions, such as in Figure 1. (Indeed, for that, see Section 9.

An important issue, however, is the rate of fluctuation in the arrival-rate function compared to the expected service-time distribution. To be concrete, we will measure time in hours, and focus on a 24-hour day, so that $T = 24$. A cycle of the sinusoidal arrival-rate function in (5.1) is $2\pi/c$; since we have set $c = 1$, a cycle is $2\pi \approx 6.3$ hours. Thus there will be about 4 cycles during the day. That roughly matches the daily cycle in Figure 1 for the six-hour period around 12:00 noon.

Since we let the mean service time be 1 and have chosen to measure time in hours, the mean service time in this example is 1 hour. That clearly is relatively long for most call centers, where the interactions are short telephone calls. If we were to change the time units in order to rectify that, making the expected service time 10 minutes, then a cycle of the arrival-rate function would become about 1 hour, making for more rapid fluctuations in the arrival rate than are normally encountered in call centers. Thus our example is more challenging than usually encountered in call centers, but may be approached in evolving contact centers if many interactions do indeed take an hour or more. (We consider a practical example directly related to Figure 1 in the Internet Supplement.) From this preliminary analysis, we should anticipate that the service times are sufficiently long in our example that the traditional PSA method is likely to perform poorly here, just as in Jennings et al. (1996), and it does. As before, we are deliberately choosing a difficult case.

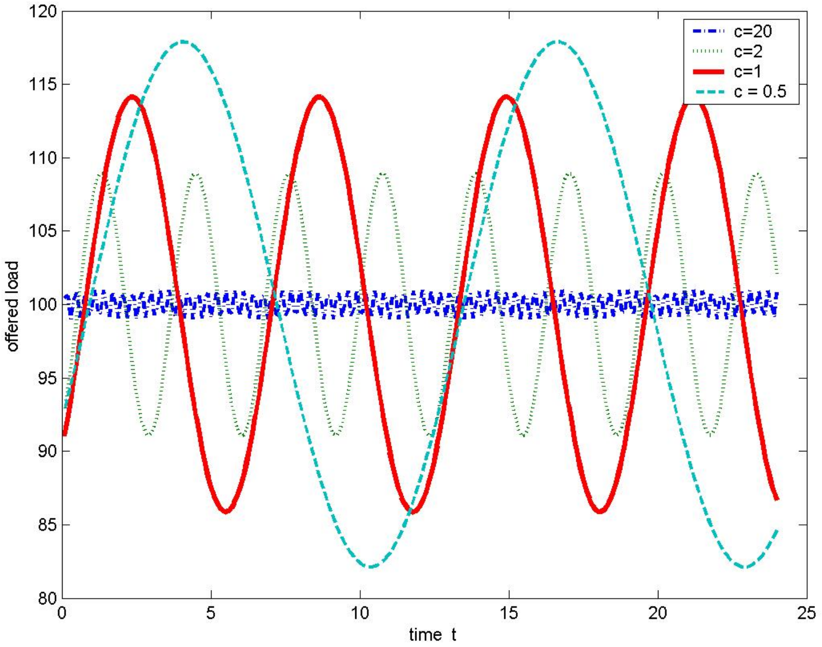
The arrival rate coincides with the PSA offered load, because the mean service time here is 1. The (infinite-server) offered load is given in (2.5). Since we have a sinusoidal arrival-rate function, we can apply Eick et al. (1993b) to give **an explicit formula for the offered-load** m_t, i.e., the mean number of busy servers in the associated infinite-server system. Since the service-time distribution is exponential, we can apply formula (15) of Eick et al. (1993b). For the sinusoidal arrival-rate function in (5.1), the offered load is

$$m_t = a + \frac{b}{1+c^2}[\sin(ct) - c \cdot \cos(ct)] = 100 + 10[\sin(t) - \cos(t)] \ . \qquad (5.2)$$

The second formula in (5.2) is based on the specific parameters: $a = 100$, $b = 20$ and $c = 1$.

In order to put our model into perspective, in Figure 3 we plot the offered load m_t in (5.2) for the sinusoidal arrival-rate function in (5.1) for the parameters $a = 100$ and $b = 20$, as in our example, but with four different values of the time-scaling parameter c: 0.5, 1, 2 and 20. The offered load coincides with 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. 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.

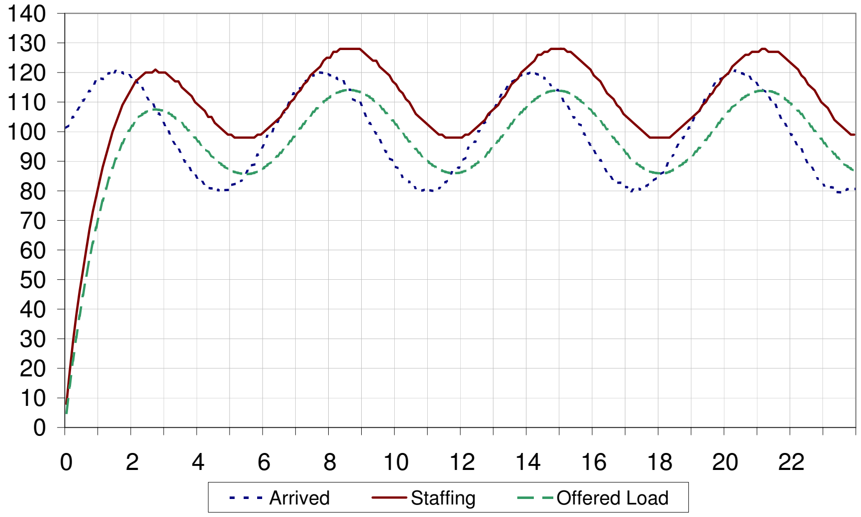
Figure 3: **The offered load m_t for the sinusoidal arrival-rate function in (5.1) with parameters $a = 100$, $b = 20$ and four possible values of c: 0.5, 1, 2 and 20.**

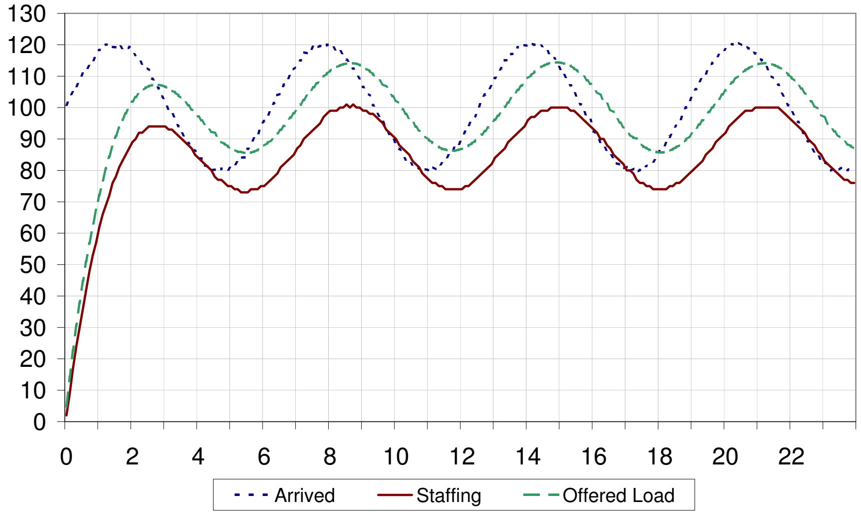


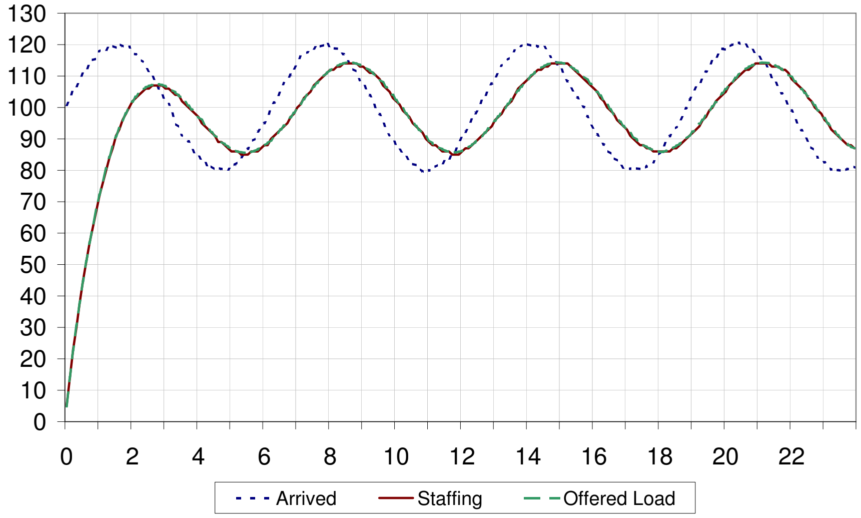
5.2. Application of the ISA

Our simulation-based iterated-staffing algorithm ISA generates staffing functions, for any given target delay probability α. In Figure 5.2 we present three graphs, showing the generated staffing functions for three regimes of operation: *Quality-Driven* (QD) - target $\alpha = 0.1$, *Efficiency-Driven* (ED) - target $\alpha = 0.9$, and *Quality-and-Efficiency-Driven* (QED) - target $\alpha = 0.5$. In each graph, we plot three curves: the arrival rate $\lambda(t)$ (dotted), the offered load m_t (dashed) and the staffing function s_t (solid).

Figure 4: **Staffing for the time-varying Erlang-A example: (1) $\alpha = 0.1$ (QD), (2) $\alpha = 0.9$ (ED), (3) $\alpha = 0.5$ (QED)**







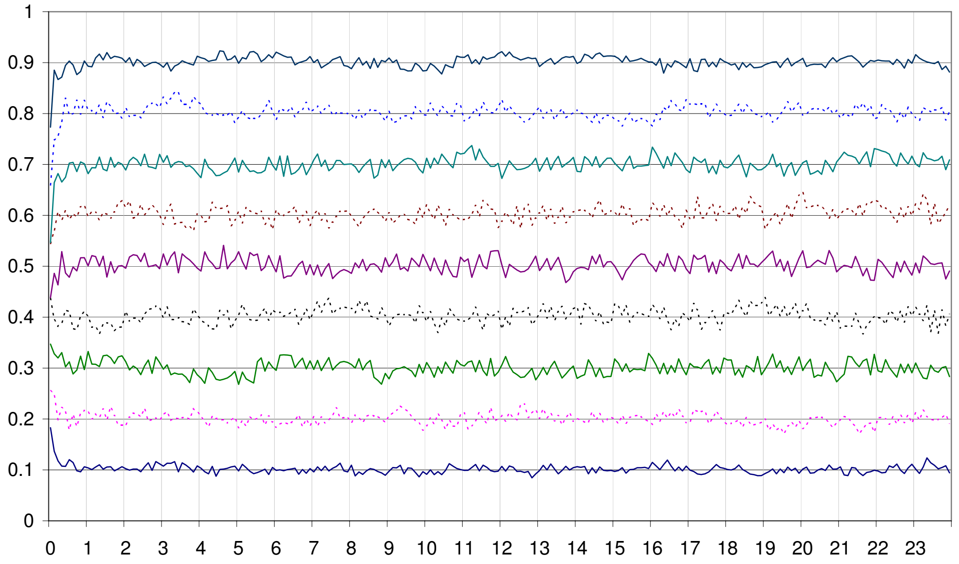
Note that we start our system empty. This allows us to observe the behavior of the transient stage. In particular, there is a ramp-up at the left side of the plot. Our methods respond appropriately to that ramp-up. That is consistent with Section 7 of Jennings et al. (1996).

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 $s_t = m_t$. The ISA-staffing s_t fell on top of the offered load m_t in the QED regime, in particular when $\alpha = 0.5$, in all our experiments. That itself is quite stunning.

5.3. Time-Stable Performance

We now show that ISA achieves time-stable performance. In Figure 5 we show the actual probability of delay obtained by applying our algorithm with target α for $\alpha = 0.1, 0.2, \ldots, 0.9$. These delay probabilities are estimated by performing multiple (5000) independent replications with the final staffing function determined by our algorithm. Under the staffing levels produced by our algorithm, **the delay probabilities are remarkably accurate and stable**; the observed delay probabilities fluctuate around the target in each case.

Figure 5: **Delay probability summary for the time-varying Erlang-A example with various α's.**



In addition to stabilizing the delay probability, other performance measures (e.g. utilization and tail probabilities) are found to be quite stable as well. Precise explanations and definitions of the performance measures are given in Section 12. In Figures 6 are 7 are summary results graphs for all 9 target α's. These two performance measures increase as α increases, so we see the 9 cases starting with $\alpha = 0.1$ at the bottom and increasing to the case $\alpha = 0.9$ at the top.

However, as the target delay probability increases toward heavy loading, the abandonment probability becomes much less time-stable, as shown in Figure 8. We discuss this phenomenon further in §10 below.

Other measures of congestion such as average waiting time and average queue length were also found to be relatively stable, but not perfectly so; e.g., see Figure 9.

Figure 6: **Utilization summary for the time-varying Erlang-A example**

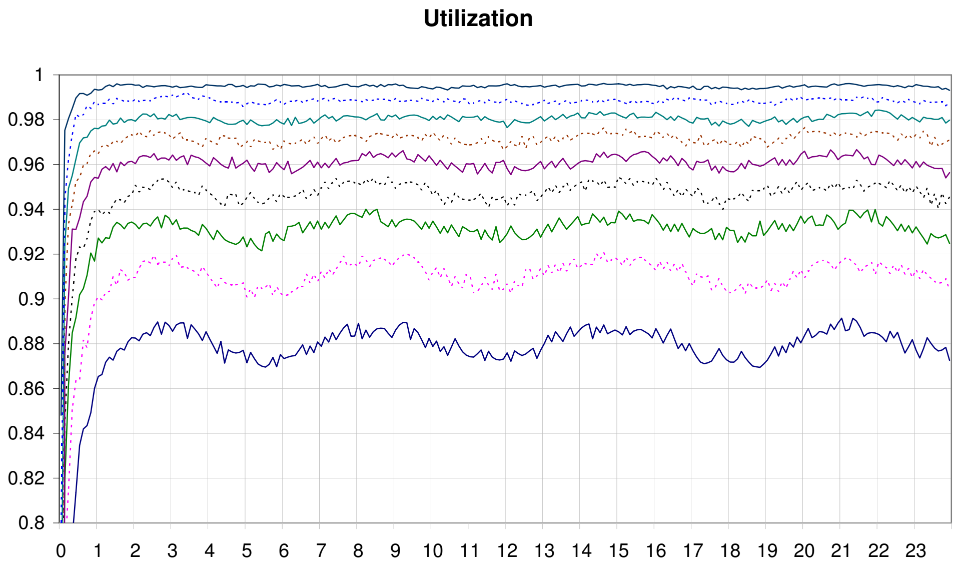


Figure 7: **Tail probability summary for the time-varying Erlang-A example**

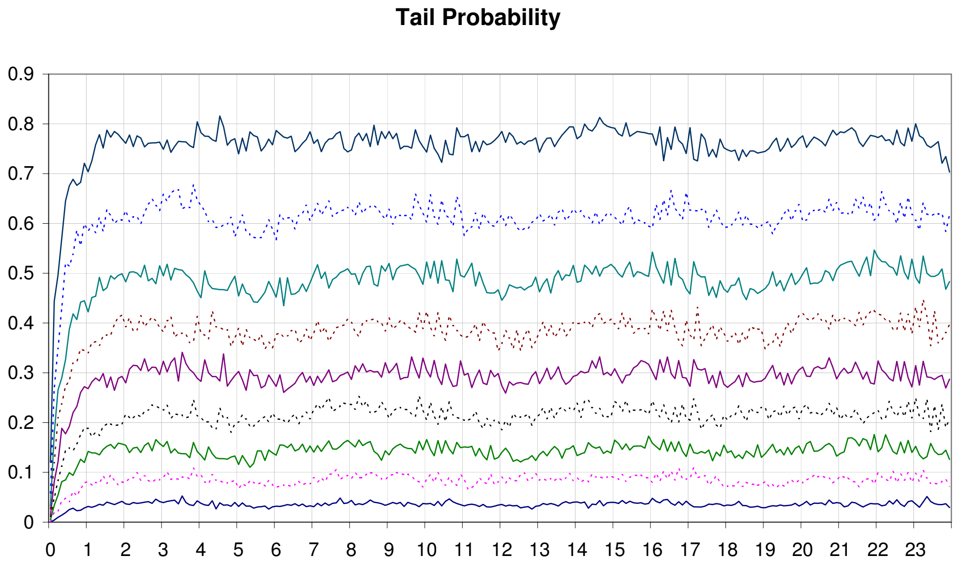


Figure 8: **Abandon probability summary for the time-varying Erlang-A example**

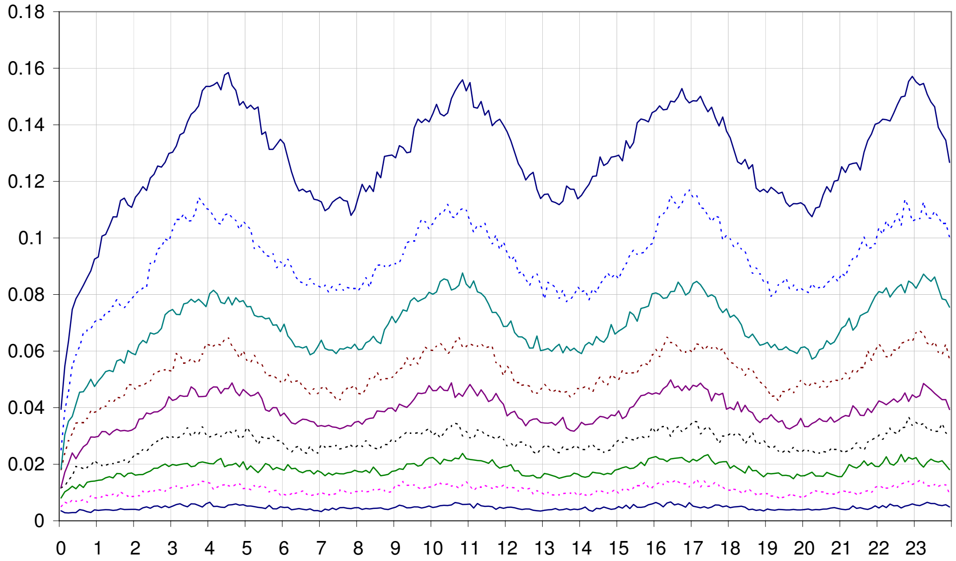
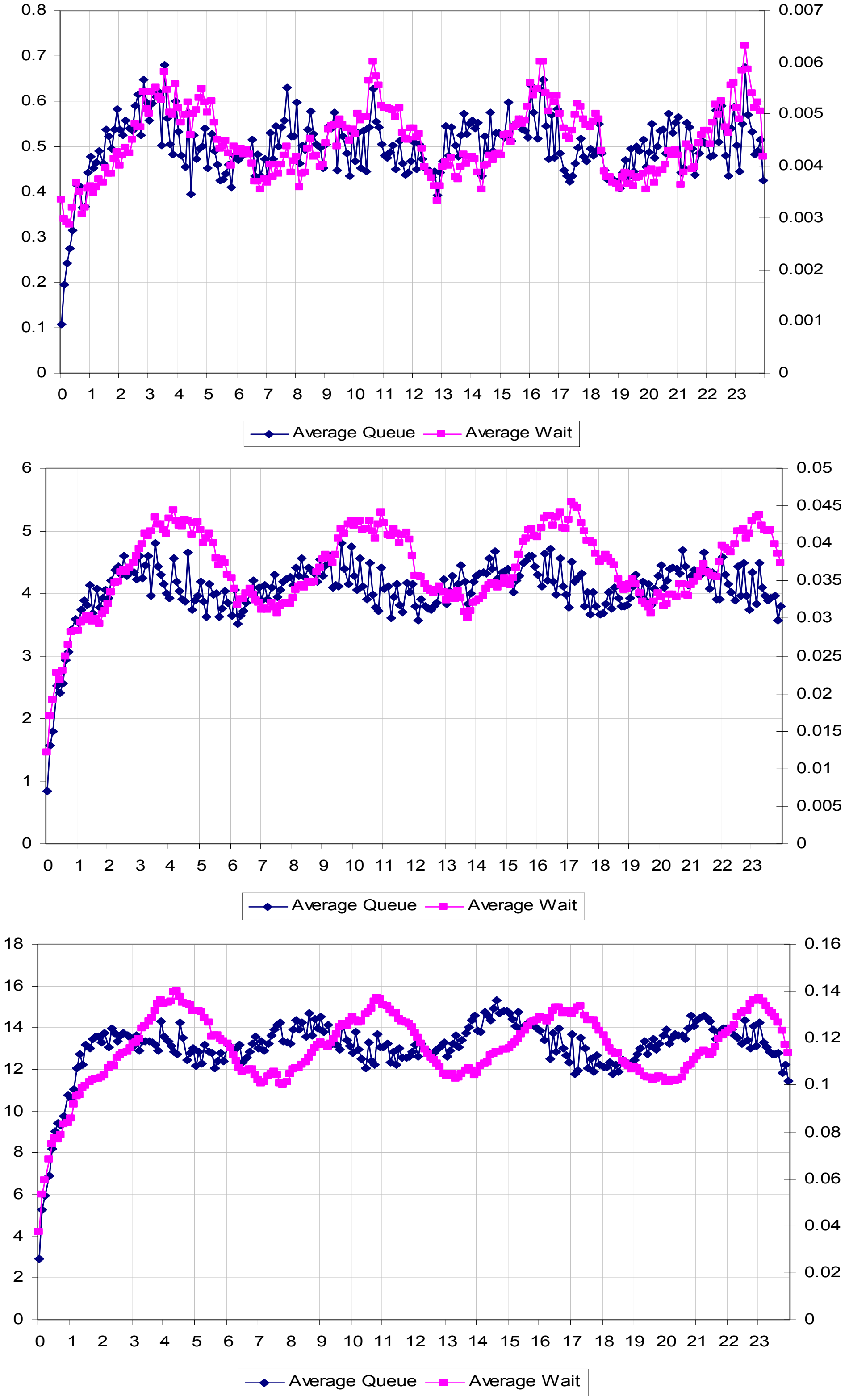


Figure 9: **Congestion measures in the time-varying Erlang-A example in the three regimes: (1) $\alpha = 0.1$ (QD), (2) $\alpha = 0.5$ (QED), (3) $\alpha = 0.9$ (ED)**



5.4. Validating the Square-Root-Staffing Formula

We now validate the square-root-staffing formula in (2.3). For that purpose, we define an **implied empirical service quality**: A function $\{\beta_t : 0 \le t \le T\}$ is defined by setting

$$\beta_t \equiv \frac{s_t - m_t}{\sqrt{m_t}}, \quad 0 \le t \le T \ , \tag{5.3}$$

where m_t is again the offered load in (2.5) and (5.2). and s_t is the staffing function obtained by the ISA algorithm. Since s_t is obtained from the ISA algorithm, the function β_t is itself obtained from the ISA algorithm. It thus becomes interesting to see if **the implied service quality is approximately constant as a function of time**, because that would imply that (5.3) approximately coincides with the square-root-staffing formula (2.3). And, indeed, it is, as shown in Figure 10. Again we consider 9 values of α ranging from 0.1 to 0.9 in steps of 0.1. As α increases, the quality of service reflected by β_t decreases. But the main point is that the empirical service quality β_t as a function of t is approximately constant as a function of t for each α over the full range from 0.1 to 0.9.

Figure 10: **Summary of the implied service quality β for the time-varying Erlang-A example.** (The implied service quality decreases as α increases through the values 0.1, 0.2, . . . , 0.9.)

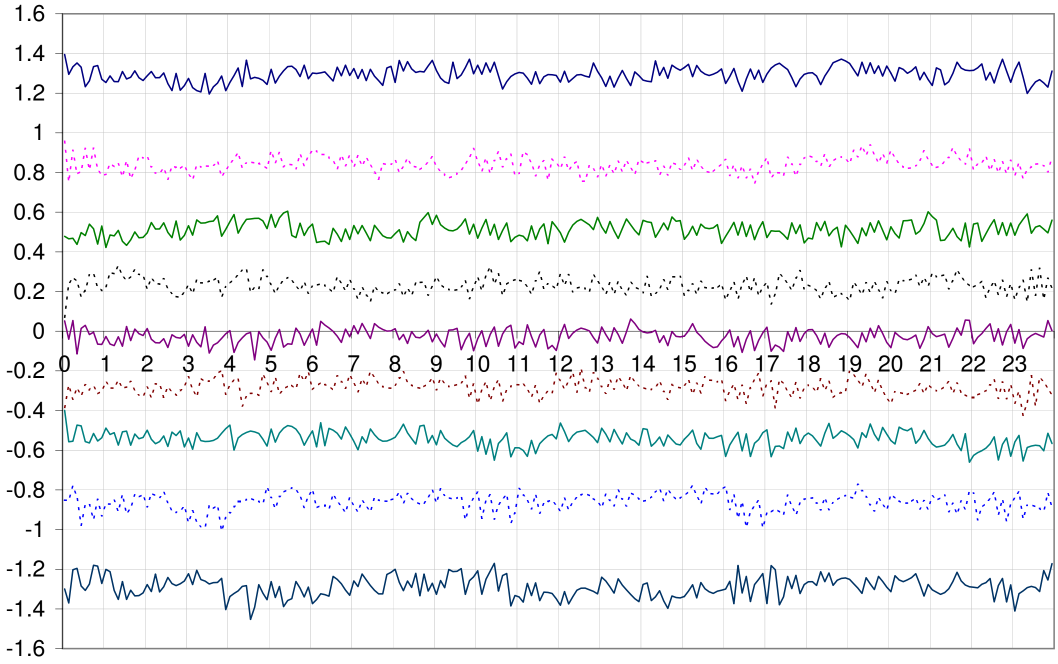


Figure 10 is extremely important because it validates the square-root-staffing formula for this example. First, Figure 5 shows that ISA is able to produce the target delay probability α for a wide range of α. Then Figure 10 shows that, when this is done, the square-root-staffing formula holds empirically. In other words, we have shown that we could have staffed directly by the square-root-staffing formula instead of by the ISA.

5.5. Relating β to α

However, one issues remains: In order to staff directly by the square-root staffing formula, *we need to be able to relate the quality of service β to the target delay probability α.* Indeed, we want a function mapping α into β. We propose a simple answer: For the time-varying Erlang-A model, **use the associated stationary Erlang-A model**, i.e., the $M/M/s+M$ model. That is tantamount to applying the **modified-offered-load approximation** to the $M/M/s+M$ model. Previously the MOL approximation has been applied only to the pure-loss and pure-delay models (without customer abandonments).

Moreover, we suggest using simple formulas obtained from the **many-server heavy-traffic limit** for the Erlang-A model in Garnett et al. (2002). The **Garnett-Mandelbaum-Reiman function, for brevity here referred to as the Garnett function** mapping β into α is

$$\alpha = \left[1 + \sqrt{\frac{\theta}{\mu}} \cdot \frac{h(\hat{\beta})}{h(-\beta)}\right]^{-1}, \qquad -\infty < \beta < \infty; \tag{5.4}$$

where $\hat{\beta} = \beta\sqrt{\theta/\mu}$, with μ the individual service rate and θ the individual abandonment rate (both here set equal to 1 now) and $h(x) = \phi(x)/(1 - \Phi(x))$ is the *hazard rate* of the standard normal distribution, with ϕ being the *probability density function* (pdf) and Φ the cdf. Of course, we want a function mapping α into β. 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 α and β. 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 11 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 11 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 next related the empirical (α, β) pairs to the Garnett function in (5.4). 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 5 and 10. In Figure 12, we plot the pairs of $(\bar{\alpha_i}, \bar{\beta_i})$ alongside the Garnett function. Needless to say, the agreement is phenomenal!

Figure 11: **The empirical conditional waiting time distribution, given positive wait, for the time-varying Erlang-A example with three delay-probability targets: (1) $\alpha = 0.1$ (QD), (2) $\alpha = 0.5$ (QED), (3) $\alpha = 0.9$ (ED)**

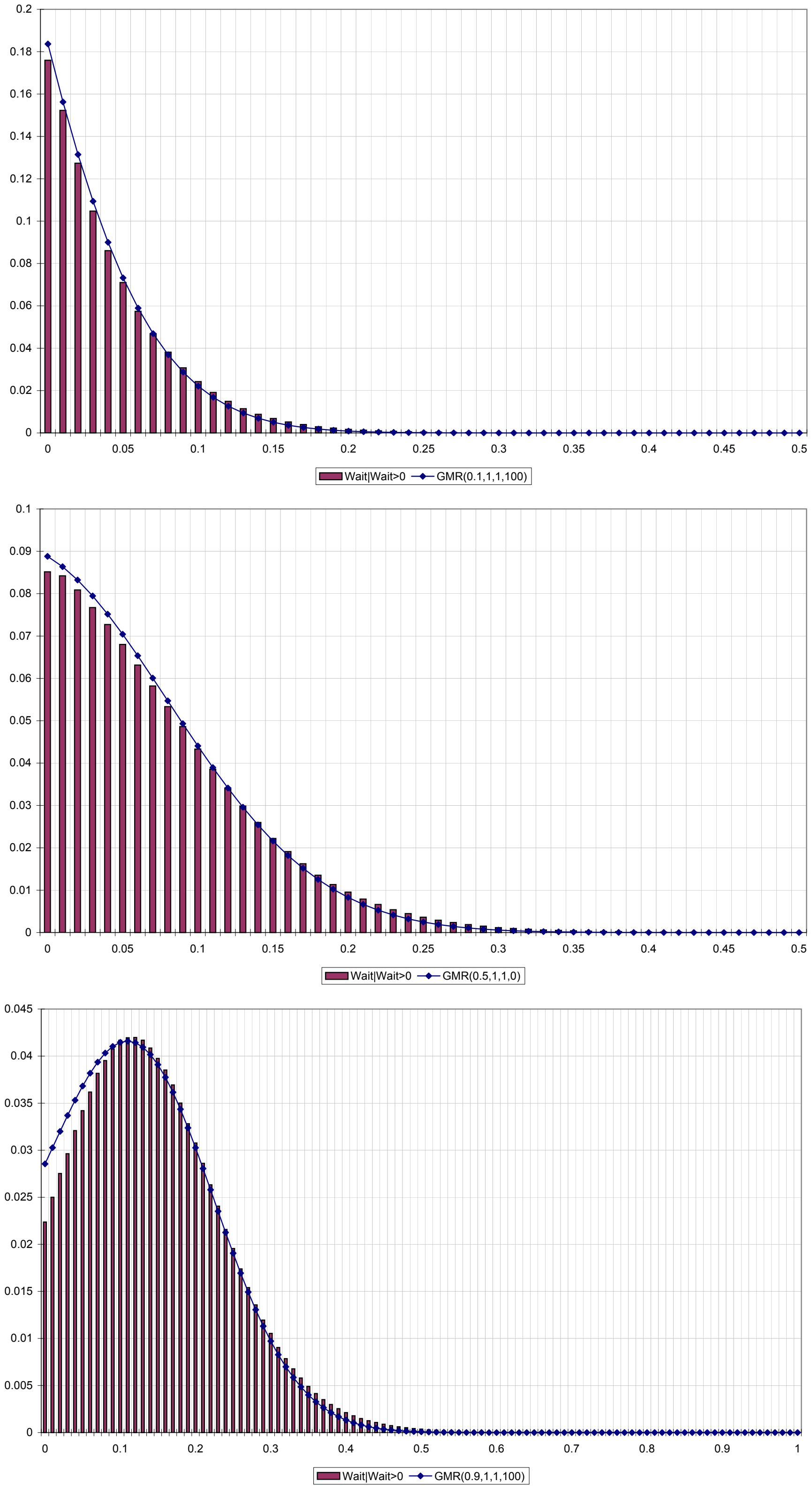
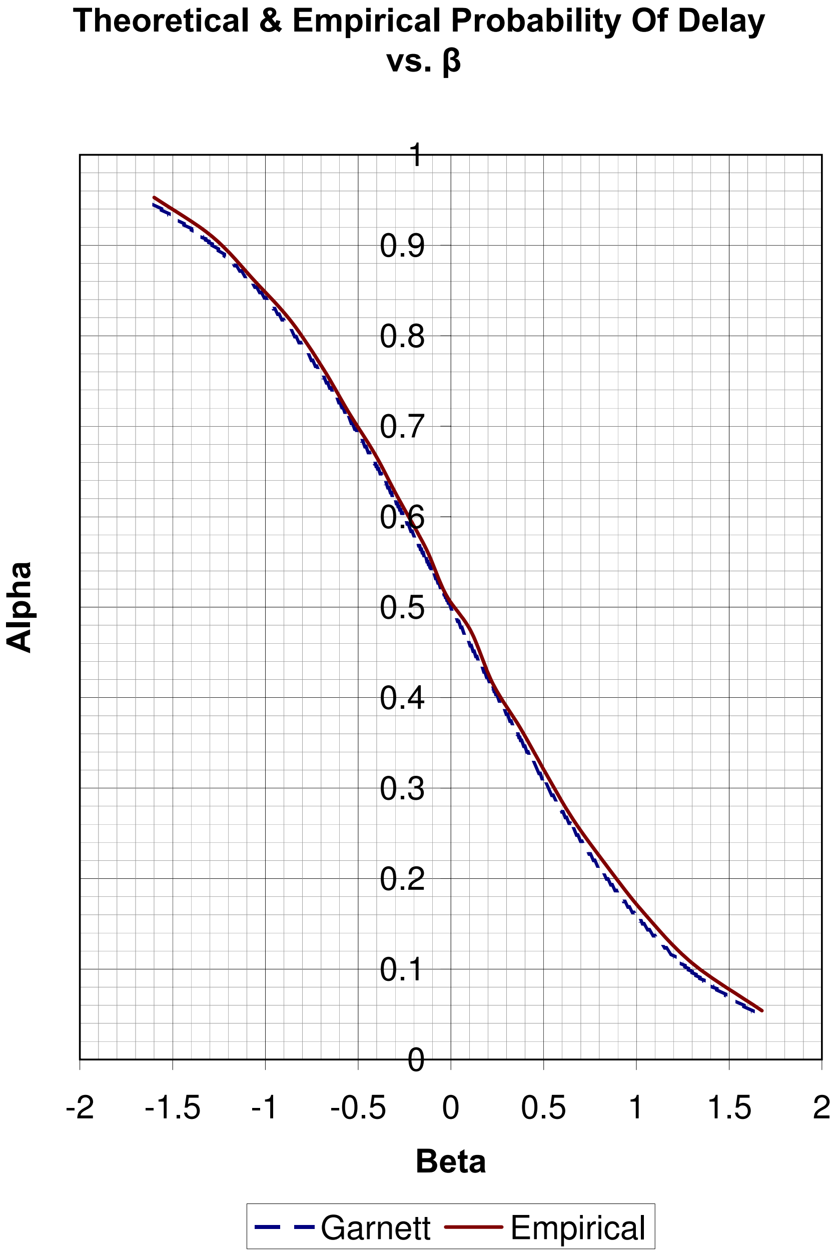


Figure 12: **A comparison of the empirical relation between α and β with the Garnett function for the time-varying Erlang-A example**

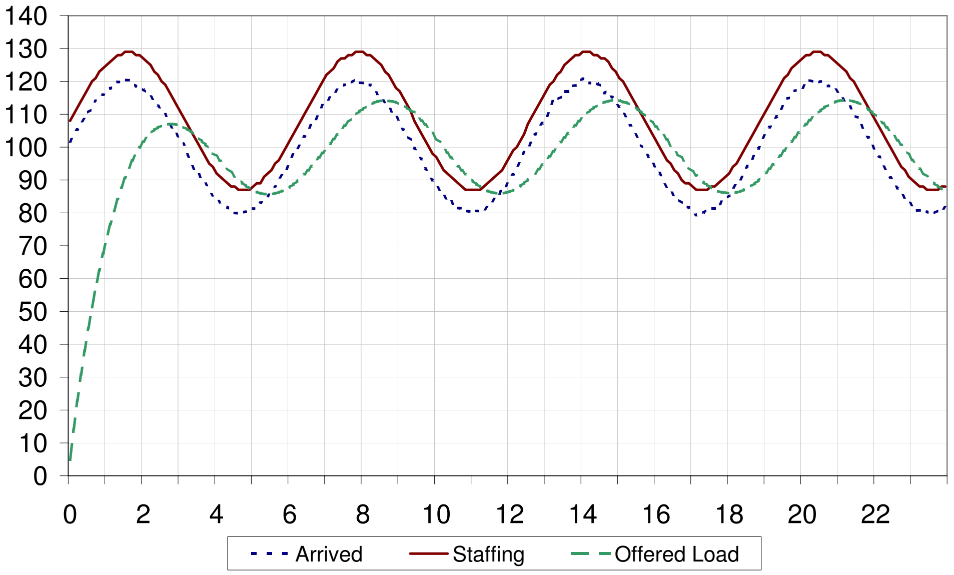


5.6. Comparison to PSA and SSA

Jennings et al. observed that their new infinite-server approximation and modified-offered-load approximation both performed much better than classical alternatives, namely, the pointwise stationary approximation (PSA) and the simple stationary approximation (SSA); SSA uses the stationary model with the overall long-run average arrival rate.

We now show that the same is true here for the time-varying Erlang-A model First we plot the arrival rates and staffing levels for PSA and SSA in Figure 5.6. Then we plot the delay probabilities and the mean queue lengths and waiting times for the two methods in Figures 5.6 and 5.6. The specific target delay probability here is $\alpha = 0.2$. That can be confirmed by looking at SSA. The arrival rate $\lambda(t)$ has average 100. For $s = 109$, the steady-state delay probability in the $M/M/s + M$ model with $\mu = \theta = 1$ is 0.196; while for $s = 108$, the steady-state delay probability is 0.225. Incidentally, the steady-state abandonment probabilities in those two cases are 0.0104 and 0.0124, respectively. In contrast, in the nonstationary environment we see that the time-dependent delay probablity is as much as 0.75 From the customer's perspective, the unreliability may be as bad as the high congestion itself.

Figure 13: **Staffing levels for (1) the pointwise-stationary approximation (PSA) and (2) the simple-stationary approximation (SSA) for the time-varying Erlang-A example**



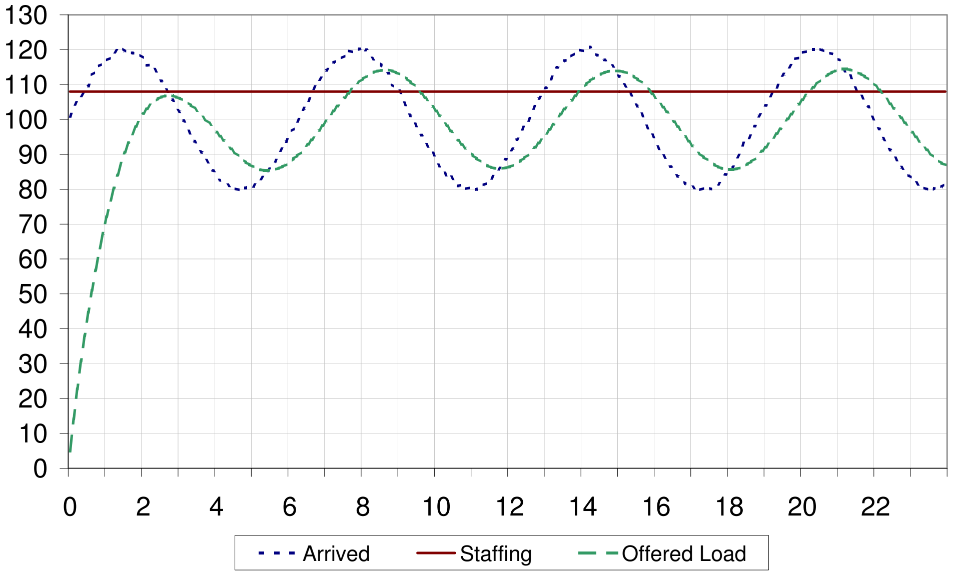
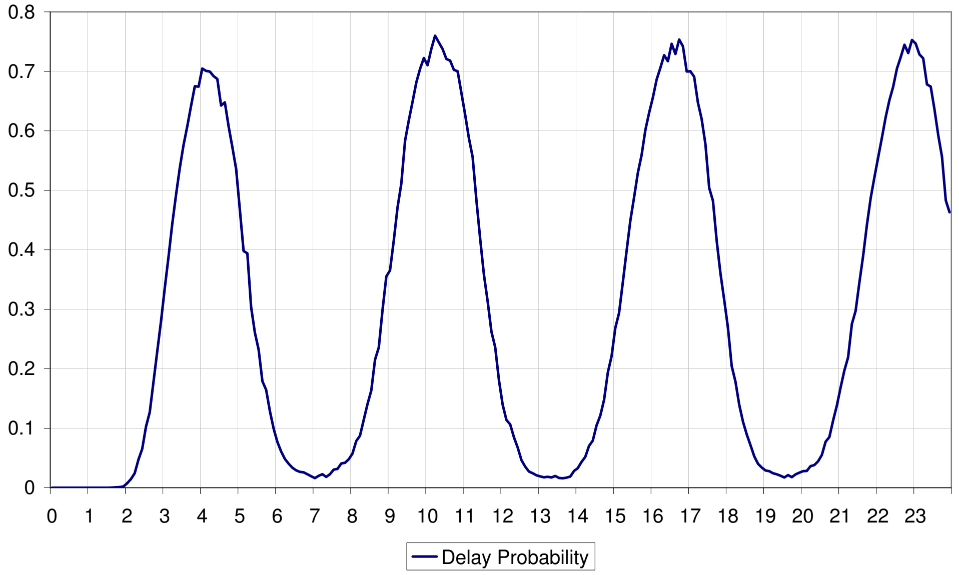


Figure 14: **Delay probabilities for (1) the pointwise-stationary approximation (PSA) and (2) the simple-stationary approximation (SSA) for the time-varying Erlang-A example**



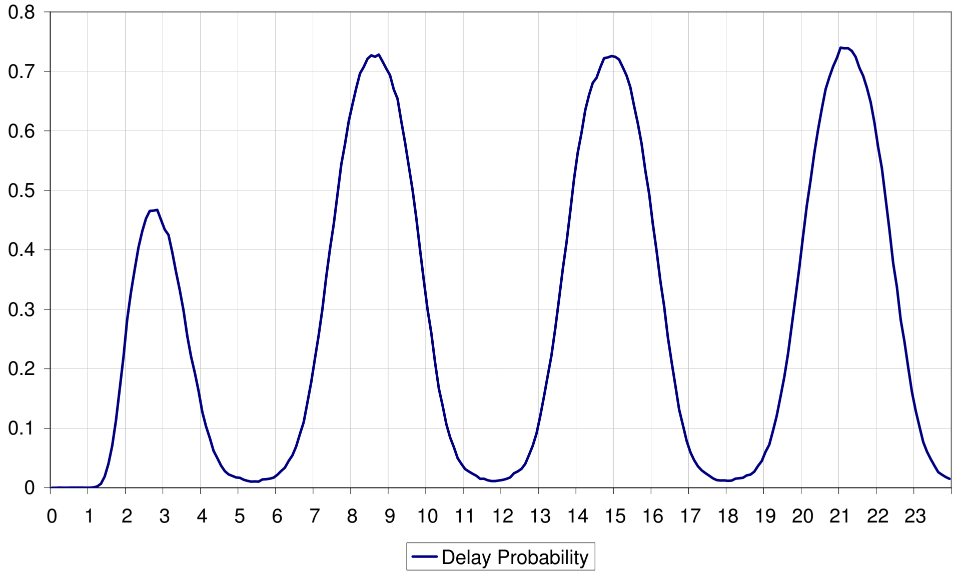
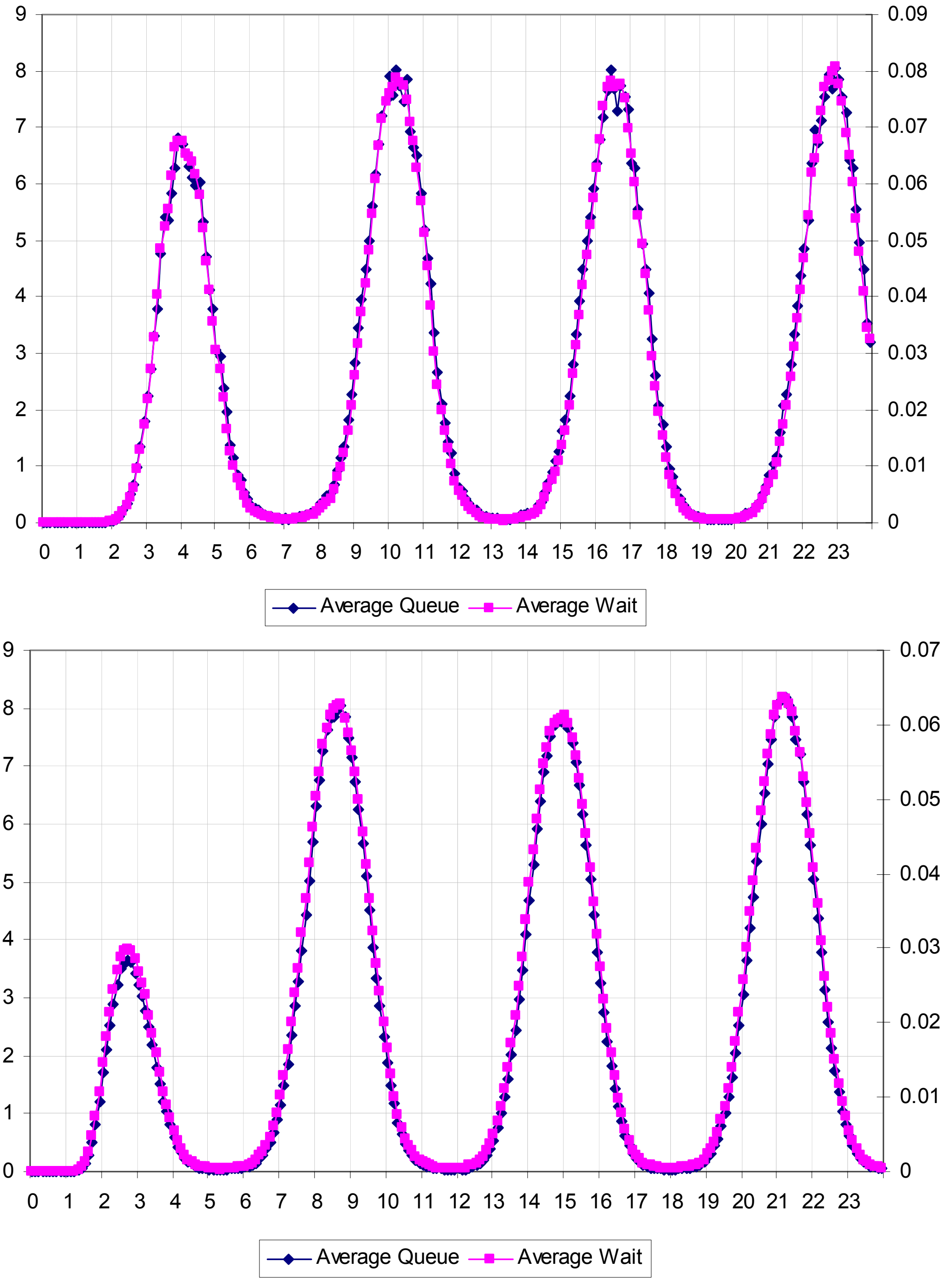


Figure 15: **Mean queue lengths and mean waiting times for (1) the pointwise-stationary approximation (PSA) and (2) the simple-stationary approximation (SSA) for the time-varying Erlang-A example**

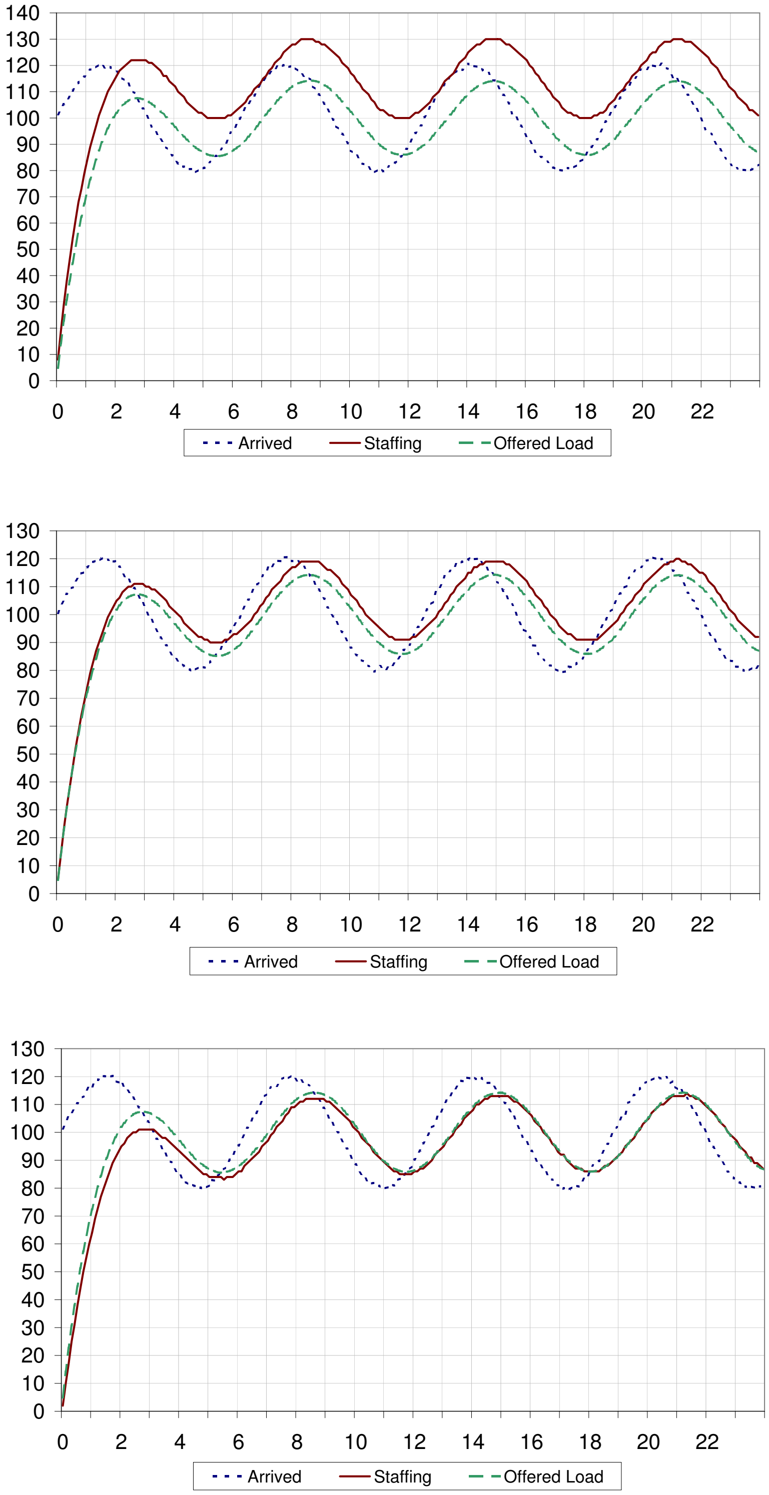


6. The Time-Varying Erlang-C Model

For comparison, we now show the performance of ISA for the same system described in §5 only without abandonment (with infinite patience) - the $M_t/M/s+t$ or Erlang-C model. As expected, the required staffing levels are higher than with abandonment, for all target delay probabilities; compare Figure 16 with Figure 5.2 in Section 5. For example, for $\alpha = 0.5$, the maximum staffing level becomes about 120 instead of 115.

For both the Erlang-A and Erlang-C models, the ISA staffing level decreases as the target delay-probability increases (as the performance requirement becomes less stringent) However, the staffing tends to coincide with the offered load in the Erlang-C example only in the ED regime, when $\alpha = 0.9$, as opposed to in the QED regime, when $\alpha + 0.5$. That illustrates how abandonment allows greater efficiency, while still meeting the delay-probability target.

Figure 16: **The final staffing function found by ISA for the time-varying Erlang-C example with three different delay-probability targets: (1) $\alpha = 0.1$ (QD), (2) $\alpha = 0.5$ (QED), (3) $\alpha = 0.9$ (ED)**



6.1. Time-Stable Performance

As before, we achieve accurate time-stable delay probabilities when we apply the ISA; see Figure 17, where again we consider target delay probabilities $0.1, 0.2, \ldots, 0.9$. The empirical service quality β_t is stabilizing as well, as can be seen from Figure 18, which shows results for the same 9 target delay probabilities. As in Figure 10, the empirical service quality decreases as the target delay probability increases.

However, the empirical service quality β_t stabilizes at a much slower rate, especially for lower values of β (larger values of α). (The approach to steady-state is known to be slower in heavy traffic.) Nevertheless, the steady-state values can be seen at the right in Figure 18.

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). Corresponding plots for other performance measures appear in Figures 19 - 21 and 22. As stated in Section 5, precise explanations and definitions of the performance measures are given in Section 12.

Figure 17: **Delay probability summary for the Erlang-C example**

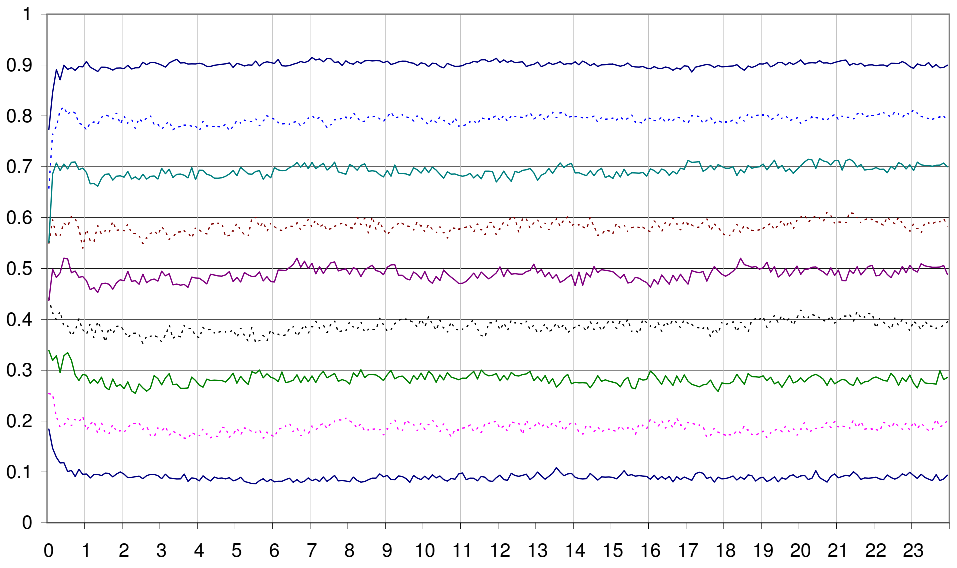


Figure **??** shows that here the time until system reaches (dynamic) steady-state is much longer compared to a system with abandonment. In fact, steady-state was not yet reached after 24 time-units.

Figure 18: **Implied service quality β summary for the Erlang-C example** (The implied service quality decreases as α increases through the values 0.1, 0.2, . . . , 0.9.)

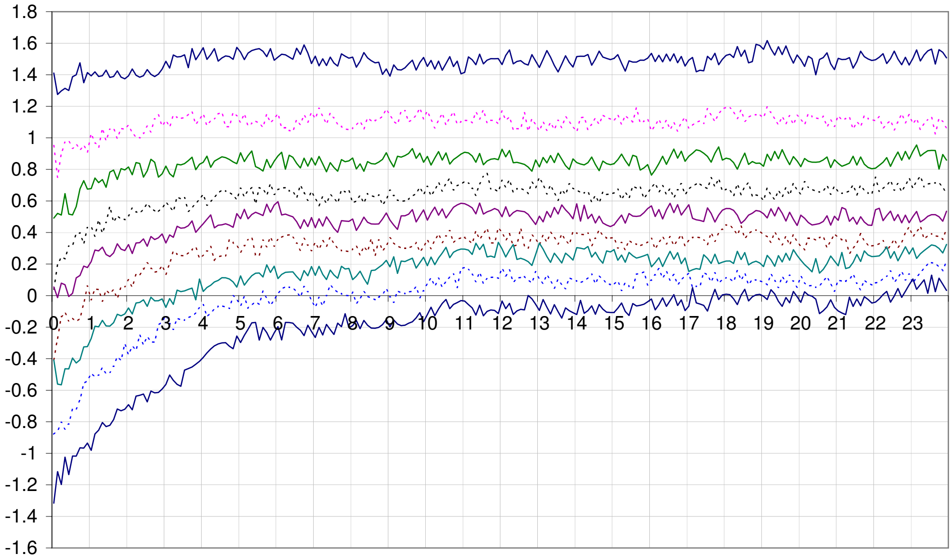


Figure 19: **Utilization summary for the Erlang-C example**

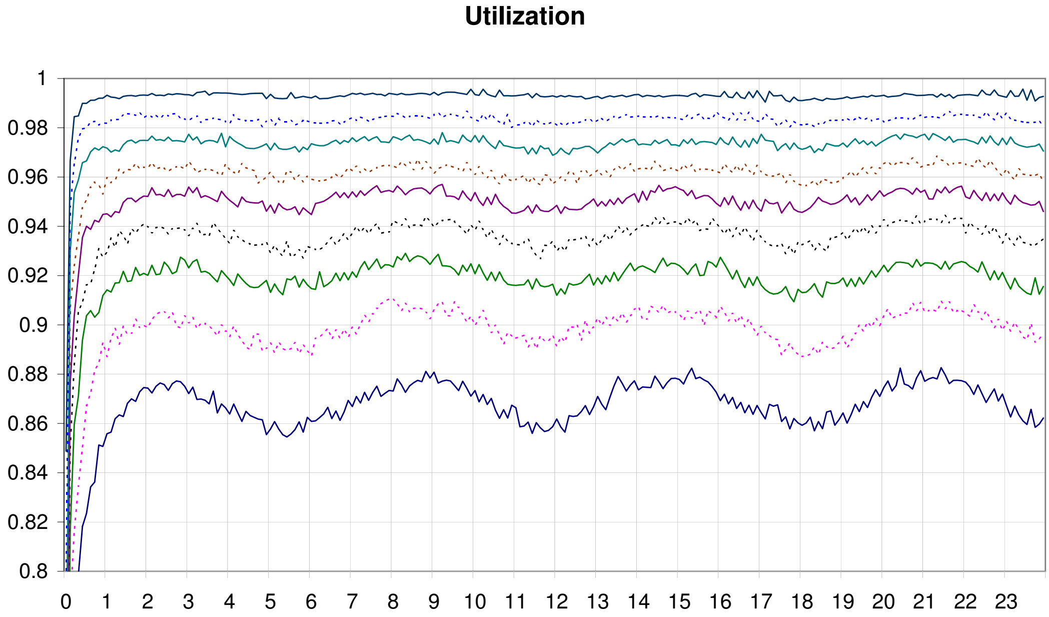


Figure 20: **Tail probability summary for the Erlang-C example**

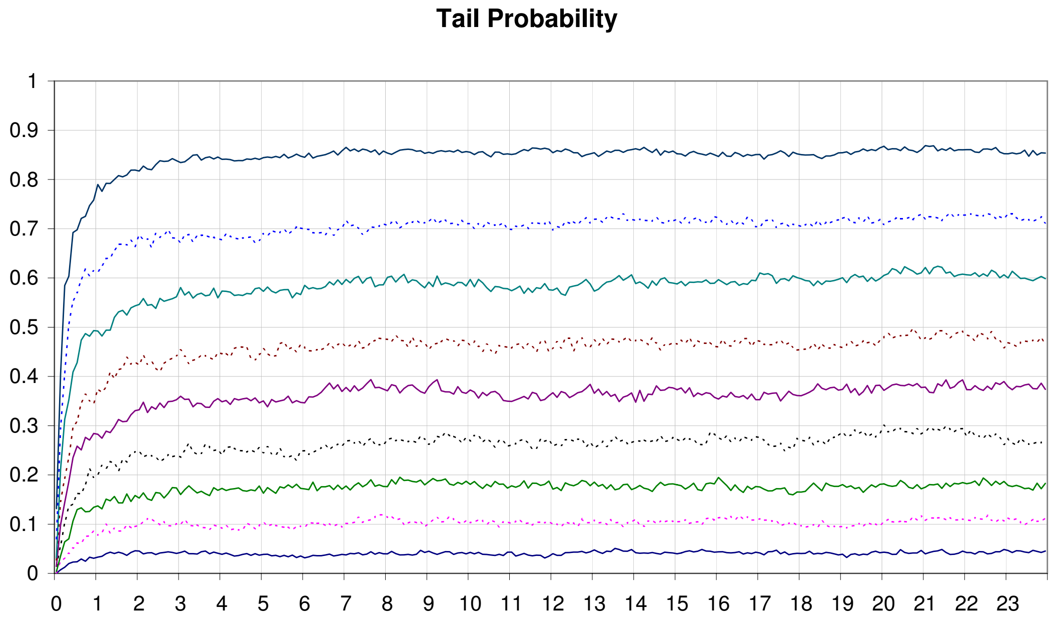


Figure 21: **Mean queue length and waiting time in the Erlang-C model with target α=0.5**

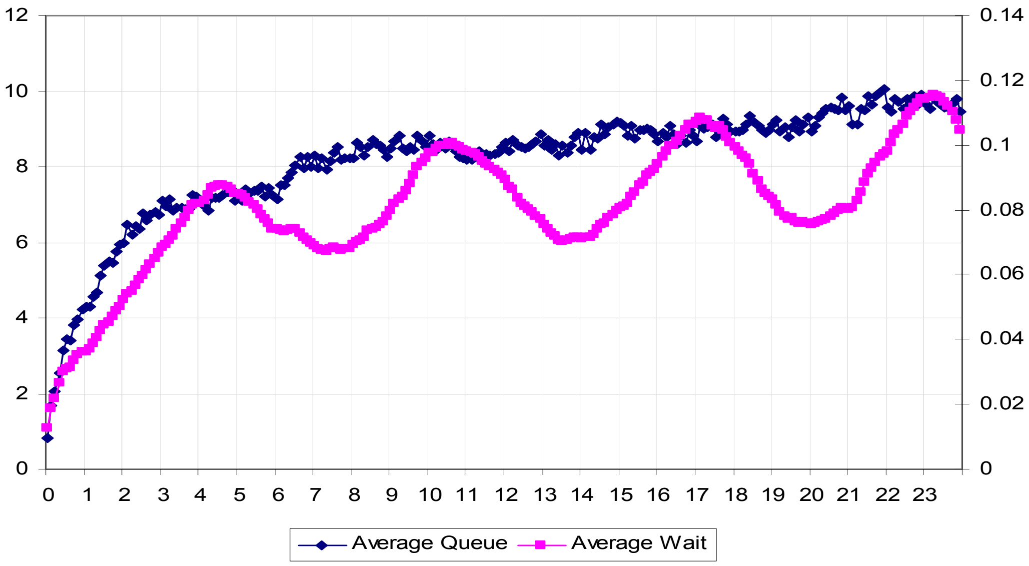
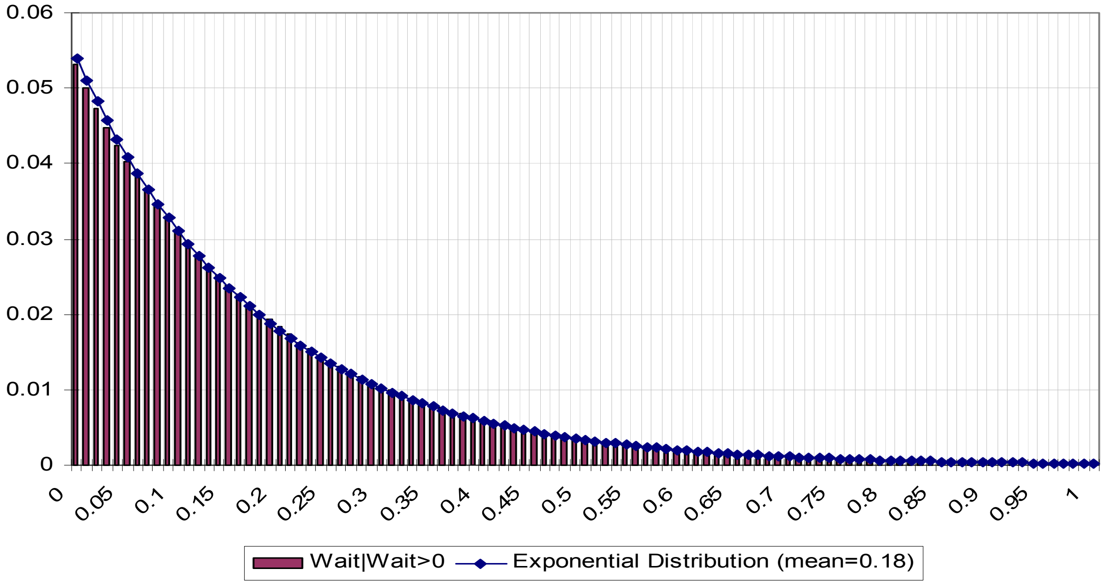


Figure 22: **The conditional distribution of the waiting time given delay in the Erlang-C model with target α=0.5**



6.2. 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 (2.3). We thus repeat the various experiments we did in §5. Recall that, for the *stationary* $M/M/s$ queue, the conditional waiting-time $(W \mid W > 0)$ is (exactly) exponentially distributed. The empirical conditional waiting-time distribution given wait, in our *time-varying* queue and over *all* customers, also fits the exponential distribution very well; see Figure **??**. The mean of the plotted exponential distribution was taken to be the overall average waiting time of those who were actually delayed during $[0, T]$.

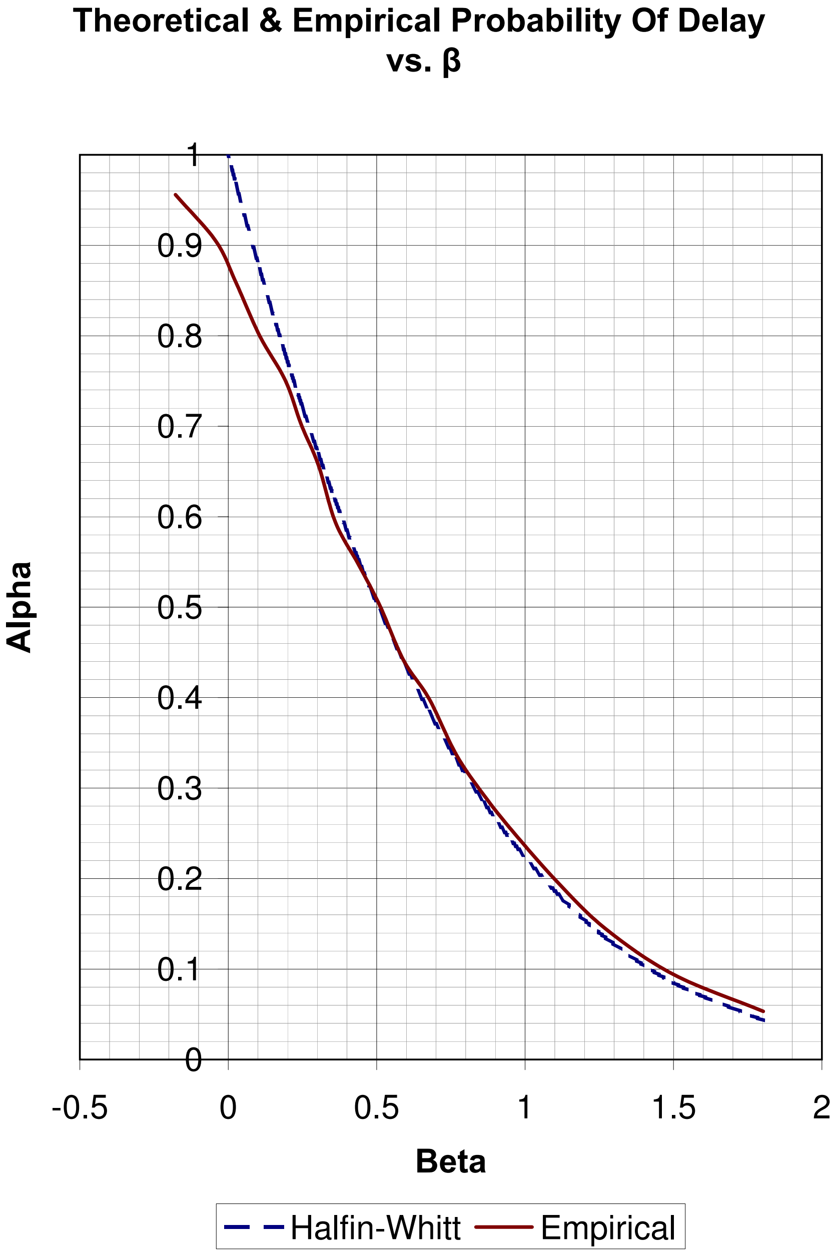
Here, the relation between α and β is compared with the **Halfin-Whitt function** from Halfin and Whitt (1981), namely,

$$P(delay) \equiv \alpha \equiv \alpha(\beta) \approx \left[1 + \beta \cdot \frac{\Phi(\beta)}{\phi(\beta)}\right]^{-1}, \quad 0 < \beta < \infty \ , \tag{6.1}$$

where ϕ is again the pdf associated with the standard normal cdf Φ. The Halfin-Whitt function in (6.1) is obtained from the Garnett function in (5.4) by letting $\theta \to 0$.

Just as we use the Garnett function to relate the target delay probability α to the quality-of-service parameter β in the square-root-staffing formula in (2.3) for the $M_t/M/s_t+M$ model, so we use the Halfin-Whitt function to relate α to β in the square-root-staffing formula in (2.3) for the $M_t/M/s_t$ model. And that essentially corresponds to the refinement performed in Section 4 of Jennings et al. (1996). The results in Figure 23 are again remarkable.

Figure 23: **Comparison of empirical results with the Halfin-Whitt approximation for the Erlang-C example**

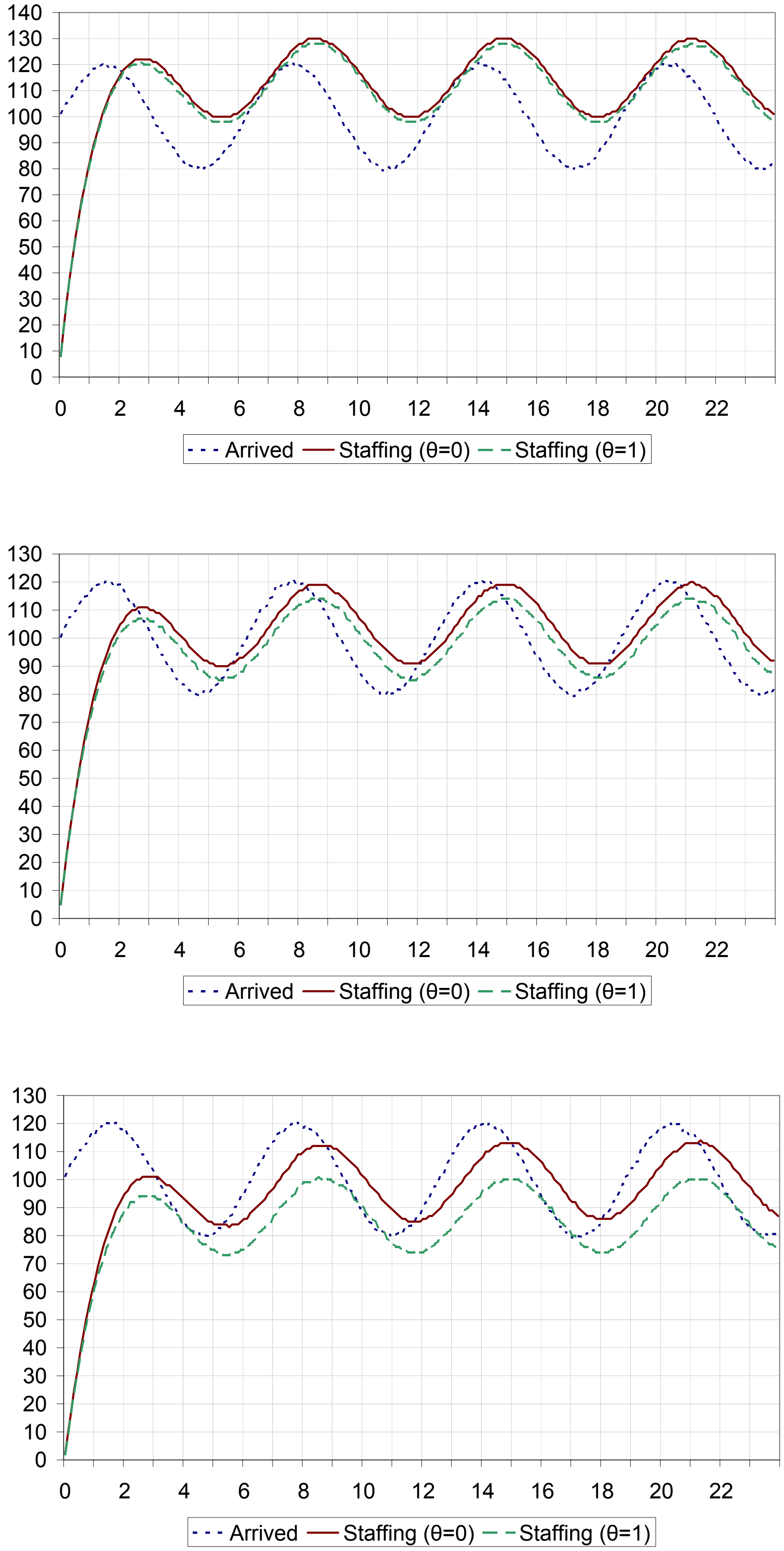


6.3. Benefits of Taking Account of Abandonment

We now show the benefit of staffing a system taking account of abandonment, assuming that abandonment in fact occurs. When compared to the model without abandonment, abandonment in the model reduces the required staff. In Figure 24 we show the difference between staffing levels for the two models introduced in §5 and §6, in the three regimes of operation: QD, QED and ED.

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 α increases.

Figure 24: **Staffing levels with and without customer abandonment ($\theta = 0$ and $\theta = 0$): (1) $\alpha = 0.1$ (2) $\alpha = 0.5$ (3) $\alpha = 0.9$**



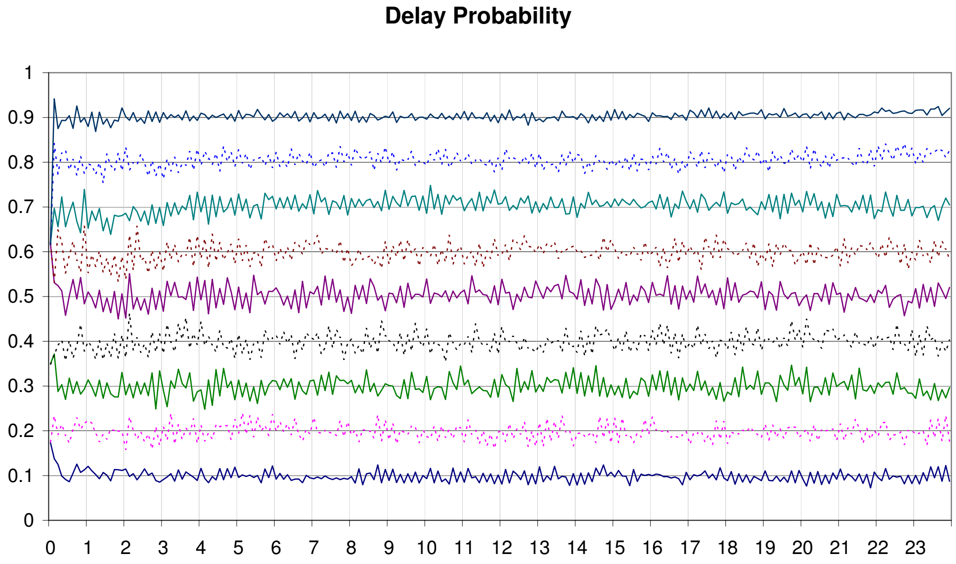
7. The Challenging Example from Jennings et al.

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.

This example is not greatly different from the Erlang-C example we have just considered in §6, but note that the frequency is 5 times greater here or, equivalently, the sinusoidal cycle is five times shorter. Thus the fluctuation in the arrival rate is even greater than in the example we have considered.

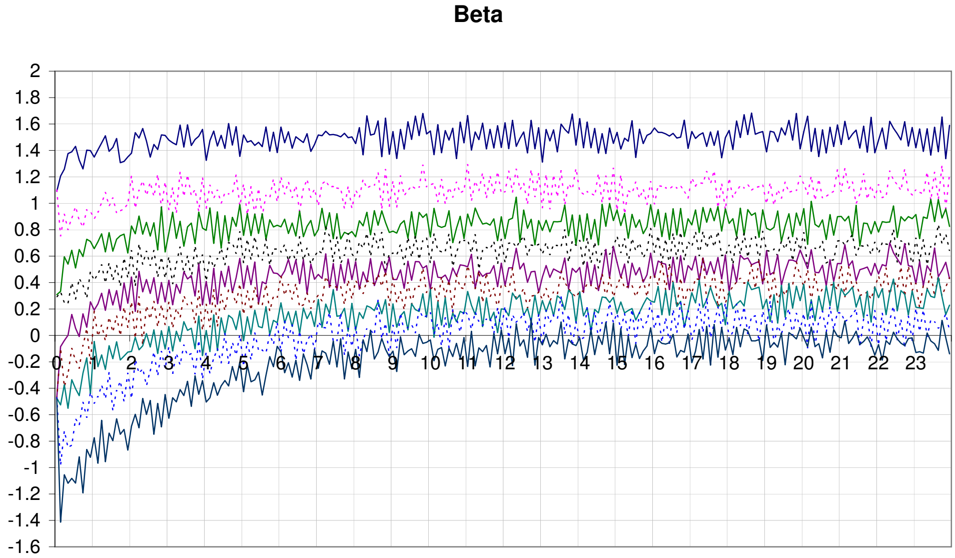
Figures 25 and 26 show that ISA also achieves time-stable performance for this example, for the full range of target delay probabilities, ranging from 0.1 to 0.9, just as before.

Figure 25: **Delay probability summary for the challenging example**



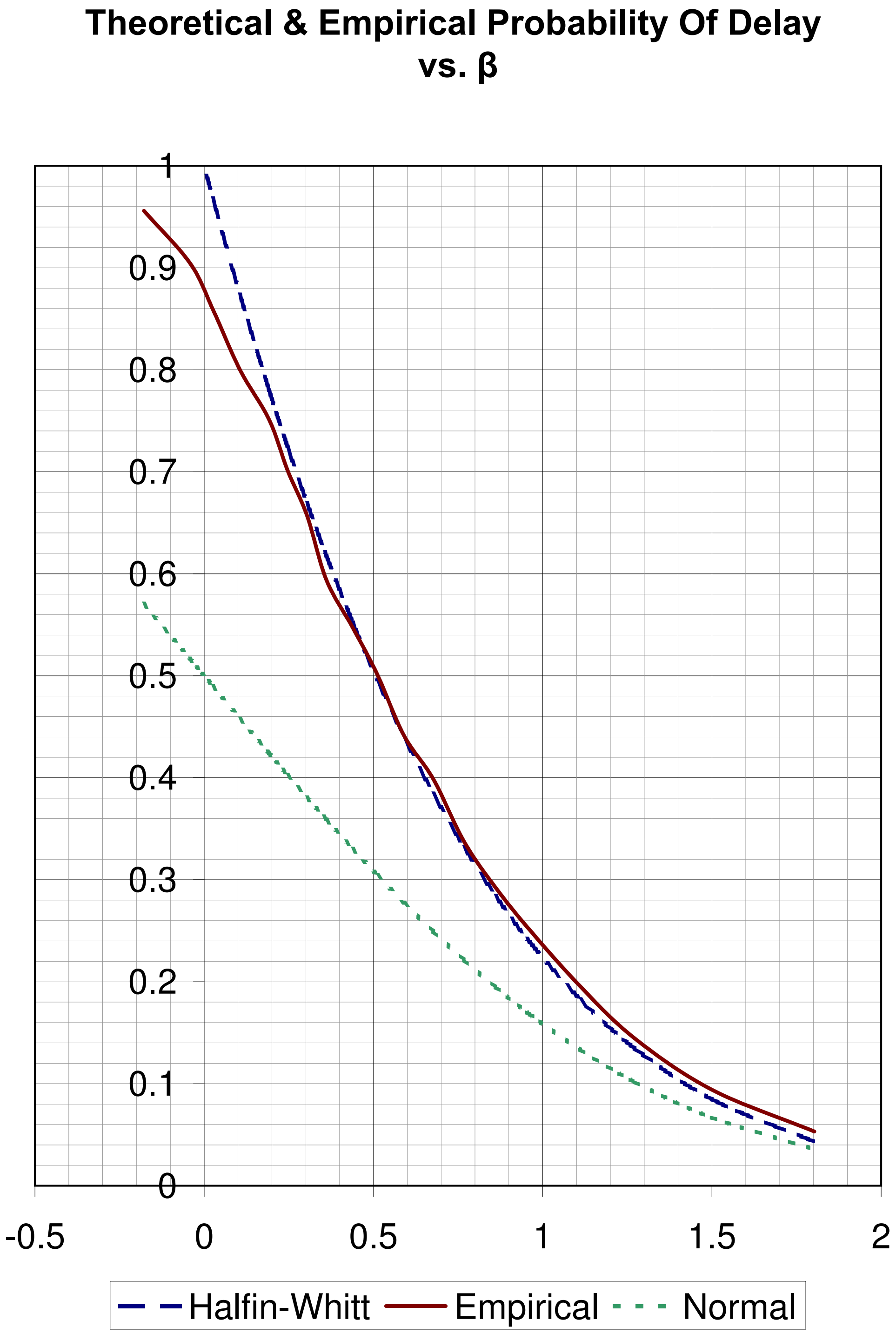
We now compare the empirical results with the Halfin-Whitt and normal approximations, paralleling Figures 12 and 23. We do so for this example below in Figure 27. Again the results are spectacular. In Figure 27 we use the Halfin-Whitt function in (6.1), just as in Figure

Figure 26: **Implied service quality β summary for the challenging example**



23. We also include the normal tail probability in (2.4), 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 Halfin-Whitt function, as we noted before.

Figure 27: **Comparison of empirical results with the Halfin-Whitt and normal approximation for the challenging example**



8. The Time-Varying Erlang-A Model with More and Less Patient Customers

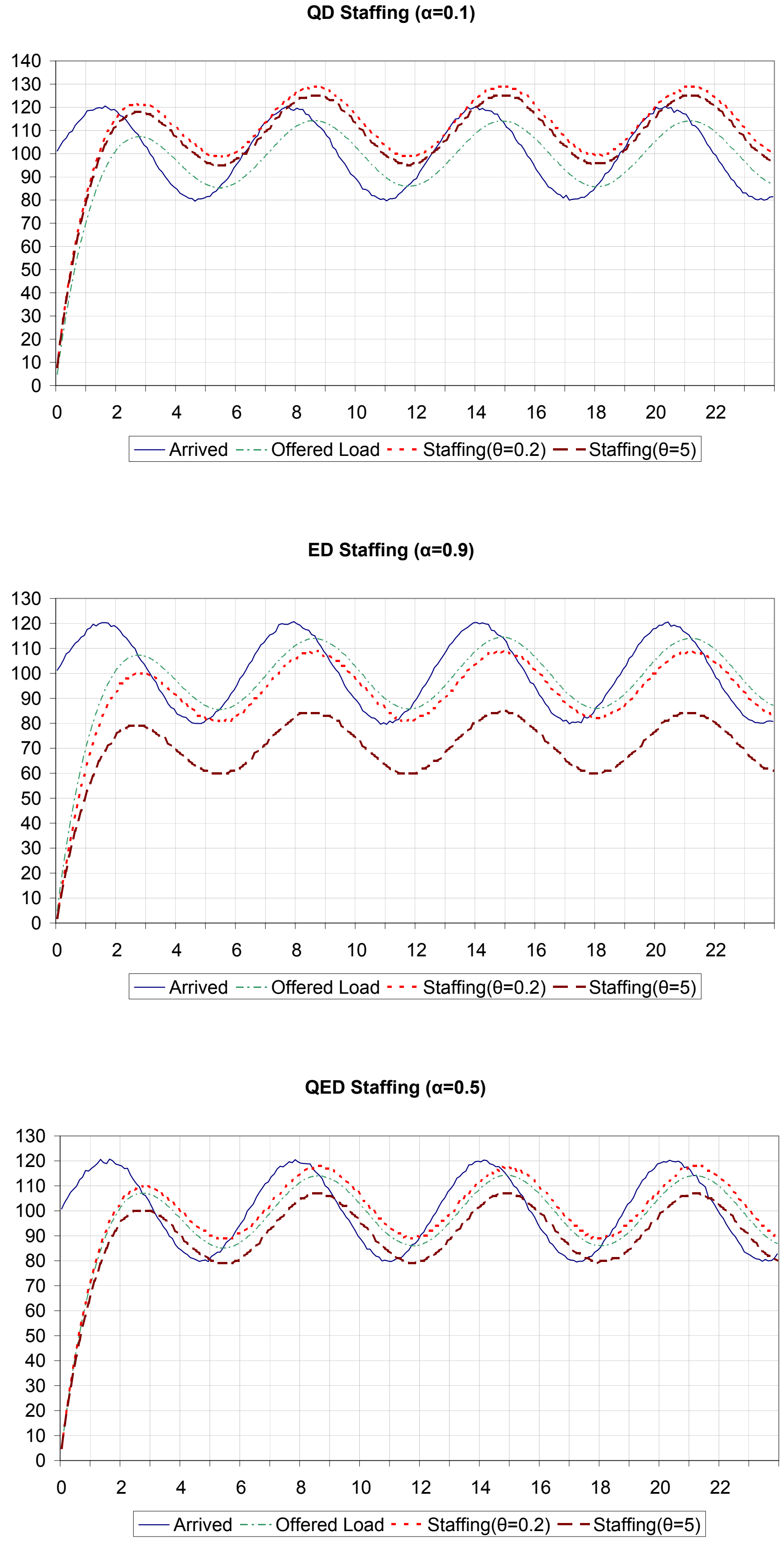
We now return to the time-varying Erlang-A model ($M_t/M/s_t + M$) considered in Section 5, except we change the patience parameter, i.e., the individual abandonment rate θ.

8.1. More and Less Patient Customers

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; 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 quality β 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 28 below. We compare the time-dependent abandonment $P_t(Ab)$ in these two scenarios in Figure 29. Note that the gap between the required staffing levels in the two cases - $\theta = 0.2$ and $\theta = 5.0$ - grows as the delay-probability target α increases, being quite small when $\alpha = 0.1$, but being very dramatic when $\alpha = 0.9$.

Figure 28: **Staffing for time-varying Erlang-A with more patient ($\theta = 0.2$) and less patient ($\theta = 5.0$) customers: (1) $\alpha = 0.1$ (QD), (2) $\alpha = 0.9$ (ED), (3) $\alpha = 0.5$ (QED)**

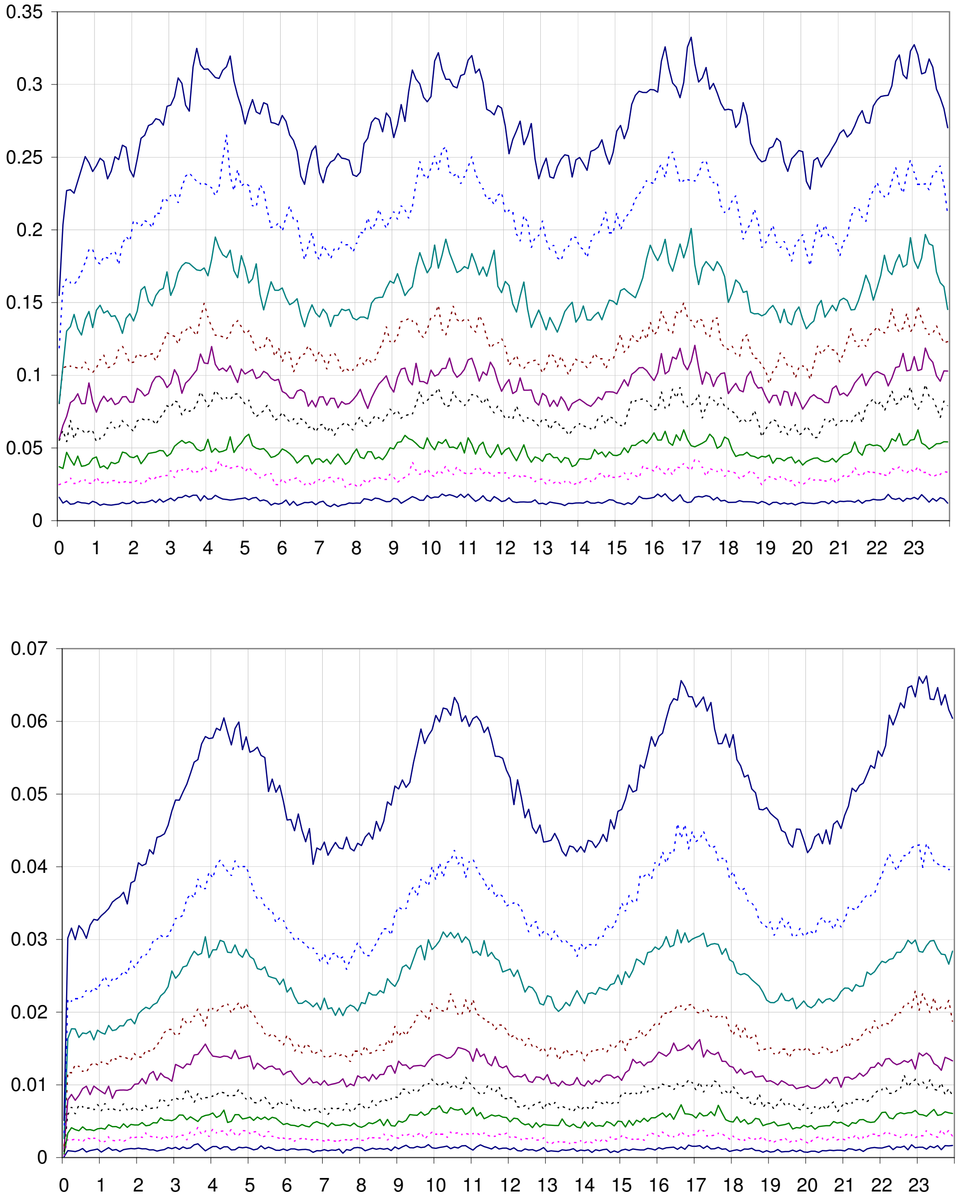


QD Staffing (α=0.1)

ED Staffing (α=0.9)

QED Staffing (α=0.5)

Figure 29: **Abandonment probabilities for the time-varying Erlang-A example with the new patience parameters: (1) θ=5 (2) θ=0.2**



We compare the empirical (α, β) pairs produced by ISA to the Garnett function in (5.4) for these two cases in Figure 30. 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 formula in (2.3), obtaining the required service quality β from the target delay probability α by using the inverse of the Garnett function in (5.4), which reduces to the Half-Whitt function in (6.1) 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 31 below.

Figure 30: **Comparison of the empirical results from ISA with the Garnett approximation for the time-varying Erlang-A example with the new patience parameters: θ=5 and θ=0.2**

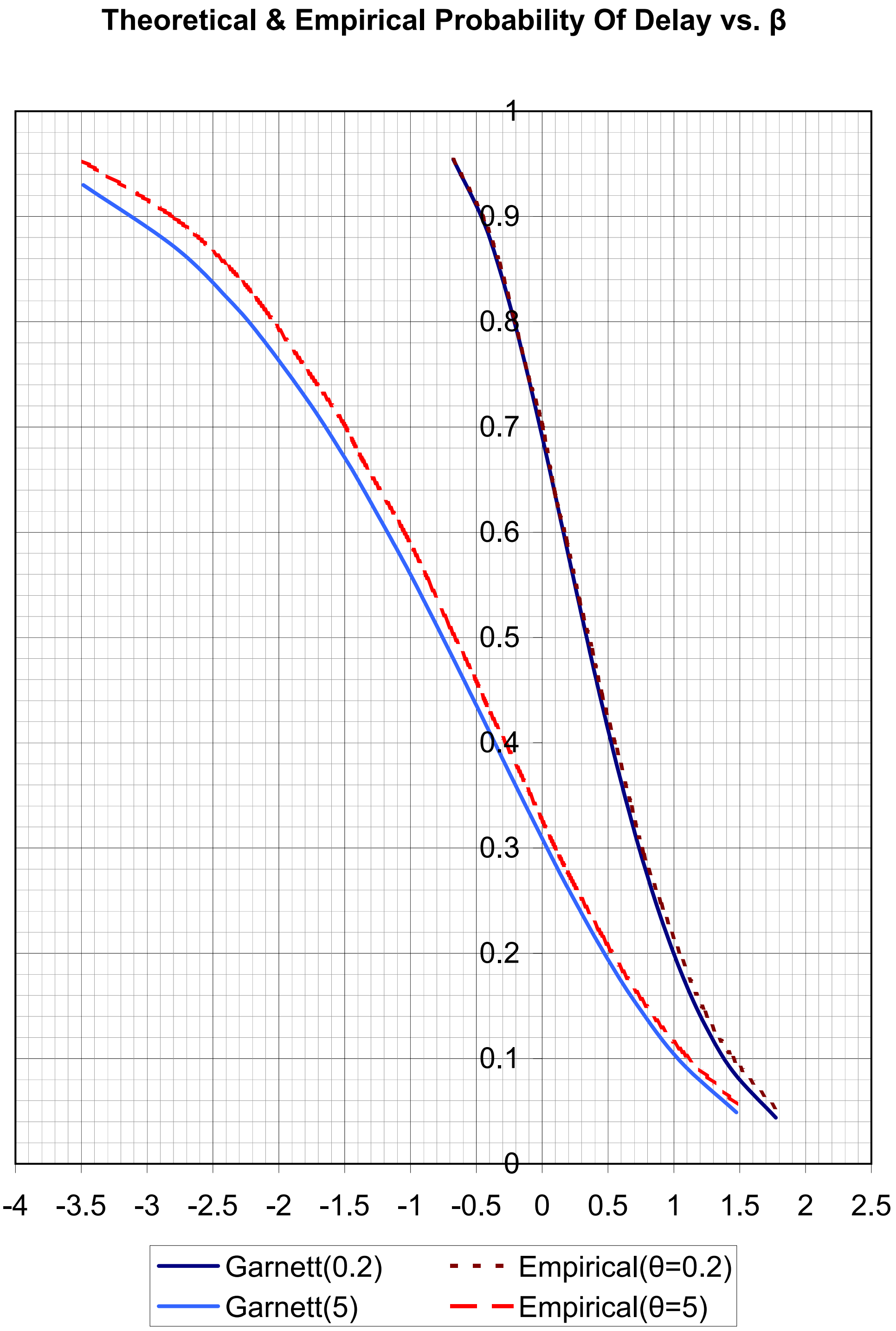
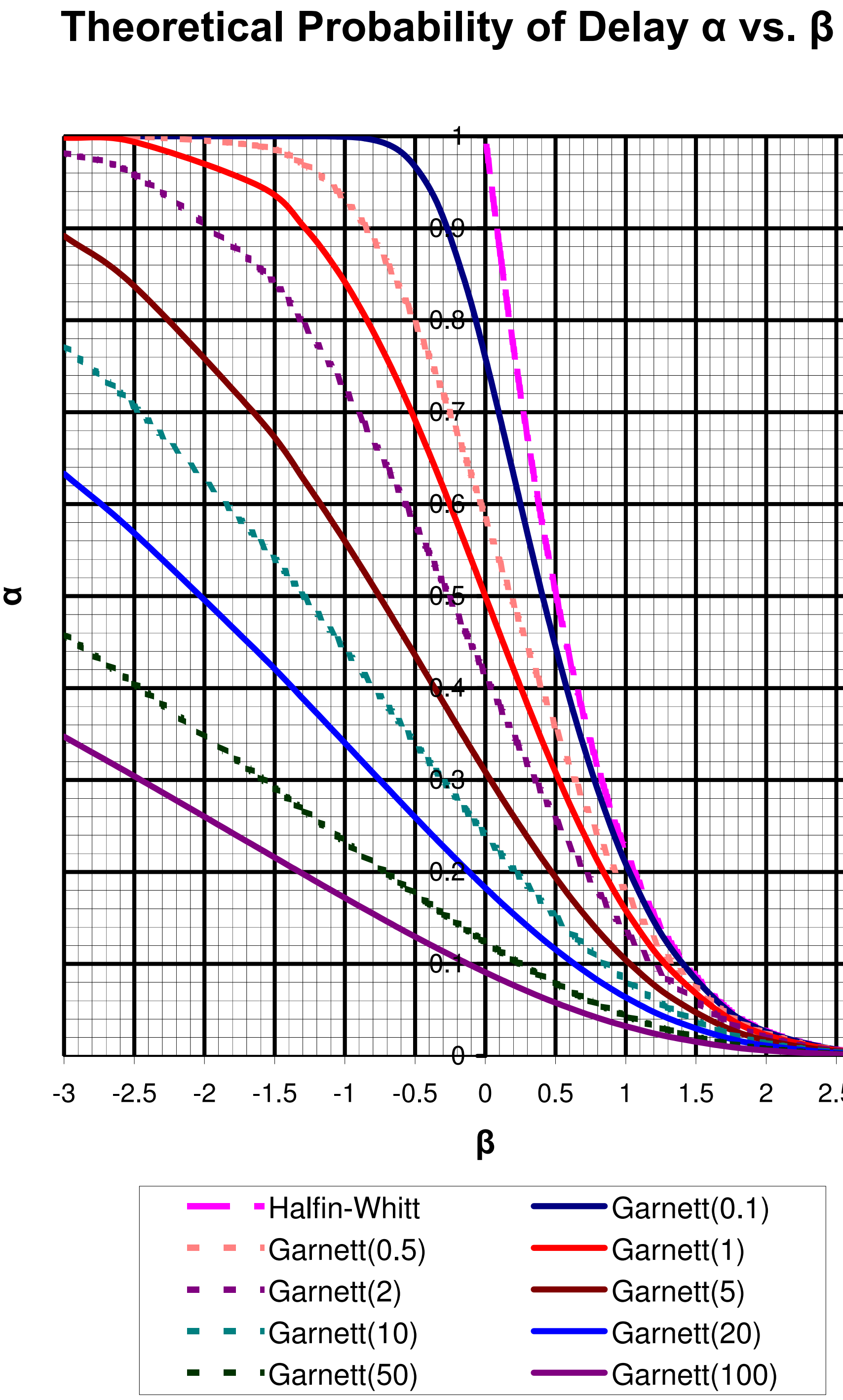


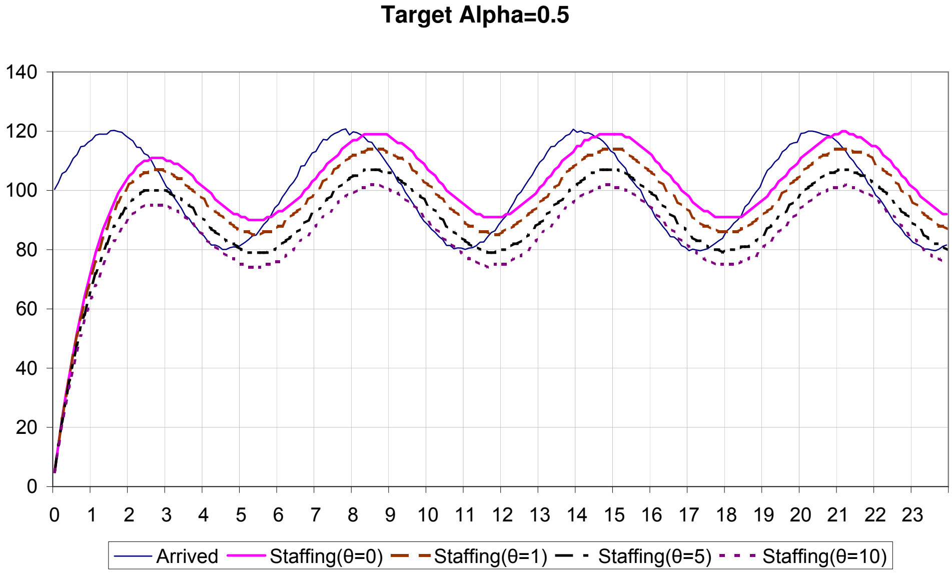
Figure 31: **The Halfin-Whitt/ Garnett functions**



8.2. Benefits of Taking Account of Abandonment Again

Following §6.3, 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 θ increases, bringing additional savings. In Figure 32 we show the comparison for delay probability $\alpha = 0.5$, which we consider to be a reasonable operational target.

Figure 32: **Staffing levels for the time-varying Erlang-A example for a range of (im)patience parameters**



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.

8.3. 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).

9. A Realistic Example Related to Figure 1

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 5. Results are shown below.

Figure 33: **Delay probability summary for the realistic example**

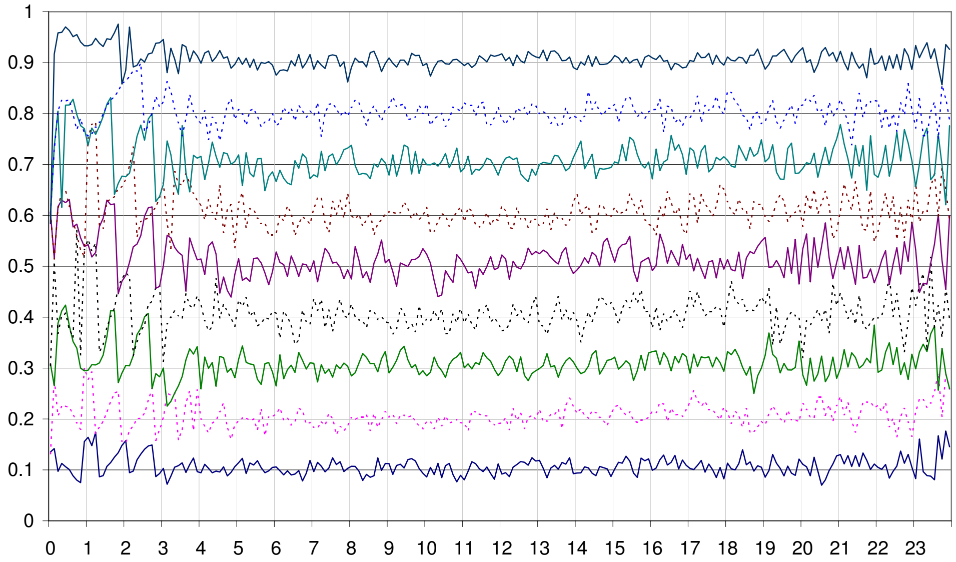


Figure 34: **Implied service quality β summary for the realistic example**

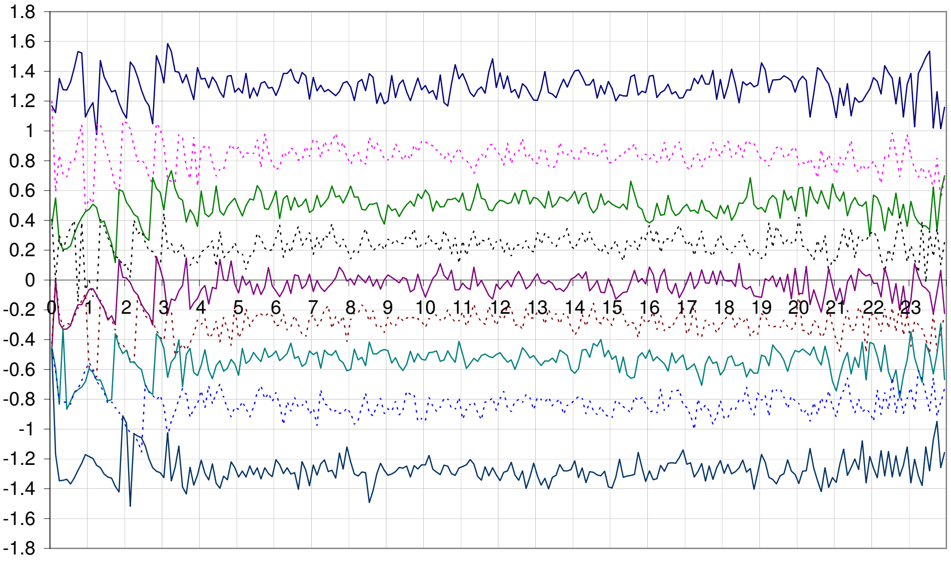


Figure 35: **Abandon probability summary for the realistic example**

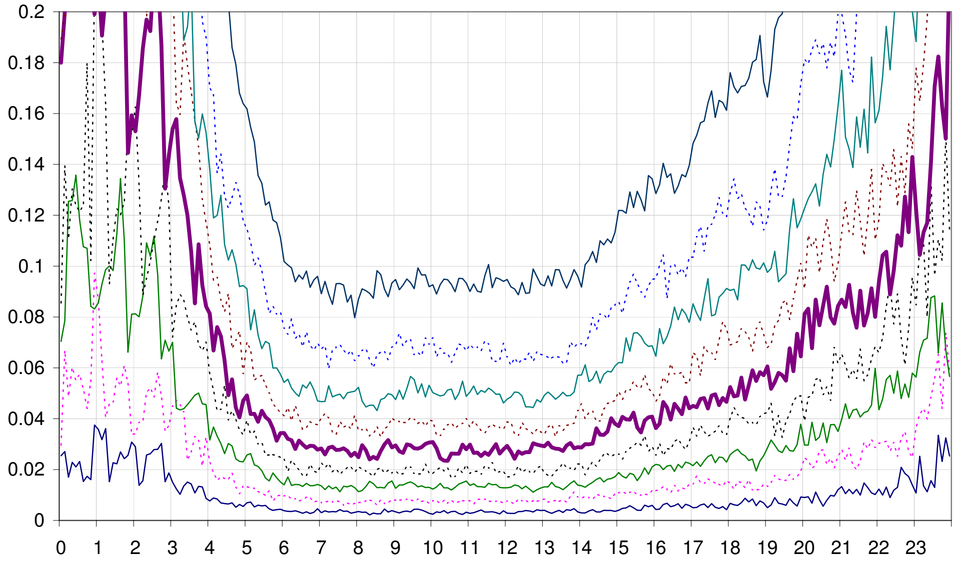
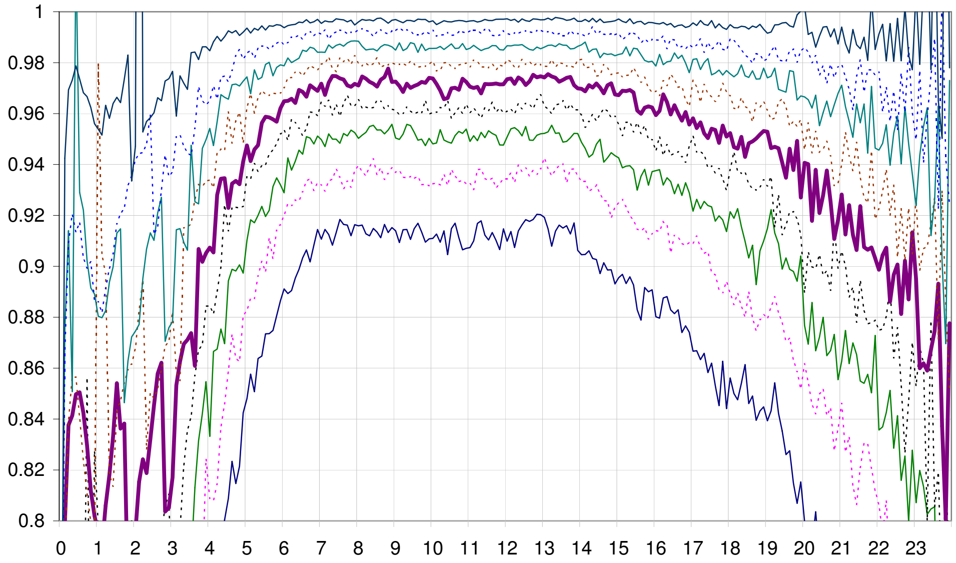


Figure 36: **Utilization summary for the realistic example**

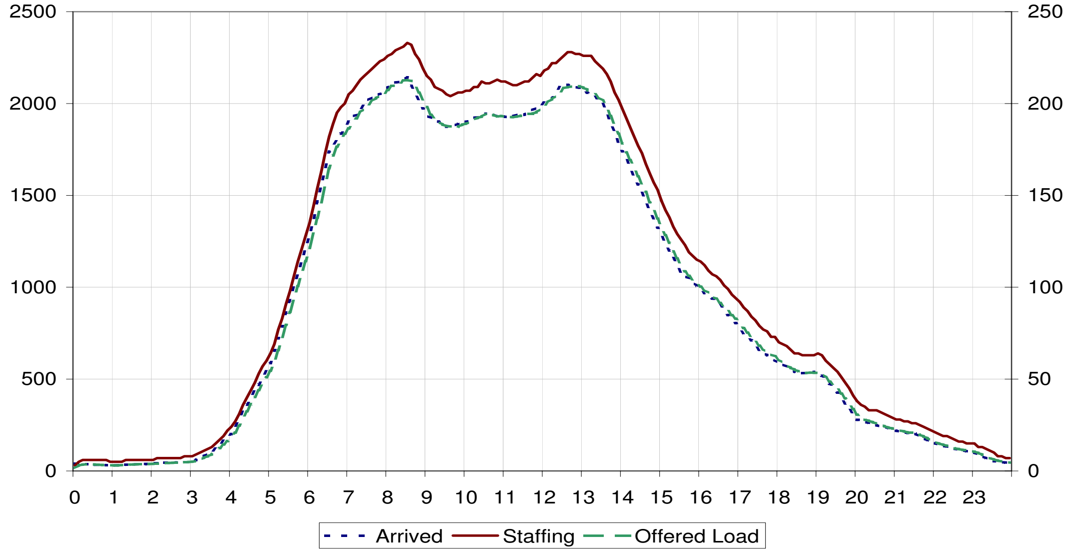


At first, we are struck by the observation that the algorithm is not as successful as before, because the target delay probability is not achieved accurately at the beginning and at the end of the day. Moreover, not all performance measures are stable over the entire day. However, this bad behavior is 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 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 quality (up to delay probability of 0.5), performance measures are varying in quite a small range, that would look appealing to most system designers.

In Figures 9- 9 we describe the performance of ISA in the three regions: QD ($\alpha = 0.1$), ED ($\alpha = 0.9$) and QED ($\alpha = 0.5$). 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 that the MOL approximation essentially coincides with the elementary PSA approximation. Consequently, the square-root-staffing formula (2.3) will perform the same using either the infinite-server offered load or the PSA offered load. Moreover, the ISA performs the same as the square-root-staffing formula with either the infinite-server offered load or the PSA offered load.

Consequently, ISA does not differ much from PSA. However, for the time-varying Erlang-A model, staffing using PSA is actually not so routine. We need to apply the steady-state distribution of the $M/M/s + M$ model or a suitable approximation.

Figure 37: **The realistic example in the QD regime (target α=0.1): (1) staffing level, offered load and arrival function, (2) average queue length and average waiting time (in average service time)**



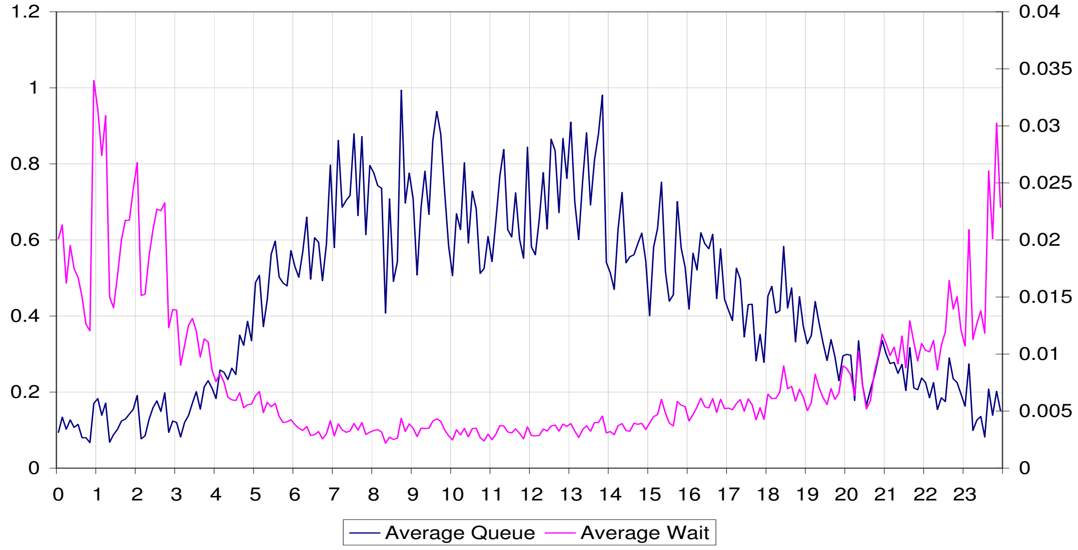
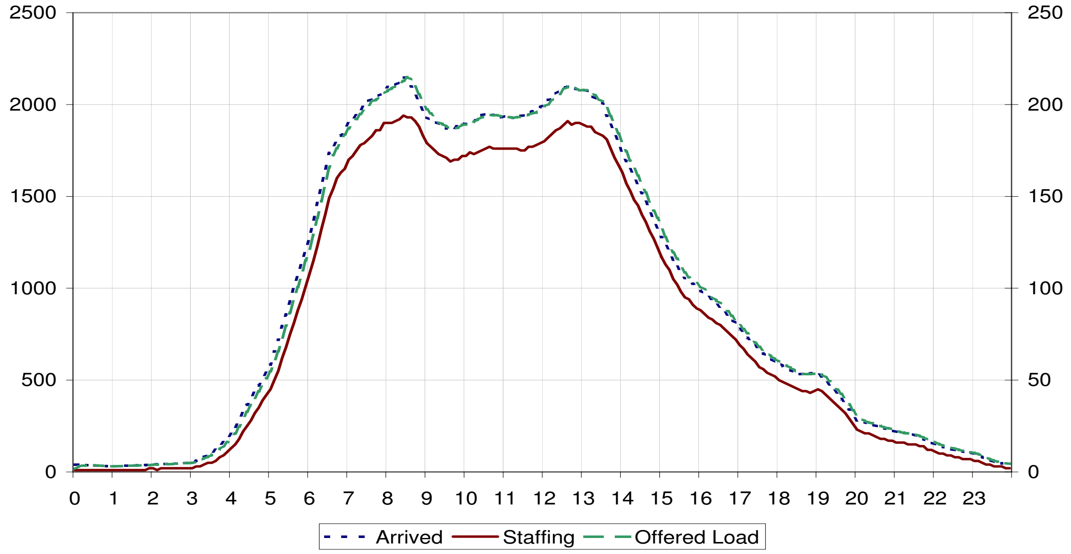


Figure 38: **The realistic example in the ED regime (target α=0.9): (1) staffing level, offered load and arrival function, (2) average queue length and average waiting time (in average service time)**



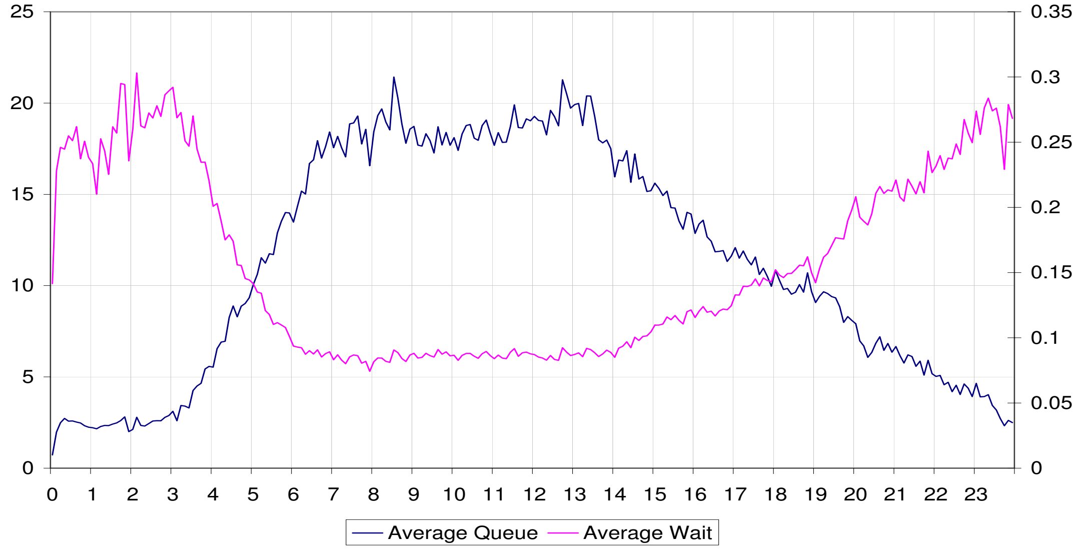
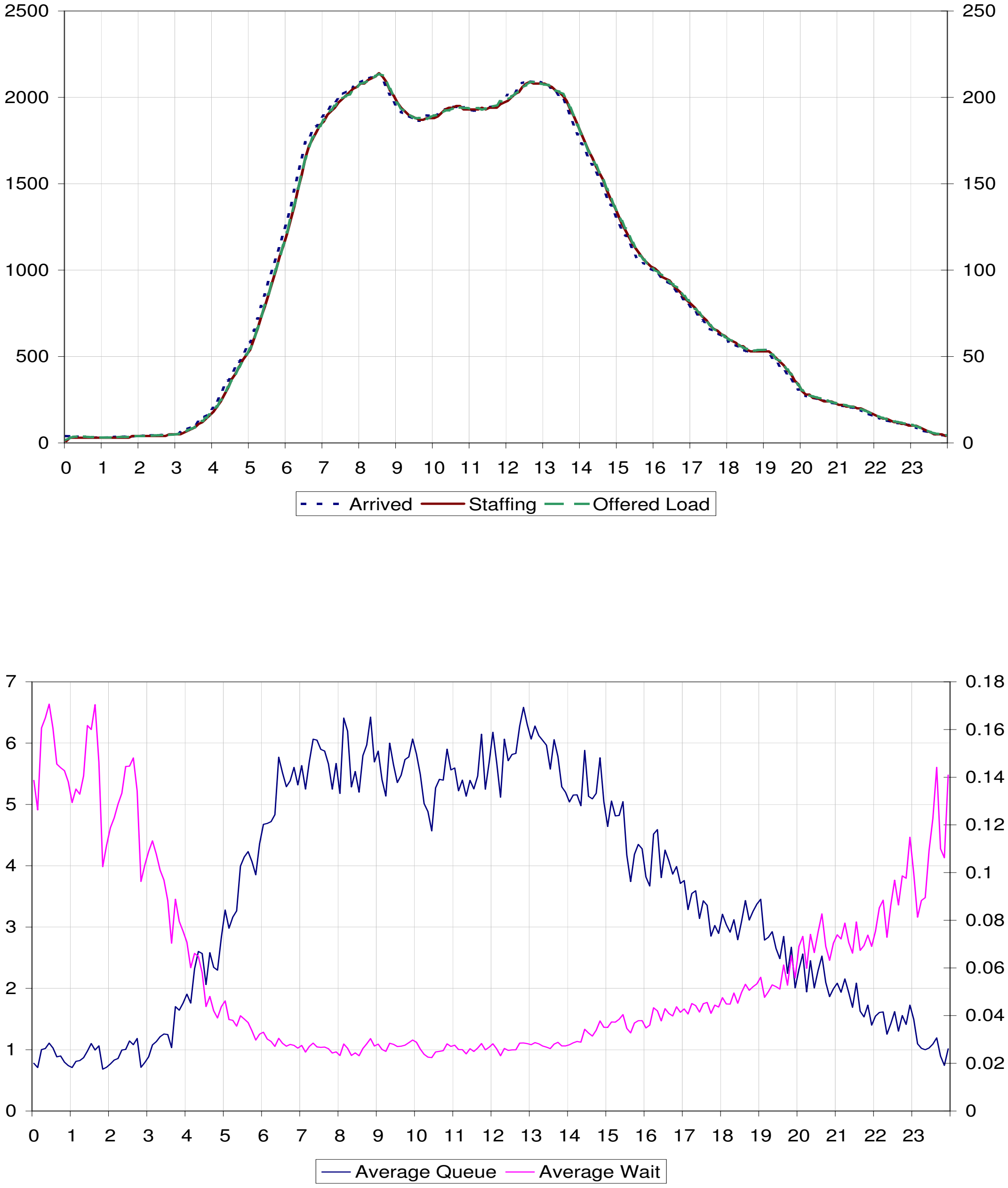
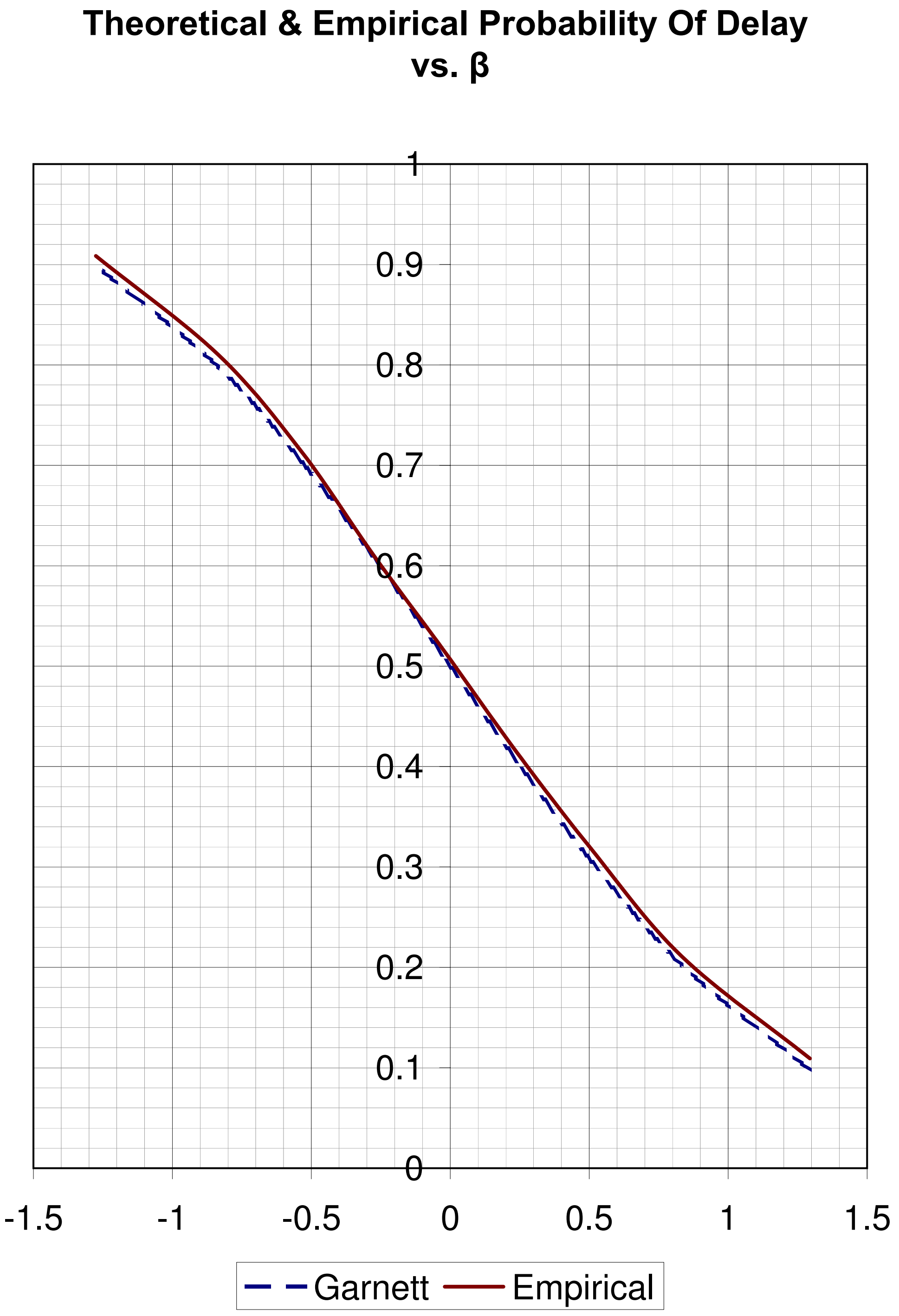


Figure 39: **The realistic example in the QED regime (target α=0.9): (1) staffing level, offered load and arrival function, (2) average queue length and average waiting time (in average service time)**



The three regimes of operation in Figures 9-9 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 9 shows, once again, that the staffing falls right on top of the offered load in the QED regime. Figure 9 shows that the excellent matching between the Garnett function and the empirical results is preserved also in this example.

Figure 40: **Comparison of empirical results with the Garnett approximation for the realistic example**



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

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

10.1. Connections to Other Models

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

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

$$\{L_t : t \geq 0\} \stackrel{\mathrm{d}}{=} \{L_t^\infty : t \geq 0\} \ . \tag{10.1}$$

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

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

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

$$L_t \stackrel{\mathrm{d}}{=} L_\infty^\infty \stackrel{\mathrm{d}}{=} L_\infty \ . \tag{10.2}$$

In (10.2), the second random variable L_∞^∞ is the steady-state number of busy servers in the stationary $M/M/\infty$ with arrival rate $\hat{\lambda}_t$ in (2.9), with m_t again the expected number in system in the time-dependent infinite-server model in (2.5). Since S has an exponential distribution, $S_e \stackrel{\mathrm{d}}{=} S$. In (10.2), the third random variable L_∞ is the steady-state number in system in the $M/M/s+M$ model with the constant number of servers equal to s_t and the arrival rate again being $\hat{\lambda}_t$ in (2.9).

10.2. The Delay Probability

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

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

$$\begin{aligned} P(W_t > 0) &= P(L_t \geq s_t) = P(L_t^\infty \geq s_t) = P(Poisson(m_t) \geq s_t) \\ &\approx P\left(N(0,1) > \frac{s_t - m_t}{\sqrt{m_t}}\right) \ , \end{aligned} \tag{10.3}$$

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

10.3. Approximations for the Waiting-Time Distribution

However, the virtual abandonment probability P_t^{ab} and the expected virtual waiting time $E[W_t]$ fluctuate much more than the delay probability; e.g., see Figure 8. We will explain that greater fluctuation.

We actually can mathematically analyze the time-dependent virtual waiting time W_t and the time-dependent virtual abandonment probability P_t^{ab}. Here is an important initial observation: Conditional on the event that $W_t > 0$, whose probability we have analyzed above, W_t is distributed (exactly) as the first passage time of the (Markovian) stochastic process $\{L_u : u \geq t\}$ from the initial value L_t encountered at time t down to the staffing function $\{s_u : u \geq t\}$, provided that we ignore all future arrivals after time t. In other words, W_t is distributed as the first passage time of the pure-death stochastic process with state-dependent death rate μL_u for $u \geq t$ down from the initial value L_t to the curve $\{s_u : u \geq t\}$. (Of course, $W_t = 0$ if $L_t < s_t$.) As a consequence, the distribution of W_t and the value of P_t^{ab} depend on only L_t and the future staffing levels, i.e., $\{s_u : u \geq t\}$. The time-dependent arrival-rate function contributes nothing further. It is easy to see that we can establish stochastic bounds on the distribution of W_t if the staffing level is monotone after time t.

We can go further if we make approximations. Even though exact relations are difficult to obtain, it is not difficult to generate very good approximations for the case in which the number of servers tends to be large, e.g., as in the specific example in the previous subsection. Then, W_t tends to be very small, so that it is often reasonable to assume that the staffing level remains constant at s_t in the time shortly after t. In other words, to study W_t and P_t^{ab}, we make the approximation $s_u \approx s_t$ for all $u > t$. We make this approximation, not because the staffing level should be nearly constant for all u after t, but because we think we only need to consider times u slightly greater than t. We are thinking of applications in which the time-dependent arrival-rate function is continuous, and the staffing changes relatively slowly.

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

$$W_t \approx W_\infty \quad \text{and} \quad P_t^{ab} \approx P_\infty^{ab} \ , \tag{10.4}$$

where the constant staffing level in the stationary $M/M/s + M$ model on the righthand sides is chosen to be s_t and the constant arrival rate is chosen to be $\hat{\lambda}_t$ in (2.9). Hence, we propose (10.4) as approximations.

Given approximations (10.4), we can use established results for the stationary $M/M/s+M$ model, e.g., as in Garnett et al. (2002) and Whitt (2005). For example, algorithms to compute the (exact) distribution of W_∞ are given there, including the corresponding conditional distributions obtained when we condition on whether or not the customer eventually is served.

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

In this subsection we turn to an issue not included in the main paper. 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 λ_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 (2.3) with a fixed target delay probability α. 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^{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 $E[W_t]$ and the virtual abandonment probability P_t^{ab}.

To proceed, we exploit the approximations in (10.4). From the approximation for the mean, we obtain the associated approximation

$$E[W_t] \approx E[W_\infty] \tag{10.5}$$

where the constant staffing level in the stationary $M/M/s+M$ model on the righthand side is chosen to be s_t and the constant arrival rate is chosen to be $\hat{\lambda}(t) \equiv \mu m_t$ in (2.9).

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

$$\sqrt{m_t} P_t^{ab} \to \eta \quad \text{and} \quad \sqrt{m_t} E[W_t] \to \frac{\eta}{\theta} \tag{10.6}$$

as $m_t \to \infty$, where

$$\eta \equiv \alpha E[N(0,1) - \beta | N(0,1) > \beta] = \alpha \left(\frac{\phi(\beta)}{\Phi^c(\beta)} - 1 \right) > 0 \tag{10.7}$$

and $\theta = \mu$.

The important practical conclusion we deduce from (10.6) is that we see that $\sqrt{m_t} P_t^{ab}$ and $\sqrt{m_t} E[W_t]$ will be asymptotically constant (time-stable and nondegenerate) as m_t increases if we are in the QED regime. However, in general, consistent with Figure 8, the performance measures P_t^{ab} and $E[W_t]$ themselves need not be asymptotically time-stable. In order for them to be asymptotically time-stable too, we need to impose extra conditions upon the mean function m_t itself.

We actual see the greatest departures from time-stability of P_t^{ab} and $E[W_t]$ for the $M_t/M/s_t + M$ model (e.g., in Figure 8) when the target delay probability is large. In those cases, it is evident that the system actually should be regarded as being 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 $E[W_t]$, 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.

11. Algorithm Dynamics

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

11.1. Empirical Observations

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

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

$$s_t^{(n)} \leq s_t^{(m)} \qquad \text{for all} \quad m < n\,. \tag{11.1}$$

An example of the monotone dynamics is shown in Figure 41, 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 ($\mathbf{r} = 10$).

In contrast, with *oscillating dynamics*, $s_t^{(n)}$ is oscillating for all t; i.e. there exist 2 subsequences $\{s_t^{(k)}\}_{k=2n}^{\infty}$ and $\{s_t^{(l)}\}_{l=2n+1}^{\infty}$, such that $s_t^{(2n)} \downarrow s_t^{(\infty)}$ and $s_t^{(2n+1)} \uparrow s_t^{(\infty)}$. Within the oscillating framework, there is monotonicity. An example of the oscillating dynamics can be viewed in Figure 42, where staffing levels are shown for the first three iterations for the same case except there is no abandonment ($\theta = 0$ and $\mathbf{r} = 0$).

11.2. Theoretical Explanation

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

Figure 41: **Monotone algorithm dynamics when $\theta > \mu$: staffing levels in the 1^{st}, 2^{nd} and last iterations. μ=1, θ=10.**

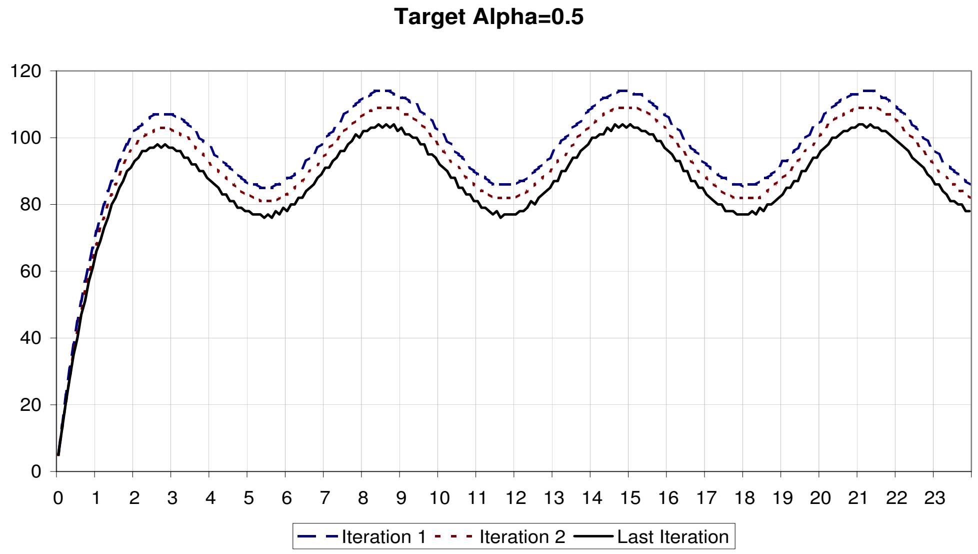
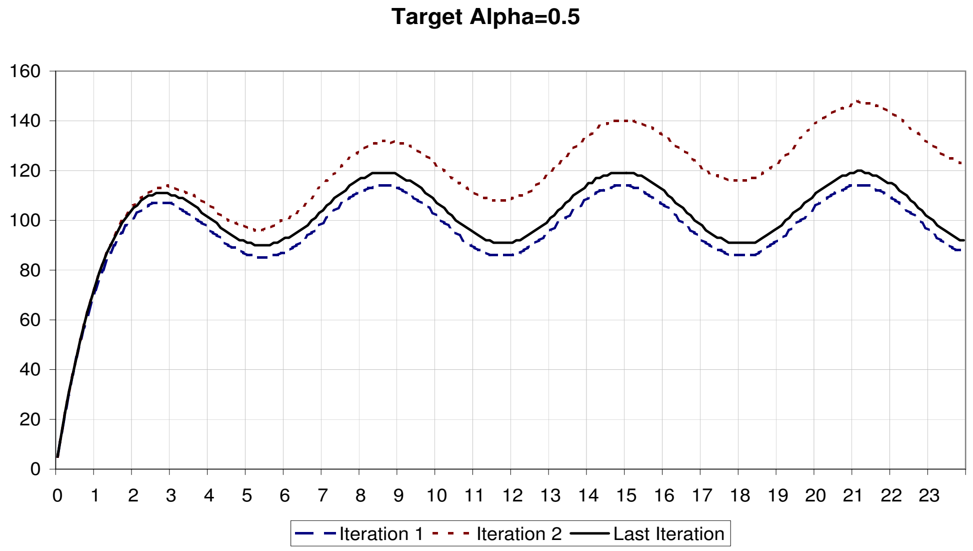


Figure 42: **Oscillating algorithm dynamics when $\theta < \mu$: staffing levels in the 1^{st}, 2^{nd} and last iterations. μ=1, θ=0**



arrival process is the same. However, the death rates depend systematically on the number of servers s_t. When $\mathbf{r} > 1$ ($\mathbf{r} < 1$), the death rates at time t decrease (increase) as s_t increases. Hence, if we disregard statistical error, caused by having to estimate the delay probabilities associated with each staffing function, we can actually prove that the algorithm converges for the $M_t/M/s_t + M$ model. To do so, we use sample-path stochastic order, as in Whitt (1981). We only need ordinary stochastic-order for each time t, but in order to get that, we need to properly address what happens before time t as well.

Here is the **key stochastic-order property** for the $M_t/M/s_t + M$ model: If $s_t^{(1)} \le s_t^{(2)}$ for all t, $0 \le t \le T$, and $\mathbf{r} > 1$, then

$$\{L_t^{(1)} : 0 \le t \le T\} \le_{st} \{L_t^{(2)} : 0 \le t \le T\} \ , \tag{11.2}$$

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

$$E\left[f\left(\{L_t^{(1)} : 0 \le t \le T\}\right)\right] \le_{st} E\left[f\left(\{L_t^{(2)} : 0 \le t \le T\}\right)\right] \tag{11.3}$$

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 (11.2) 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 (11.2).

The sample-path ordering obtained ensures that a departure occurs in the lower process whenever it occurs in the upper process and the two sample paths are equal. As indicated above, the two processes are given identical arrival streams. Then we construct all departures (service completions or abandonments) from those of the lower process at epochs when the two sample paths are equal. Suppose that at time t the sample paths are equal: $L_t^{(1)} = L_t^{(2)} = k$. Then, at that t, the death rates in the two birth and death processes are necessarily ordered by $\delta_1(k) \ge \delta_2(k)$. We only let departures occur in process 2 when they occur in process 1, so the two sample paths can never cross over. When a departure occurs in process 1 with both sample paths in state k, we let a departure also occur in process 2 with probability $\delta_2(k)/\delta_1(k)$, with no departure occurring in process 2 otherwise. This keeps the sample paths ordered w.p. 1 for all t. At the same time, the two stochastic processes individually have the correct finite-dimensional distributions. The construction is just like the thinning of a Poisson process used in the simulation of a nonhomogeneous Poisson process.

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

$$L_t^{(1)} \leq_{st} L_t^{(2)} \quad \text{for all} \quad t \ , \tag{11.4}$$

where now $\leq_{st}$ denotes conventional stochastic order for real-valued random variables, just as in Chapter 9 of Ross (1996); also see Müller and Stoyan (2002). We only need the more elementary stochastic order in (11.4), but we use the more sophisticated sample-path stochastic order in (11.2) to get it. The stochastic order is equivalent to the tail probabilities being ordered; i.e., (11.4) is equivalent to $P(L_t^{(1)} > x) \leq P(L_t^{(2)} > x)$ for all x, which implies the ordering for the staffing functions at time t. In particular, suppose that

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

Since

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

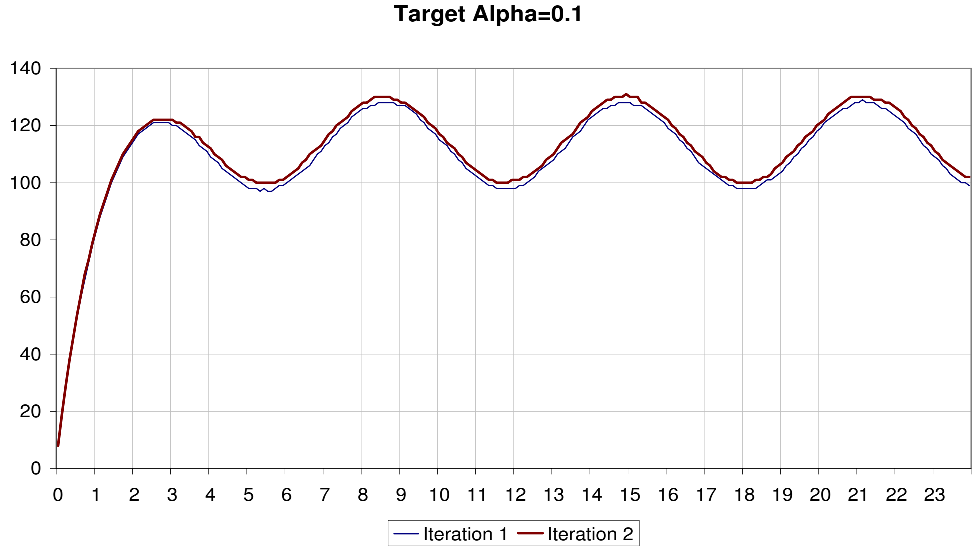
necessarily $s_t^{(1)} \leq s_t^{(2)}$.

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

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

We also observed that the **target delay probability** α strongly influenced the dynamics. In particular, higher values of α cause larger oscillations in the oscillating case, and slower convergence to the limit in all cases. This phenomenon is illustrated in Figures 43 and 44. 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 43) and $\alpha = 0.5$ (Figure 44) for the case of

Figure 43: **Algorithm dynamics: range of staffing level for target α=0.1**



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

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

We observed that $I_{t_2} \geq I_{t_1}$ for all $t_2 > t_1$. An illustration can be viewed in Figure 45. 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 44: **Algorithm dynamics: range of staffing level for target α=0.5**

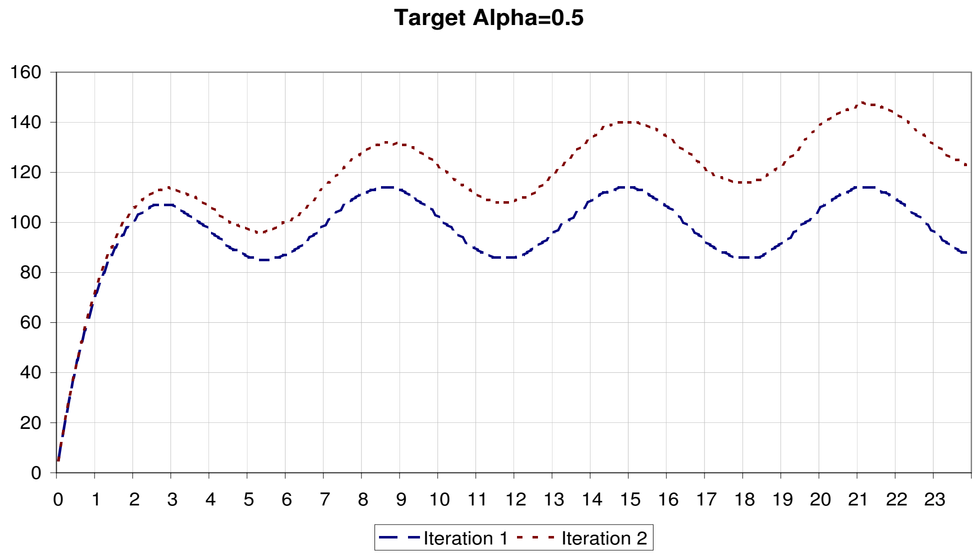
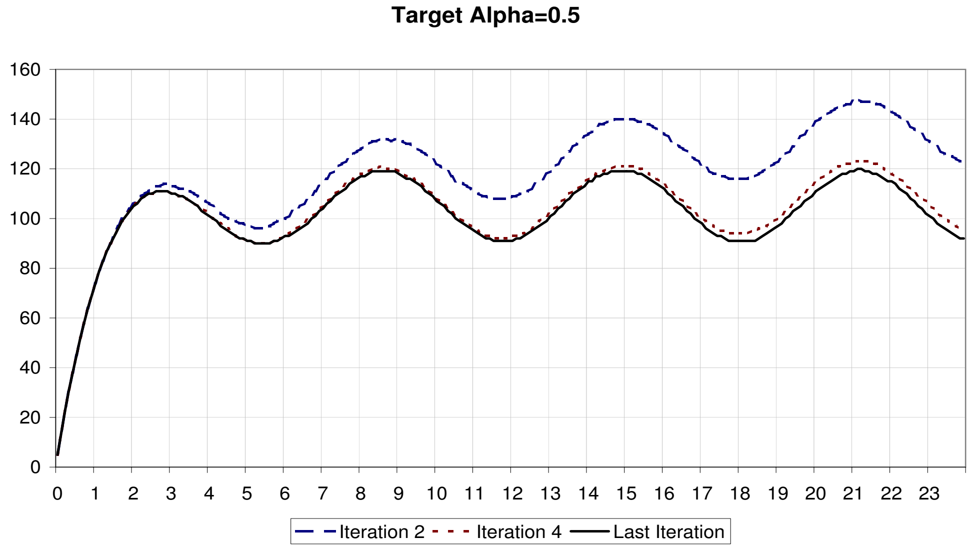


Figure 45: **Algorithm dynamics: evolution of convergence during algorithm runtime**



12. 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^{th} replication, the estimator is:

$$\hat{\alpha}_k(t) = \frac{\sum_i 1\{\mathit{customer\_i\_entered\_queue\_at\_interval\_t}\}}{\sum_i 1\{\mathit{customer\_i\_entered\_system\_at\_interval\_t}\}} \equiv \frac{\hat{Q}_k(t)}{\hat{S}_k(t)}\ . \tag{12.1}$$

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

$$\hat{w}_k(t) = \frac{\sum_i w_i 1\{\mathit{customer\_i\_entered\_system\_at\_interval\_t}\}}{\sum_i 1\{\mathit{customer\_i\_entered\_system\_at\_interval\_t}\}} \tag{12.2}$$

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 $1\{L_t^{(\infty)} - s_t^{(\infty)} \geq 5\}$ are averaged over all replications. (Here $L_t^{(\infty)}$ and $s_t^{(\infty)}$ are the values at time t obtained from the last iteration of ISA.)

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:

$$\hat{\rho}_k(t) = \frac{\sum_{i=1}^{s_t^{(\infty)}} b_i}{s_t^{(\infty)} \cdot \Delta} \tag{12.3}$$

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.

13. A Uniform-Acceleration Perspective

We can create a rigorous framework for the square-root-staffing formula 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.

13.1. Limits for a Family of Multi-Server Queues with Abandonment

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

$$\lambda^\eta_t = \eta \cdot \lambda_t \quad \text{and} \quad s^\eta_t = \eta \cdot s^{(f)}_t + \sqrt{\eta} \cdot s^{(d)}_t + o(\sqrt{\eta}\,). \tag{13.1}$$

(The superscripts f and d on $s^{(d)}_t$ and $s^{(d)}_t$ indicate the "fluid-approximation" term and the "diffusion-approximation" term, respectively.)

Unlike the uniform acceleration scalings that lead to the pointwise stationary approximation, this one is inspired by the scalings of Halfin and Whitt (1981), Garnett et al. (2002) and Mandelbaum, Massey and Reiman (1998). 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 η. By limit theorems developed in Mandelbaum, Massey and Reiman (1998), 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 13.1. *Consider the family of multiserver queues with abandonment having the growth conditions for its parameters as defined above. If we set*

$$s^\eta_t = \eta \cdot m_t + \sqrt{\eta} \cdot s^{(d)}_t + o(\sqrt{\eta}\,) \tag{13.2}$$

i.e., if we use (13.1) *with* $s^{(f)}_t = m_t$, *where*

$$\frac{d}{dt} m_t = \lambda_t - \mu_t \cdot m_t, \tag{13.3}$$

then

$$\lim_{\eta\to\infty} P\left(L_t^\eta \geq s_t^\eta\right) = p\left(L_t^{(d)} \geq s_t^{(d)}\right), \tag{13.4}$$

where $L^{(d)} = \left\{ L_t^{(d)} \mid t \geq 0 \right\}$ *is a diffusion process, which is the unique sample-path solution to the integral equation*

$$\begin{aligned} L_t^{(d)} &= L_0^{(d)} + \int_0^t (\mu_u - \theta_u) \cdot (s_u^{(d)})^- du \\ &\quad - \int_0^t \left(\theta_u \cdot (L_u^{(d)})^+ - \mu_u \cdot (L_u^{(d)})^-\right) du + B\left(\int_0^t (\lambda_u + \mu_u \cdot m_u) du\right) \end{aligned} \tag{13.5}$$

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 $L^{(d)}$. Notice that since λ_t and μ_t are given, then so is m_t. Thus server staffing for this model can only be controlled by the selection of $s^{(d)}$. Also notice that the diffusion $L^{(d)}$ is independent of $s^{(d)}$ *as long as* $\theta_t = \mu_t$ or $s_t^{(d)} \geq 0$ 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 MOL server-staffing heuristic.

Corollary 13.1. *If* $\theta = \mu$ *and* $s_t^\eta = \eta \cdot m_t + \Phi^{-1}(1-\alpha) \cdot \sqrt{\eta \cdot m_t}$, *where*

$$\frac{1}{\sqrt{2\pi}} \int_{\Phi^{-1}(1-\alpha)}^{\infty} e^{-x^2/2} dx = \alpha, \tag{13.6}$$

then we have

$$\lim_{\eta\to\infty} \left(L_t^\eta \geq s_t^\eta\right) = \alpha \tag{13.7}$$

for all $t > 0$.

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

Corollary 13.2. *The differential equation for the mean of* $L^{(d)}$ *is*

$$\frac{d}{dt} E\left[L_t^{(d)}\right] = (\mu_t - \theta_t) \cdot (s_t^{(d)})^- - \theta_t \cdot E\left[(L_t^{(d)})^+\right] + \mu_t \cdot E\left[(L_t^{(d)})^-\right]. \tag{13.8}$$

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

$$\begin{aligned} \frac{d}{dt} \mathsf{Var}\left[L_t^{(d)}\right] &= -2\theta_t \cdot \mathsf{Var}\left[(L_t^{(d)})^+\right] - 2\mu_t \cdot \mathsf{Var}\left[(L_t^{(d)})^-\right] \\ &\quad -2(\theta_t + \mu_t) \cdot E\left[(L_t^{(d)})^+\right] \cdot E\left[(L_t^{(d)})^-\right] + \lambda_t + \mu_t \cdot m_t. \end{aligned} \tag{13.9}$$

Proof of Theorem 13.1 . Define the function $f^{\eta}_t(\cdot)$, where

$$f^{\eta}_t(x) = \eta \cdot \lambda_t - \theta_t \cdot (\eta \cdot x - s^{\eta}_t)^+ - \mu_t \cdot (\eta \cdot x \wedge s^{\eta}_t). \tag{13.10}$$

Now we have

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

However

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

combining these results gives us the asymptotic expansion

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

as $\eta \to \infty$.

It follows that $f^{\eta}_t = \eta \cdot f^{(f)}_t + \sqrt{\eta} \cdot f^{(d)}_t + o(\sqrt{\eta})$, where

$$f^{(f)}_t(x) = \lambda_t - \theta_t \cdot (x - m_t)^+ - \mu_t \cdot (x \wedge m_t) \tag{13.11}$$

and

$$f^{(d)}_t(x) = (\theta_t - \mu_t) \cdot \left((s^{(d)}_t)^+ \cdot 1_{\{x>m_t\}} - (s^{(d)}_t)^- \cdot 1_{\{x \geq m_t\}} \right) . \tag{13.12}$$

Now

$$\Lambda f^{(f)}_t(x; y) = (\theta_t - \mu_t) \cdot \big(y \cdot 1_{\{x<m_t\}} - y^- \cdot 1_{\{x=m_t\}} \big) - \theta_t \cdot y \ , \tag{13.13}$$

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. Hence we have

$$\Lambda f^{(f)}_t(m_t; y) = \mu_t \cdot y^- - \theta_t \cdot y^+ \ \text{ and } \ f^{(d)}_t(m_t) = (\mu_t - \theta_t)(s^{(1)}_t)^- \tag{13.14}$$

Finally, we have

$$L^{(d)}_t = L^{(d)}_0 + \int_0^t \left(\Lambda f^{(f)}_t \left(m_u; L^{(d)}_u \right) + f^{(d)}_t \left(m_u \right) \right) du \tag{13.15}$$

$$
\begin{aligned}
& +B\left(\int_0^t (\lambda_u + \mu_u \cdot m_u) du\right) \\
= \;& L_0^{(d)} - \int_0^t \left(\theta_u \cdot ((L_u^{(d)})^+ + (s_u^{(d)})^-) - \mu_u \cdot ((L_u^{(d)})^- + (s_u^{(d)})^-)\right) du \\
& +B\left(\int_0^t (\lambda_u + \mu_u \cdot m_u) du\right). \qquad (13.16)
\end{aligned}
$$

13.2. Case 1: $\theta_t = \mu_t$ for all t

We then have

$$L_t^{(d)} = L_0^{(d)} - \int_0^t \mu_u \cdot L_u^{(d)} du + B\left(\int_0^t (\lambda_u + \mu_u \cdot m_u) du\right). \qquad (13.17)$$

It follows that $L^{(d)}$ is a zero-mean Gaussian process (if $L_0^{(d)} = 0$) and

$$\frac{d}{dt}\mathsf{Var}\left[L_t^{(d)}\right] = -2\mu_t \cdot \mathsf{Var}\left[L_t^{(d)}\right] + \lambda_t + \mu_t \cdot m_t. \qquad (13.18)$$

Moreover, if $m_0 = \mathsf{Var}\left[L_0^{(d)}\right]$, then $\mathsf{Var}\left[L_t^{(d)}\right] = m_t$ for all $t \geq 0$.

We remark that the simplification in this special case is to be expected, because we know from Section 10 that the $M_t/M_t/s_t + M_t$ model in this case reduces to the infinite-server $M_t/M_t/\infty$ model, which in turn - by making a time change - can be transformed into a $M_t/M/\infty$ model, for which the time-dependent distribution is known to be Poisson for all t, with the mean m_t in (2.5).

13.3. Case 2: $\theta_t = 0$

We then have

$$L_t^{(d)} = L_0^{(d)} + \int_0^t \mu_u \cdot \left((L_u^{(d)})^- + (s_u^{(d)})^-\right) du + B\left(\int_0^t (\lambda_u + \mu_u \cdot m_u) du\right). \qquad (13.19)$$

with

$$\frac{d}{dt}E\left[L_t^{(d)}\right] = \mu_t \cdot \left(E\left[(L_t^{(d)})^-\right] + (s_t^{(d)})^-\right) \qquad (13.20)$$

and

$$\frac{d}{dt}\mathsf{Var}\left[L_t^{(d)}\right] = -2\mu_t \cdot \left(\mathsf{Var}\left[(L_t^{(d)})^-\right] + E\left[(L_t^{(d)})^+\right] \cdot E\left[(L_t^{(d)})^-\right]\right) + \lambda_t + \mu_t \cdot m_t. \qquad (13.21)$$

To conclude this section, we summarise the implications for our proposal to staffing at the offered load in the QED regime. Here is the implication: Asymptotically, controlling the delay for this queue with abandonment is a *second order* staffing effort (selecting $s_t^{(d)}$) 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 for the multi-server queue without abandonment. This is also the one case where the diffusion $L^{(d)}$ is Gaussian.

14. Summary and Directions for Future Research

14.1. Summary

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 for the time-varying Erlang-A model (with or without abandonment), covering the ED, QD and QED operational regimes. All experiments were done with 9 target delay probabilities, ranging from $\alpha = 0.1$ (QD) to $\alpha = 0.9$ (ED).

In the process, we found that the square-root-staffing formula in (2.3) based on the modified-offered-load approximation (using arrival rate $\hat{\lambda}(t)$ in (2.9)) is remarkably effective too. As in Jennings et al., that approach is facilitated by applying many-server heavy-traffic limits, in particular, using the Garnett-Mandelbaum-Reiman function in (5.4).

Finally, we found that the simple approach of "staffing to the offered load" is remarkably effective in the QED regime (when $\alpha = 0.5$). That was substantiated time and again by having the ISA staffing function s_t fall on top of the time-dependent offered load (the infinite-server mean m_t in (2.5)). When the service times are short, as for the realistic example related to Figure 1 discussed in §9, the offered load m_t tends to agree closely with the PSA offered load $\lambda(t)E[S]$; then staffing to the offered load reduces to the "naive deterministic approximation": staffing to the PSA offered load $\lambda(t)E[S]$.

14.2. Next Steps

Here are some natural "next-steps":

1. As discussed in Section 5, for the time-varying Erlang-A model, 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. Of special interest for actual call centers are multi-class models with skill-based routing.

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 11, 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 - the one with $\theta = \mu$ in Section 10 - we have shown in §§10 and 13 that our staffing methods are asymptotically correct as the scale increases. In §13 we exploited the mathematical framework of service networks in Mandelbaum et. al.(1998). It would be nice to prove much more generally that, under proper scaling, the actual time-dependent probability of delay indeed converges to the specified target as scale increases.

References

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

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

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

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

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

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

[7] Green, L. V., Kolesar, P. J., Whitt, W. 2005. Coping with Time-Varying Demand when Setting Staffing Requirements for a Service System. Columbia University. Available at: http://www.columbia.edu/∼ww2040/Coping.pdf

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

[9] Ingolfsson, A. 2005. Modeling the $M(t)/M/s(t)$ Queue with an Exhaustive Discipline. Available at: $http://www.bus.ualberta.ca/aingolfsson/working_papers.htm$

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

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

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

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

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

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

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

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

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

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

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

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

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

[23] Wallace, R. B., Whitt, W. 2004. A Staffing Algorithm for Call Centers with Skill-Based Routing. Submitted for publication, 2004. Available at: http://www.columbia.edu/∼ww2040/recent.html

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

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

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

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

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

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