MANAGEMENT SCIENCE
DOI 10.1287/mnsc.1070.0821ec
pp. ec1–ec20

e-companion

ONLY AVAILABLE IN ELECTRONIC FORM

Electronic Companion—"Staffing of Time-Varying Queues to Achieve Time-Stable Performance" by Zohar Feldman, Avishai Mandelbaum, William A. Massey, and Ward Whitt, *Management Science*, DOI 10.1287/mnsc.1070.0821.

1. Overview

In §2 we indicate how key performance measures are defined and estimated. In §3 we investigate how ISA performs for the Erlang-C model, without customer abandoment. Except for the abandonment, it is the same model as in §4 of the paper. In §4 we return to the model with abandonment and consider the cases of more and less patient customers. Specifically, we let the abandonment rate be $\theta = 0.2$ and $\theta = 5.0$ instead of $\theta = 1.0$. We show that ISA and MOL are equally effective in these other scenarios. Finally, in §5 we present some asymptotic analysis that provides additional theoretical support. As stated in the paper, Feldman et al. (2005) is a longer unabridged version, e.g., containing 47 figures.

2. Estimating the Performance Measures

In this work we examine several performance measures. Since we have time-varying arrivals, care is needed in their definition and estimation. In this subsection we describe our estimation procedure. See §2 of the main paper for a description of the ISA algorithm.

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, thinking of time measured in hours. In our examples the mean service times was 1 hour. We make staffing changes every $\Delta = 0.1$ hour.

Given each staffing function, we estimate the time-dependent number of customers in the system by performing 5000 independent replications. For replication k, the **delay probability** in interval t, $\hat{\alpha}_k(t)$, is estimated by the number $\hat{Q}_k(t)$ of customers who cannot be served immediately upon arrival and thus join the queue divided by the number $\hat{S}_k(t)$ of arriving customers during the t time interval. We obtain the overall estimator $\hat{\alpha}(t)$ by averaging $\hat{\alpha}_k(t)$ 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 $\hat{w}_k(t)$ of the **average waiting time** in interval t is defined in an analogous way by the sum of the waiting times (until starting service) for all arrivals in that time interval divided by the total number of arrivals in that time interval. 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. For each replication, it is the actual value observed at the end of the time interval. The overall average queue length is 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, where $L_t^{(\infty)}$ and $s_t^{(\infty)}$ are the number in system and the staffing level at time t obtained from the last iteration of ISA.

For replication k, the estimator $\hat{\rho}_k(t)$ of the **server utilization** *in interval* t is the proportion of busy-servers during the time-interval, accounting for servers who are busy only a fraction of the interval:

$$\hat{\rho}_k(t) = \frac{\sum_{i=1}^{s_t^{(\infty)}} b_i}{s_t^{(\infty)} \cdot \Delta} \tag{2.1}$$

where b_i denotes the busy time of server i in interval t and Δ is the length of the time interval. Again, the overall estimator $\hat{\rho}(t)$ is the average over all replications.

3. The Time-Varying Erlang-C Model

For comparison with the experiments for the time-varying Erlang-A ($M_t/M/s_t + M$) model in §4 of the main paper, we now show the performance of ISA for the same system but without abandonment (with infinite patience) - the $M_t/M/s_t$ or time-varying Erlang-C model. As expected, the required staffing levels are higher than with abandonment, for all target delay probabilities; compare Figure 1 with Figure 2 in the paper. For example, for $\alpha = 0.5$, the maximum staffing level becomes about 120 instead of 115. An immediate conclusion is that it is important to include abandonment in the model when it is in fact present.

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, for the Erlang-C model the staffing tends to coincide with the offered load in the ED regime, when $\alpha = 0.9$, as opposed to in the QED regime, when $\alpha = 0.5$. That shows how abandonment allows greater efficiency, while still meeting the delay-probability target.

3.1. Time-Stable Performance

As before, we achieve accurate time-stable delay probabilities when we apply the ISA; see Figure 2, where again we consider target delay probabilities $0.1, 0.2, \ldots, 0.9$.

The empirical service quality β_t^{ISA} is stabilizing as well, as can be seen from Figure 3, which shows results for the same 9 target delay probabilities. As in Figure 5 in the paper, the empirical service quality decreases as the target delay probability increases. However, the empirical service quality β_t^{ISA} 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 3.

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 4, 5, 6 and 7. Precise explanations and definitions of the performance measures are given in Section 2.

Figures 3 and 6 show that here the time until system reaches (dynamic) steady-state is much longer compared to a system with abandonment. In fact, in Figure 6 steady-state was not yet reached after 24 time-units (the full day). Steady-state is approached much more quickly with abandonment; see Figure 8 of Feldman et al. (2005).

3.2. Validating the Square-Root-Staffing Formula

Just as for the time-varying Erlang-A model, we want to validate the square-root-staffing formula in (5) of the paper. We thus repeat the experiments we did with abandonment. 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 exceptionally well; see Figure 7. 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,

$$P(delay) \equiv \alpha \equiv \alpha(\beta) \approx \left[1 + \beta \cdot \frac{\Phi(\beta)}{\phi(\beta)}\right]^{-1}, \quad 0 < \beta < \infty \ , \tag{3.1}$$

where ϕ is again the pdf associated with the standard normal cdf Φ. The Halfin-Whitt function in (3.1) is obtained from the Garnett function in (10) of the paper 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 (5) 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 (5) 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 8 are again remarkable for $\beta > 0.25$.

3.3. 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. (We do not claim that abandonment, per se, is good. Instead, we claim that it is good to take account of it if it is in fact present.) When compared to the model without abandonment, abandonment in the model reduces the required staff. In Figure 9 we show the difference between staffing levels with and without abandonment 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 day (24-hour period). 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 day, for $\alpha = 0.1, 0.5, 0.9$ respectively. The labor savings increases as α increases.

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

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

4.1. More and Less Patient Customers

We consider two new cases (both with $\mu = 1$: $\theta = 0.2$; 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.

The performance of ISA is essentially the same as for the previous case with $\theta = 1.0$. We compare the staffing levels for these alternative environments, for the three regimes QD ($\alpha = 0.1$), ED ($\alpha = 0.9$) and QED ($\alpha = 0.5$) in Figure 10 below. In both these new cases, the target delay probabilities were achieved quite accurately for all target delay probabilities ranging from $\alpha = 0.1$ to $\alpha = 0.9$; see Figure 11. The implied empirical quality of service β_t^{ISA} defined in (9) of the paper is also stable, just as with $\theta = 1.0$; see Figure 12. We compare the time-dependent abandonment $P_t(Ab)$ in these two scenarios in Figure 13. 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 $\alpha = 0.9$.

We compare the empirical (α, β) pairs produced by ISA to the Garnett function in (10) of the paper for these two cases in Figure 14. 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 MOL approximation, here manifested in the square-root-staffing formula in (5)

of the paper, obtaining the required service quality β from the target delay probability α by using the inverse of the Garnett function in (10) of the paper, which reduces to the Half-Whitt function in (3.1) 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 15 below.

4.2. Benefits of Taking Account of Abandonment Again

Following §3.3, 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 16 we show the comparison for delay probability $\alpha = 0.5$, which we consider to be a reasonable operational target.

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

4.3. 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).

5. An Asymptotic Perspective

We can create a rigorous asymptotic framework for the square-root-staffing formula by considering the system as the arrival rate is allowed to increase. We can then apply the asymptotic analysis of uniform acceleration to multi-server queues with abandonment, as in Mandelbaum, Massey and Reiman (1998)..

The underlying intuition for optimal staffing is that, for large systems, we should staff exactly for the number of customers requesting service. That is, from a first-order deterministic-fluid-model perspective, abandonment does not happen at all. Thus the associated fluid model should not be a function of any abandonment parameters. The effect of abandonment should appear as second-order diffusion-model phenomenon. Thus, abandonment parameters should only contribute to 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.

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

In this section we will consider a family of Markovian $M_t/M/s_t + M$ models indexed by a parameter η. As before, we will focus on the stochastic process representing the number of customers in the system, which is a time-varying birth-and-death process. We will identify that stochastic process with the $M_t/M/s_t + M$ model.

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

$$\lambda_t^\eta = \eta \cdot \lambda_t \quad \text{and} \quad s_t^\eta = \eta \cdot s_t^{(f)} + \sqrt{\eta} \cdot s_t^{(d)} + o(\sqrt{\eta}\,). \tag{5.1}$$

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

Unlike the uniform acceleration scalings that lead to the pointwise stationary approximation, as in Massey and Whitt (1998), 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 5.1. *Consider the family of multiserver queues with abandonment having the growth conditions for its parameters as defined above. If we set*

$$s_t^\eta = \eta \cdot m_t + \sqrt{\eta} \cdot s_t^{(d)} + o(\sqrt{\eta}\,) \tag{5.2}$$

i.e., if we use (5.1) *with* $s_t^{(f)} = m_t$*, where*

$$\frac{d}{dt} m_t = \lambda_t - \mu_t \cdot m_t, \tag{5.3}$$

then

$$\lim_{\eta \to \infty} P\left(N_t^\eta \geq s_t^\eta\right) = P\left(N_t^{(d)} \geq s_t^{(d)}\right), \tag{5.4}$$

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

$$\begin{aligned} N_t^{(d)} &= N_0^{(d)} + \int_0^t (\mu_u - \theta_u) \cdot (s_u^{(d)})^- du \\ &\quad - \int_0^t \left(\theta_u \cdot (N_u^{(d)})^+ - \mu_u \cdot (N_u^{(d)})^-\right) du + B\left(\int_0^t (\lambda_u + \mu_u \cdot m_u) du\right) \end{aligned} \tag{5.5}$$

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 $N^{(d)}$. 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 $N^{(d)}$ 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 5.1. *If* $\theta = \mu$ *and* $s_t^\eta = \eta \cdot m_t + \Phi^{-1}(1-\alpha) \cdot \sqrt{\eta \cdot m_t}$*, where*

$$\frac{1}{\sqrt{2\pi}} \int_{\Phi^{-1}(1-\alpha)}^\infty e^{-x^2/2} dx = \alpha, \tag{5.6}$$

then we have

$$\lim_{\eta \to \infty} P\left(N_t^\eta \geq s_t^\eta\right) = \alpha \tag{5.7}$$

for all $t > 0$.

Unfortunately, $N^{(d)}$ in general is *not* a Gaussian process. This also means that the following set of differential equations are not autonomous. (A differential equation is said to be autonomous if the right-hand side does not involve the variable by which we are differentiating.)

Corollary 5.2. *The differential equation for the mean of* $N^{(d)}$ *is*

$$\frac{d}{dt}E\left[N_t^{(d)}\right] = (\mu_t - \theta_t)\cdot(s_t^{(d)})^- - \theta_t \cdot E\left[(N_t^{(d)})^+\right] + \mu_t \cdot E\left[(N_t^{(d)})^-\right]. \tag{5.8}$$

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

$$\begin{aligned}\frac{d}{dt}\mathsf{Var}\left[N_t^{(d)}\right] &= -2\theta_t \cdot \mathsf{Var}\left[(N_t^{(d)})^+\right] - 2\mu_t \cdot \mathsf{Var}\left[(N_t^{(d)})^-\right] \\ &\quad -2(\theta_t+\mu_t)\cdot E\left[(N_t^{(d)})^+\right]\cdot E\left[(N_t^{(d)})^-\right] + \lambda_t + \mu_t \cdot m_t.\end{aligned} \tag{5.9}$$

Proof of Theorem 5.1. Define the function $f_t^\eta(\cdot)$, where

$$f_t^\eta(x) = \eta\cdot\lambda_t - \theta_t\cdot(\eta\cdot x - s_t^\eta)^+ - \mu_t\cdot(\eta\cdot x \wedge s_t^\eta). \tag{5.10}$$

Now we have

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

However,

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

Combining these results, we get the asymptotic expansion

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

as $\eta\to\infty$.

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

$$f_t^{(f)}(x) = \lambda_t - \theta_t\cdot(x-m_t)^+ - \mu_t\cdot(x\wedge m_t) \tag{5.11}$$

and

$$f_t^{(d)}(x) = (\theta_t-\mu_t)\cdot\left((s_t^{(d)})^+\cdot 1_{\{x>m_t\}} - (s_t^{(d)})^-\cdot 1_{\{x\ge m_t\}}\right). \tag{5.12}$$

Now

$$\Lambda f_t^{(f)}(x;y) = (\theta_t-\mu_t)\cdot\left(y\cdot 1_{\{x<m_t\}} - y^-\cdot 1_{\{x=m_t\}}\right) - \theta_t\cdot y\ , \tag{5.13}$$

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

$$\Lambda f_t^{(f)}(m_t;y) = \mu_t\cdot y^- - \theta_t\cdot y^+ \quad\text{and}\quad f_t^{(d)}(m_t) = (\mu_t-\theta_t)(s_t^{(1)})^- \tag{5.14}$$

Finally, we have

$$\begin{aligned} N_t^{(d)} &= N_0^{(d)} + \int_0^t \left(\Lambda f_t^{(f)} \left(m_u; N_u^{(d)} \right) + f_t^{(d)} \left(m_u \right) \right) du \qquad (5.15) \\ &\quad + B \left(\int_0^t (\lambda_u + \mu_u \cdot m_u) du \right) \\ &= N_0^{(d)} - \int_0^t \left(\theta_u \cdot ((N_u^{(d)})^+ + (s_u^{(d)})^-) - \mu_u \cdot ((N_u^{(d)})^- + (s_u^{(d)})^-) \right) du \\ &\quad + B \left(\int_0^t (\lambda_u + \mu_u \cdot m_u) du \right). \qquad (5.16) \end{aligned}$$

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

We then have

$$N_t^{(d)} = N_0^{(d)} - \int_0^t \mu_u \cdot N_u^{(d)} du + B \left(\int_0^t (\lambda_u + \mu_u \cdot m_u) du \right). \qquad (5.17)$$

It follows that $N^{(d)}$ is a zero-mean Gaussian process (if $N_0^{(d)} = 0$) and

$$\frac{d}{dt} \mathsf{Var} \left[N_t^{(d)} \right] = -2\mu_t \cdot \mathsf{Var} \left[N_t^{(d)} \right] + \lambda_t + \mu_t \cdot m_t. \qquad (5.18)$$

Moreover, if $m_0 = \mathsf{Var} \left[N_0^{(d)} \right]$, then $\mathsf{Var} \left[N_t^{(d)} \right] = m_t$ for all $t \geq 0$.

We remark that the simplification in this special case is to be expected, because we know from §6 of the main paper 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) of the main paper.

5.3. Case 2: $\theta_t = 0$

We then have

$$N_t^{(d)} = N_0^{(d)} + \int_0^t \mu_u \cdot \left((N_u^{(d)})^- + (s_u^{(d)})^- \right) du + B \left(\int_0^t (\lambda_u + \mu_u \cdot m_u) du \right). \qquad (5.19)$$

with

$$\frac{d}{dt} E \left[N_t^{(d)} \right] = \mu_t \cdot \left(E \left[(N_t^{(d)})^- \right] + (s_t^{(d)})^- \right) \qquad (5.20)$$

and

$$\frac{d}{dt} \mathsf{Var} \left[N_t^{(d)} \right] = -2\mu_t \cdot \left(\mathsf{Var} \left[(N_t^{(d)})^- \right] + E \left[(N_t^{(d)})^+ \right] \cdot E \left[(N_t^{(d)})^- \right] \right) + \lambda_t + \mu_t \cdot m_t. \qquad (5.21)$$

Figure 1: **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)**

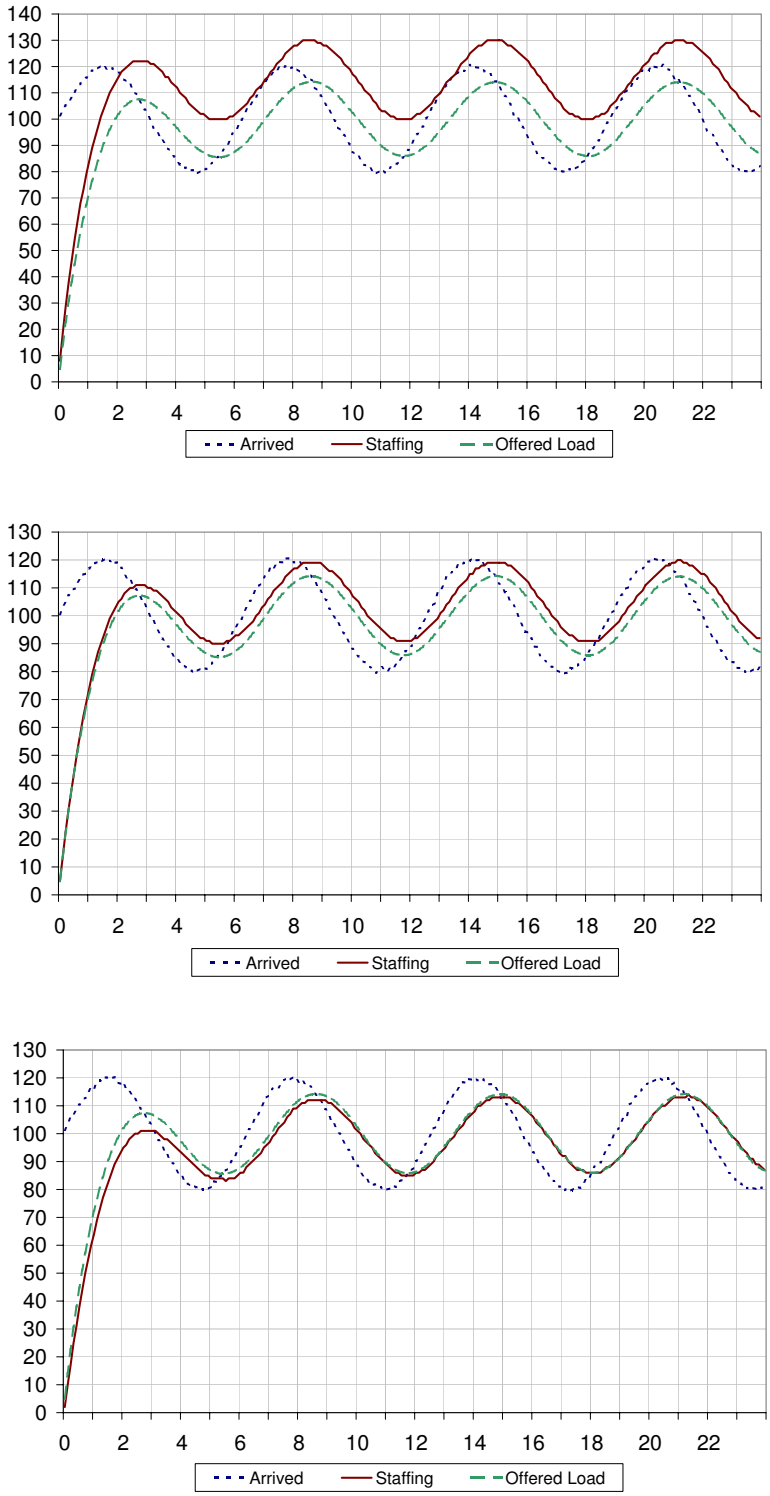


Figure 2: **Delay probability summary for the Erlang-C example**

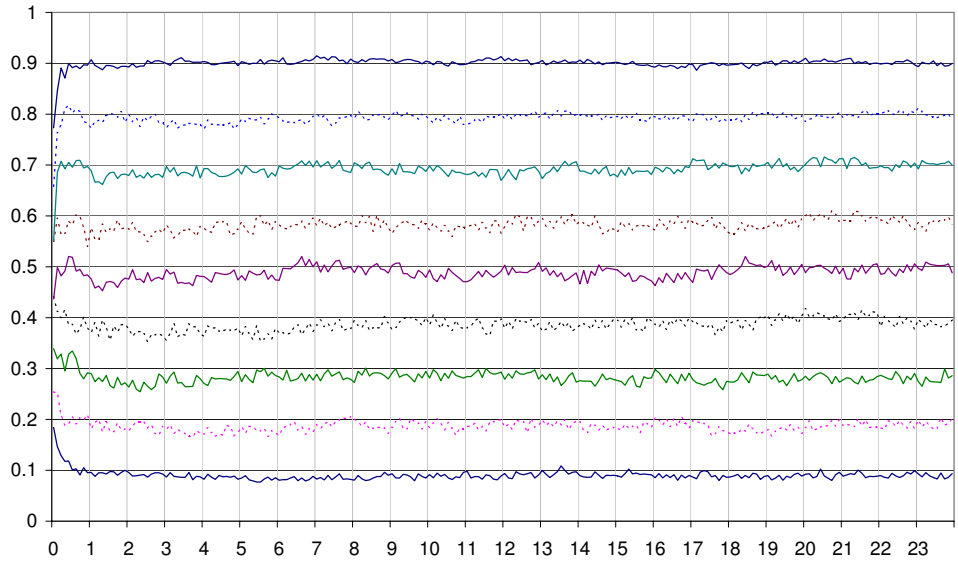


Figure 3: **Implied service quality β summary for the Erlang-C example** (The implied service quality decreases as α increases through the values 0.1, 0.2, ..., 0.9.)

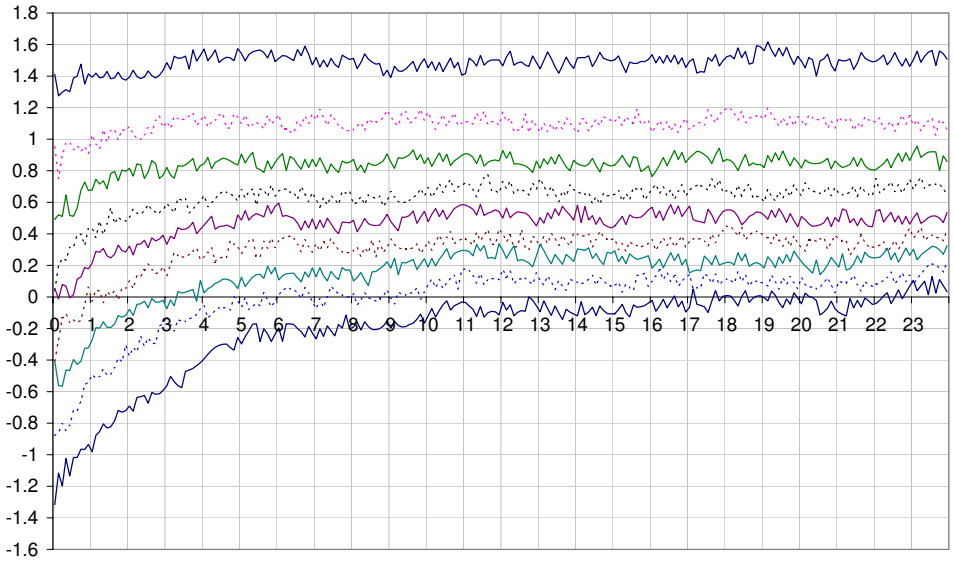


Figure 4: **Utilization summary for the Erlang-C example**

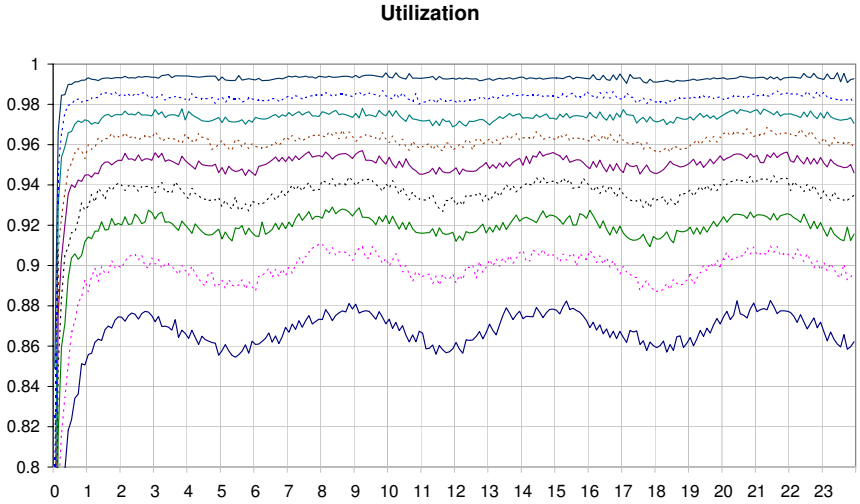


Figure 5: **Tail probability summary for the Erlang-C example**

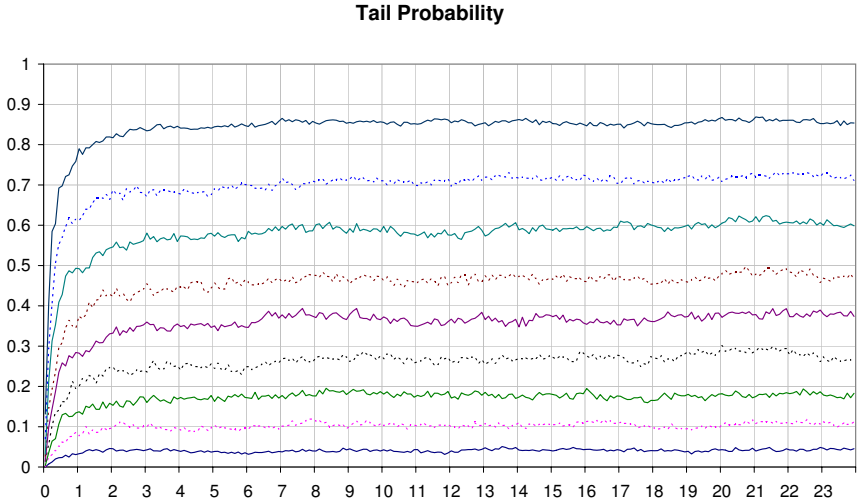


Figure 6: **Mean queue length and waiting time in the Erlang-C model with target α=0.5**

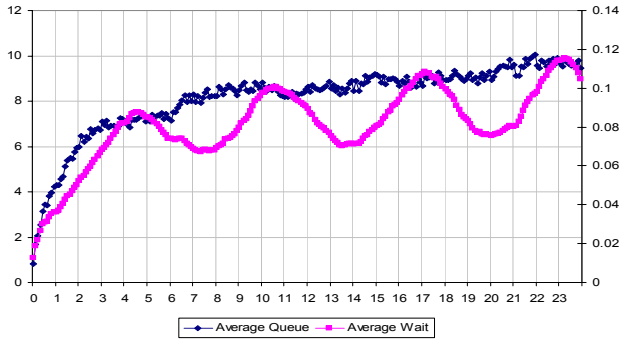


Figure 7: **The conditional distribution of the waiting time given delay in the Erlang-C model with target α=0.5**

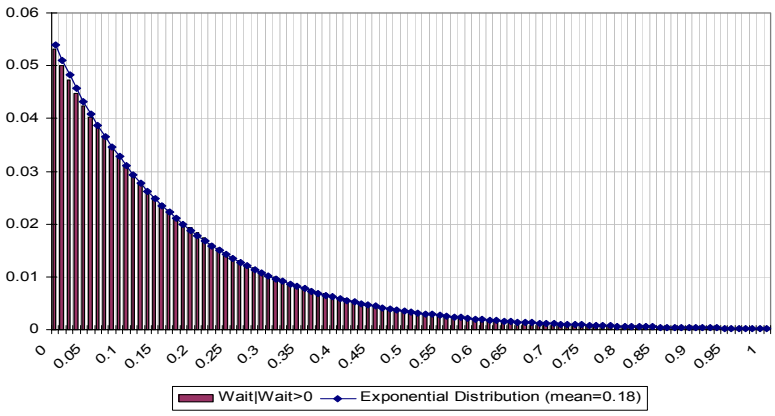


Figure 8: **Comparison of empirical results with the Halfin-Whitt approximation for the Erlang-C example**

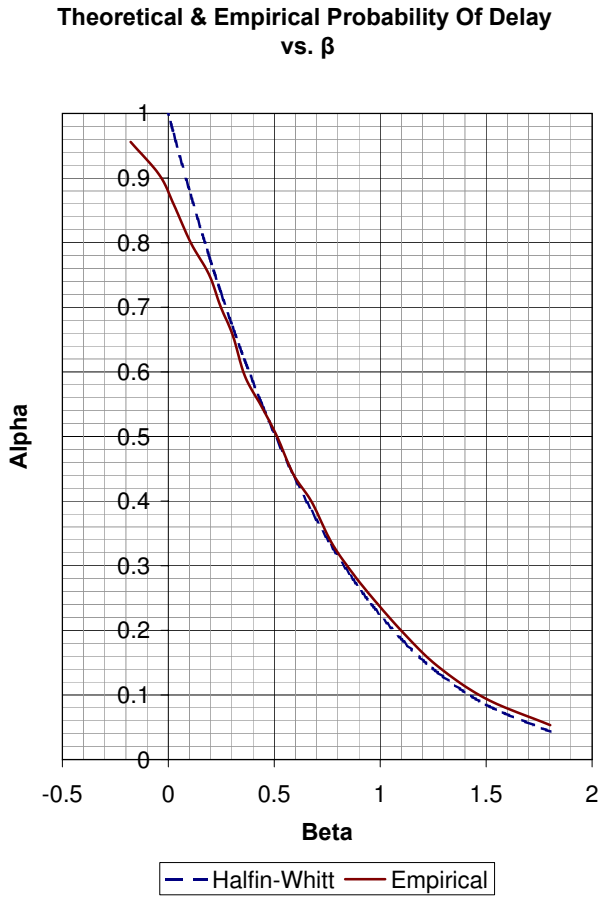
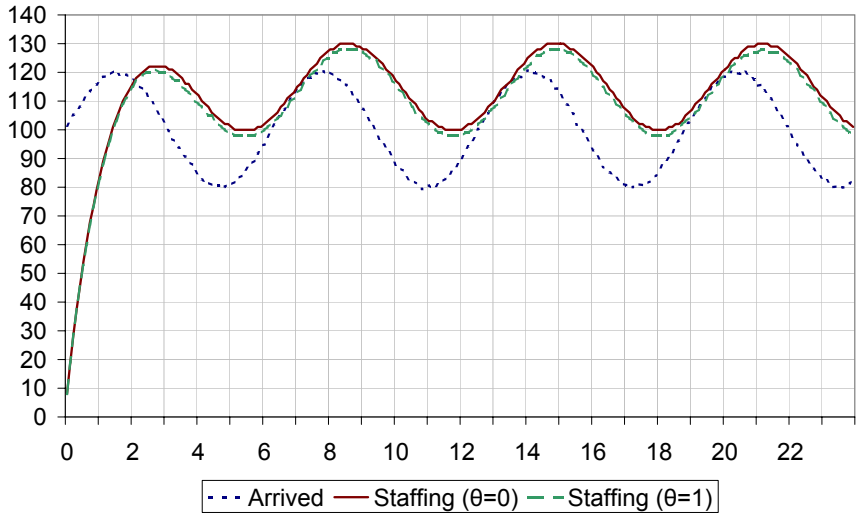
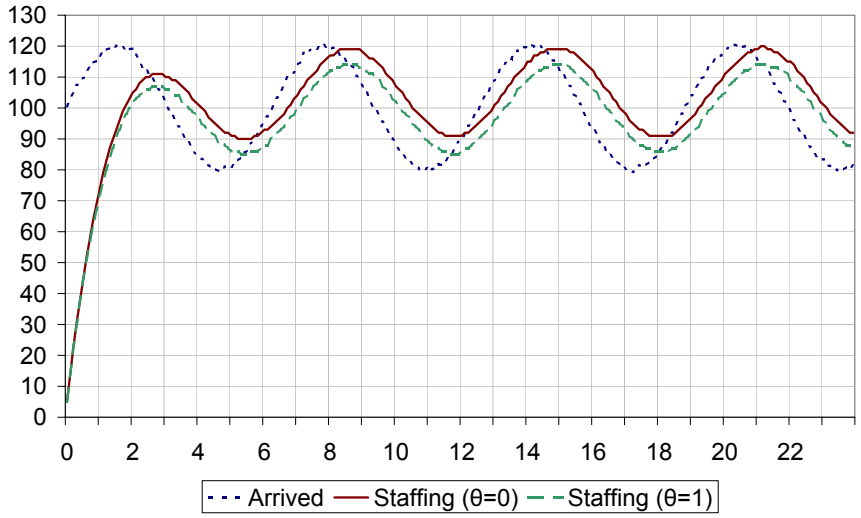


Figure 9: **Staffing levels with and without customer abandonment ($\theta = 1$ and $\theta = 0$): (1) $\alpha = 0.1$ (2) $\alpha = 0.5$ (3) $\alpha = 0.9$**





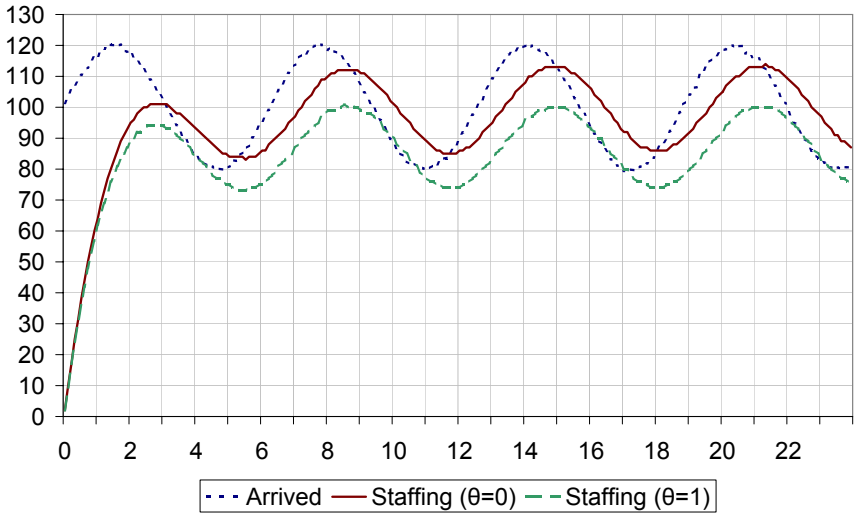
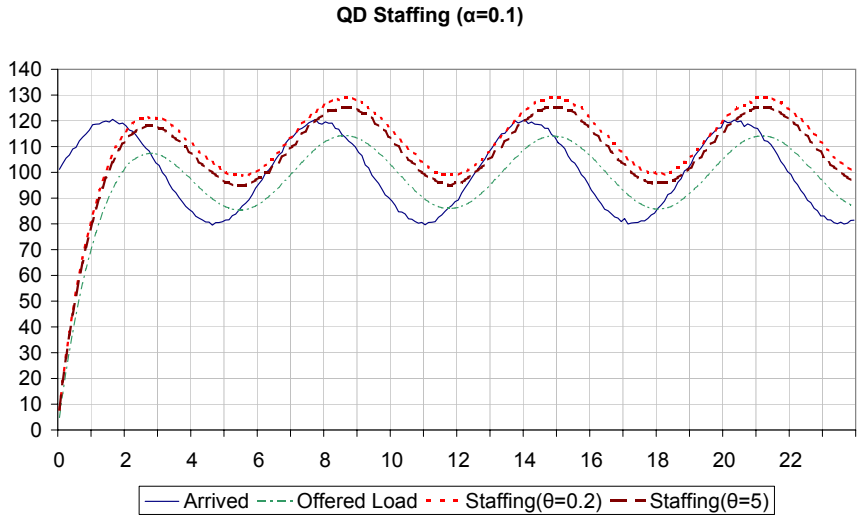
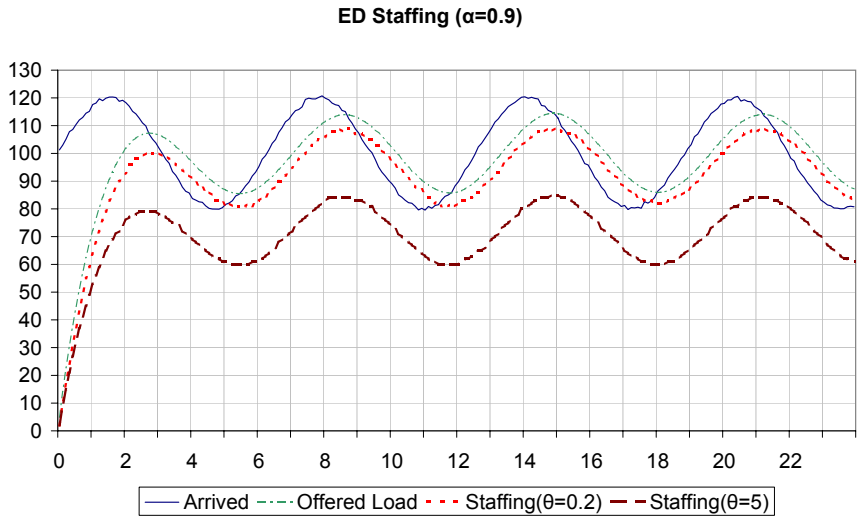


Figure 10: **Staffing for time-varying Erlang-A with more patient ($\theta = 0.2$) and less patient ($\theta = 5.0$) customers: (1) $\alpha = 0.1$ (QD), (2) $\alpha = 0.9$ (ED), (3) $\alpha = 0.5$ (QED)**





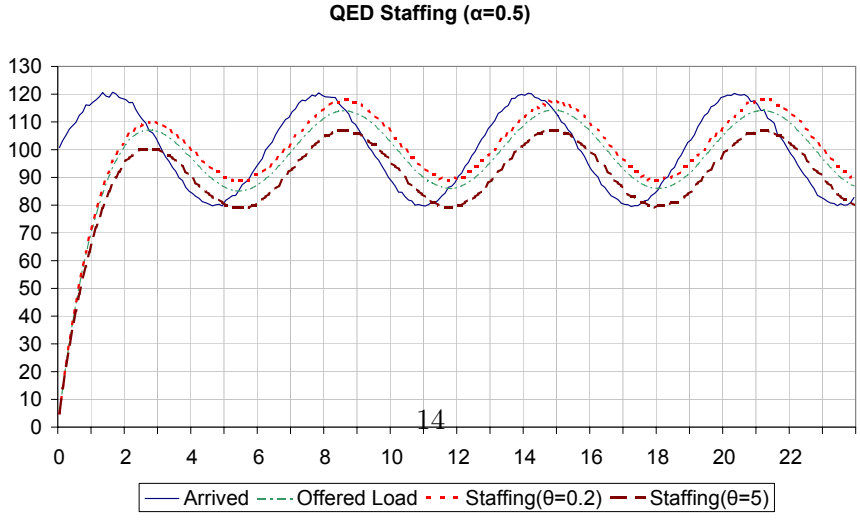
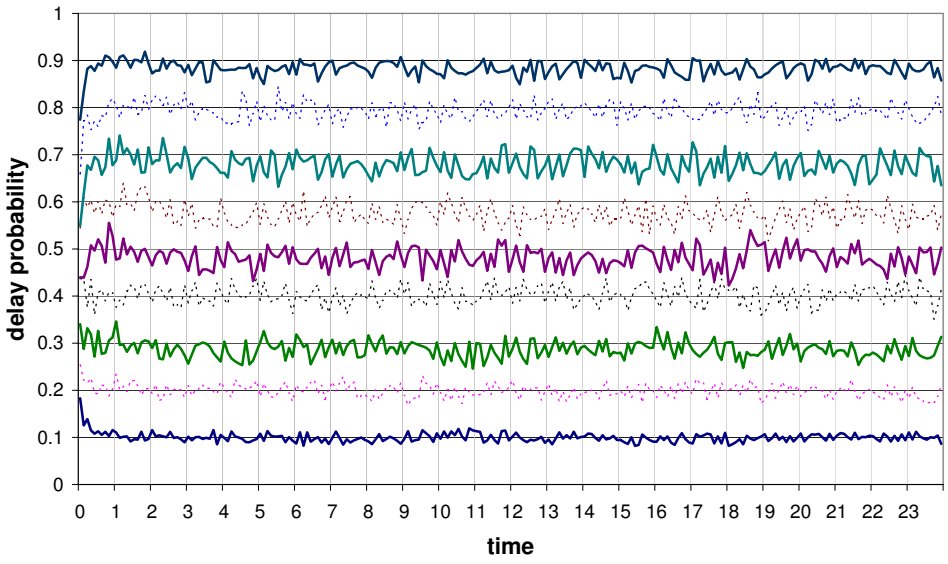


Figure 11: **Delay probabilities for the time-varying Erlang-A example with the new patience parameters: (1) θ=5 (2) θ=0.2**



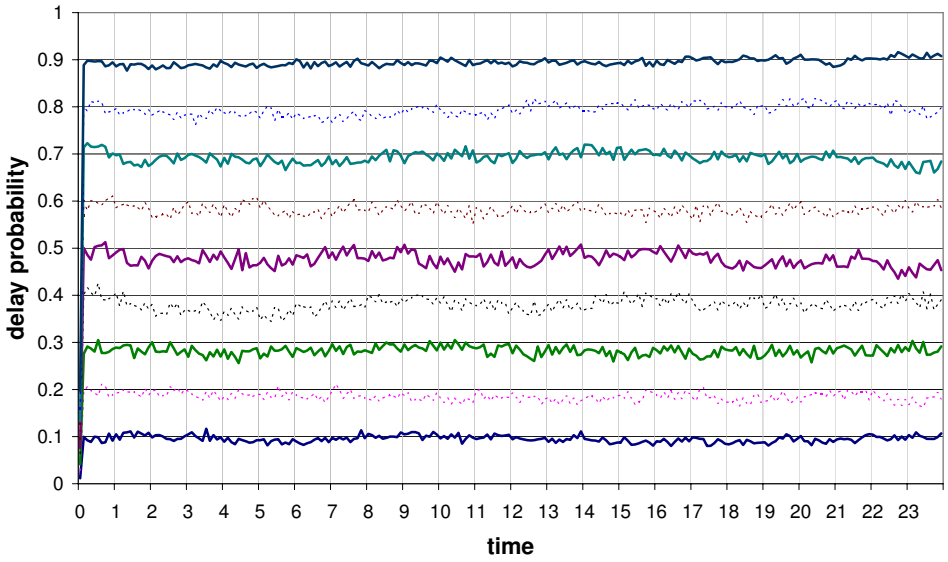
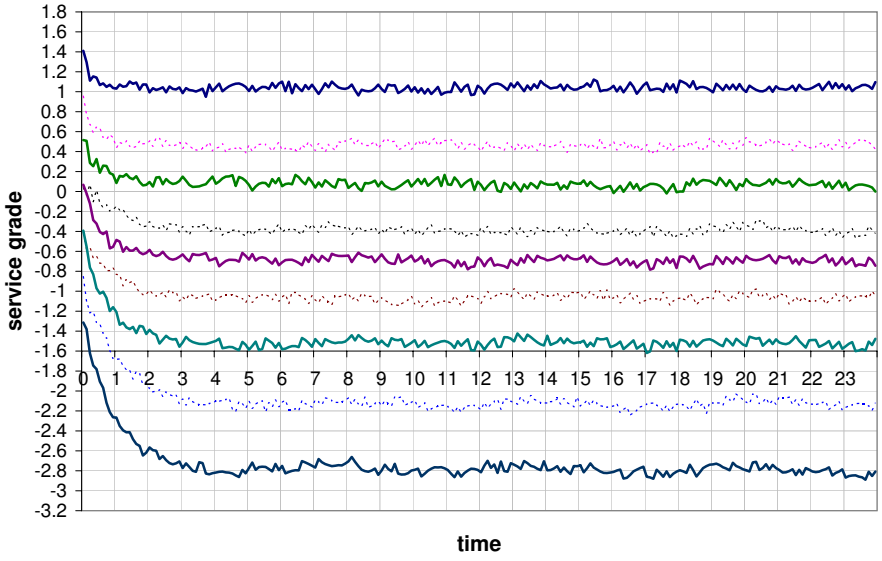


Figure 12: **Implied empirical quality of service β_t^{ISA} for the time-varying Erlang-A example with the new patience parameters: (1) θ=5 (2) θ=0.2**



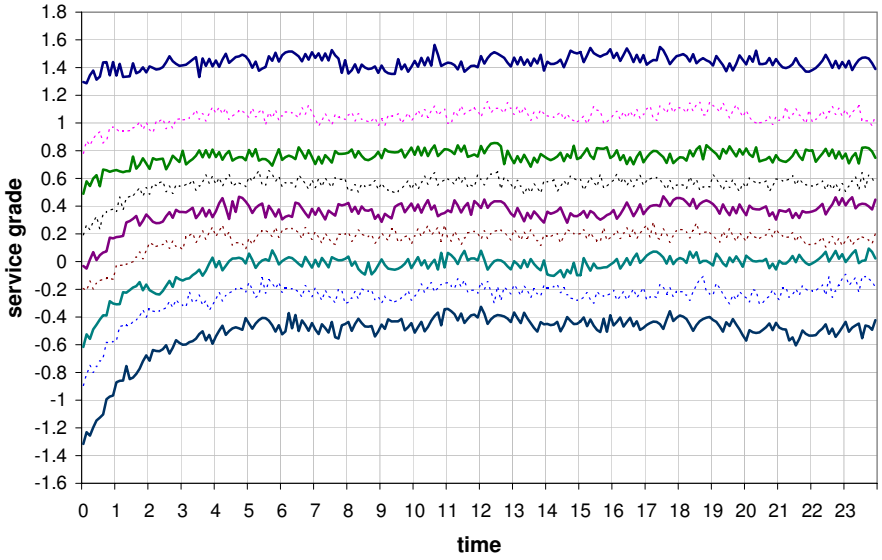
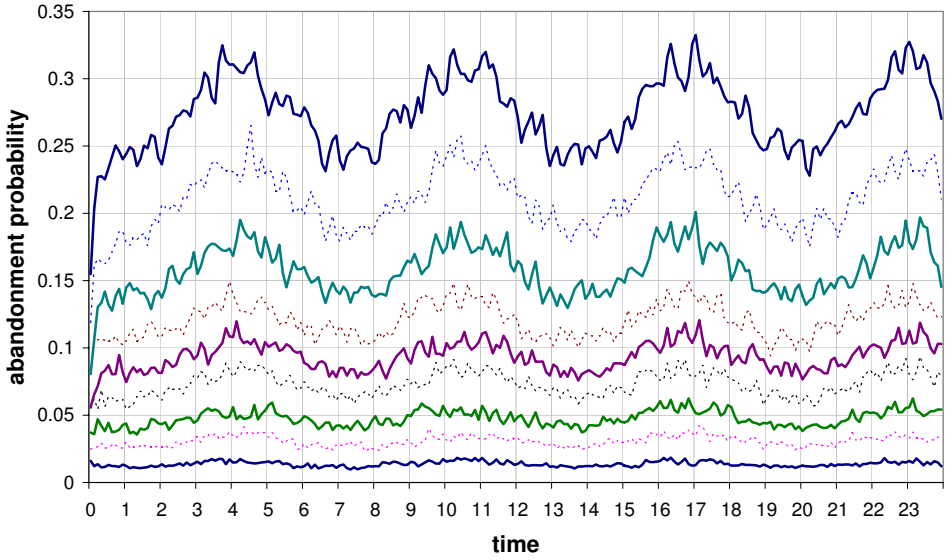


Figure 13: **Abandonment probabilities for the time-varying Erlang-A example with the new patience parameters: (1) θ=5 (2) θ=0.2**



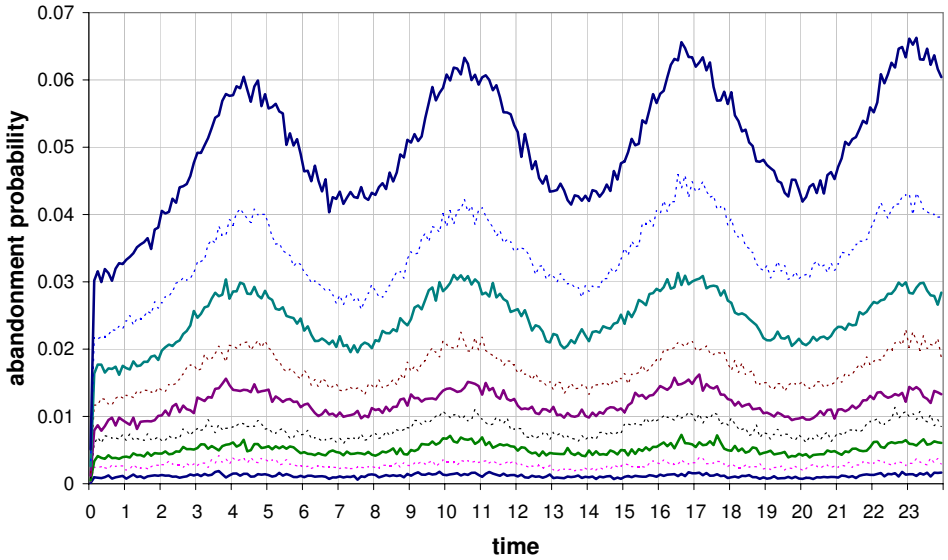


Figure 14: **Comparison of the empirical results from ISA with the Garnett approximation for the time-varying Erlang-A example with the new patience parameters: θ=5 and θ=0.2**

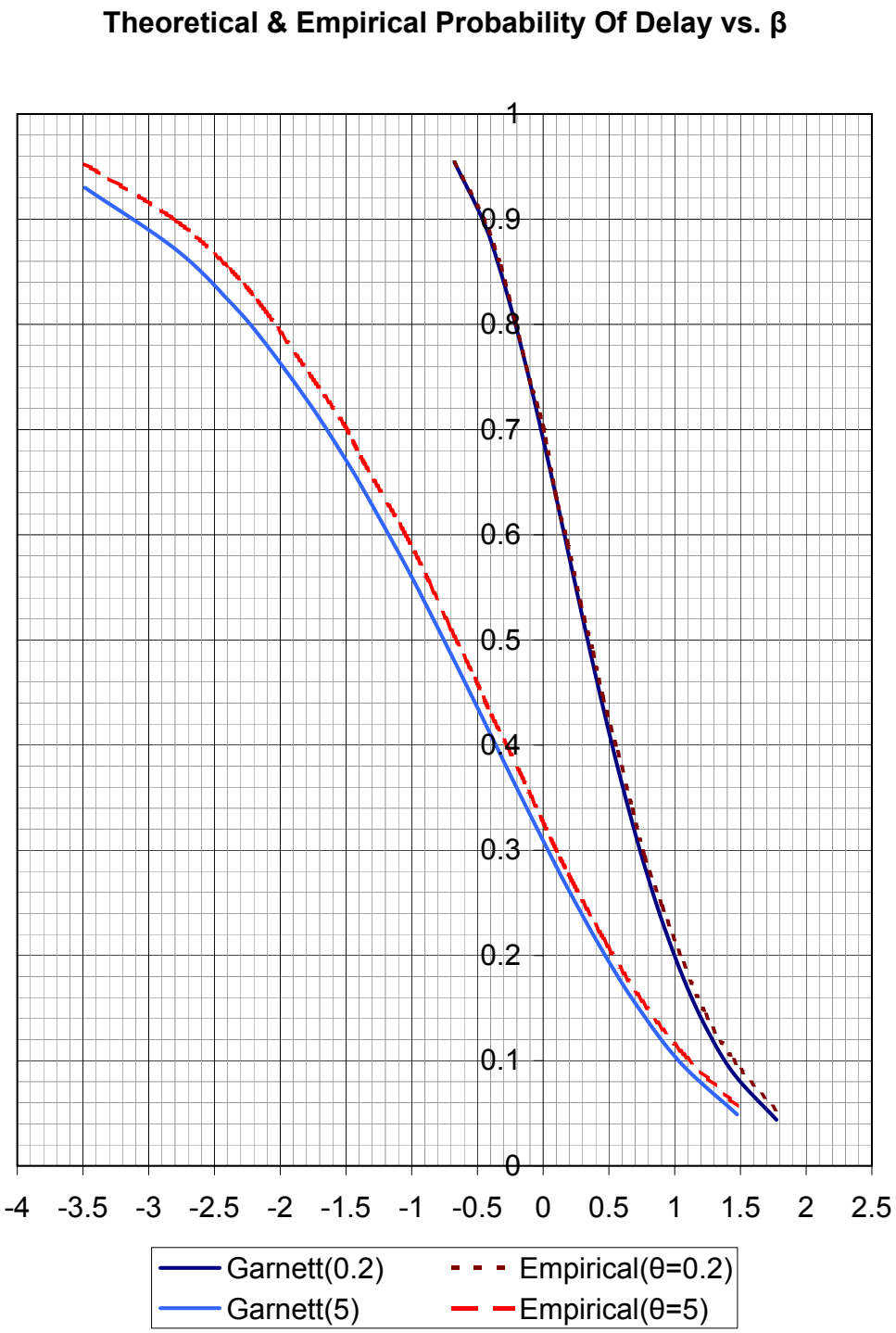


Figure 15: **The Halfin-Whitt/ Garnett functions**

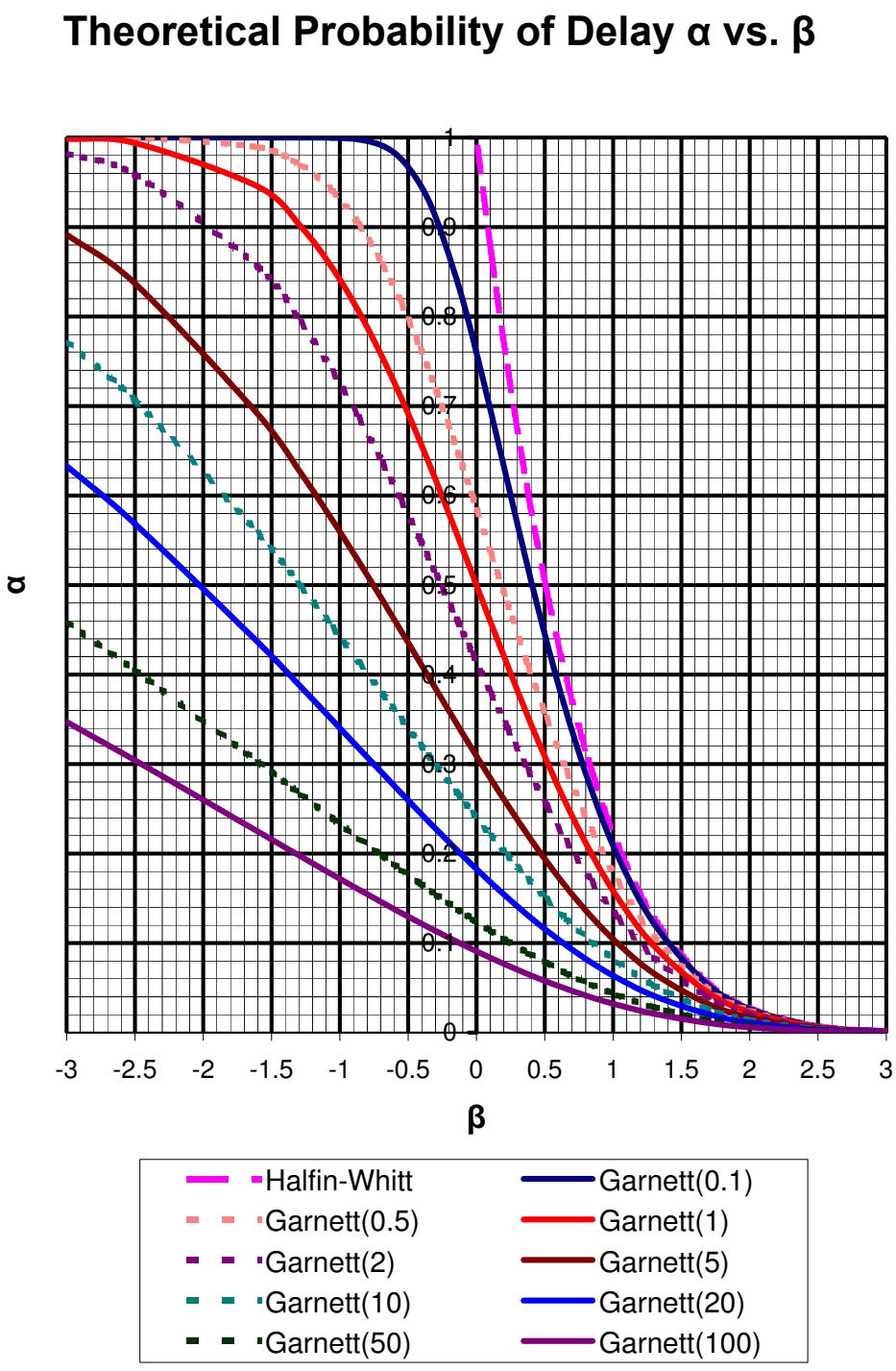


Figure 16: **Staffing levels for the time-varying Erlang-A example for a range of (im)patience parameters**