**This page is intentionally blank. Proper e-companion title page, with INFORMS branding and exact metadata of the main paper, will be produced by the INFORMS office when the issue is being assembled.**

# E-Companion

## EC.1. Overview

This e-companion has five more sections. In §EC.2 we extend §3 by giving explicit DIS performance formulas in structured special cases, when the arrival-rate function is sinusoidal and quadratic. In §EC.3 we indicate how the simulation estimates were made and give confidence intervals. In §EC.4 we present results of simulation experiments showing that the DIS-MOL staffing algorithm and the approximations for other performance measures are effective for $M_t/GI/s_t + GI$ models with non-exponential service-time and abandonment-time distributions. In §EC.5 we relate the DIS-MOL staffing for the special case of the $M_t/M/s_t + M$ model to a square root staffing formula. In §EC.6 we consider several discretization issues and real-world constraints for the DIS-OL and DIS-MOL staffing functions. First the staffing levels must be integer-valued, so they must be rounded. Second, when the staffing levels decrease, we do not remove a server until one completes the service in progress. Throughout we assume that server assignments can be switched when a server leaves, so that only the total number of servers matters. As a consequence, when the number of servers is scheduled to decrease by one when all servers are busy, then a server leaves at the time of the next service completion. In addition, we have considered specified staffing intervals, e.g., requiring that staffing changes can only be made on the half hour.

Additional material appears in a longer version available on the authors' web pages. In §EC.7, we supplement previous results in Jennings et al. (1996) and Feldman et al. (2008) showing that the pointwise stationary approximation does not perform well when the mean service time is relatively long. In §EC.8 we present the results from additional simulation experiments to show that the DIS-MOL approximation is effective for stabilizing the abandonment probability and the expected waiting time in $M_t/GI/s_t + GI$ models with non-exponential service-time and abandonment-time distributions. In §EC.9 we compare our SRS approximation for stabilizing the abandonment probability to the associated SRS formula developed by Feldman et al. (2008) for the delay probability, based on the Garnett function in (4). In §EC.10 we report simulation experiments for larger and smaller $M_t/M/s_t+M$ systems, specifically for the same sinusoidal arrival rate function in (18) for average arrival rates $a=20$, $a=50$ and $a=1000$; the main paper considered the case $a=100$. Finally, to put the staffing issue in perspective, in §EC.11 we study the sensitivity of the performance to a change of a single server. To do so, we give numerical results for the stationary $M/M/s+M$ model. These show that errors in meeting the target may be partly due to the impact of a single server.

## EC.2. DIS Approximations in Structured Special Cases

Since many service systems have daily cycles, it is natural to consider sinusoidal and other periodic arrival-rate functions, as was done in Jennings et al. (1996), Feldman et al. (2008). When we do so, we can apply Eick et al. (1993b) to obtain explicit formulas for the performance measures in that setting. For periodic arrival processes, it is natural to focus upon the dynamic steady state, which prevails because we have started the system empty at time $t=-\infty$. The position within the cycle is determined by simply defining the arrival-rate function over the entire real line, and assuming that it is periodic. In that way, time 0 corresponds to a definite place within a cycle.

Theorem EC.1. *Consider the DIS-OL approximation for the $M_t/GI/s_t+GI$ model with sinusoidal arrival-rate function $\lambda(t)=a+b\cdot\sin(ct)$ and delay target $w$. Then $Q(t)$ and $B(t)$ are independent Poisson random variables having sinusoidal means*

$$\begin{aligned} E[Q(t)] &= E[T]\,(a+\gamma\sin(ct-\theta)),\\ &\qquad for\ \theta\equiv\arctan[\phi_1(T)/\phi_2(T)],\quad \gamma\equiv b\sqrt{\phi_1(T)^2+\phi_2(T)^2},\\ E[B(t)] &= \bar{F}(w)E[S]\left(a+\tilde{\gamma}\sin[c(t-w)-\tilde{\theta}]\right),\\ &\qquad for\ \tilde{\theta}\equiv\arctan[\phi_1(S)/\phi_2(S)],\quad \tilde{\gamma}\equiv b\sqrt{\phi_1(S)^2+\phi_2(S)^2}, \end{aligned}$$

*where $\phi_1(X)\equiv E[\sin(cX_e)]$, $\phi_2(X)\equiv E[\cos(cX_e)]$ for a nonnegative random variable $X$.*

It is easy to see that the extreme values of $E[Q(t)]$ and $E[B(t)]$ occur at $t_Q=t_\lambda+\theta/c$ and $t_B=t_\lambda+\tilde{\theta}/c$, where $t_\lambda=\pi/2\gamma+n\pi/\gamma$ for $n$ integer are times at which the extreme values of $\lambda(t)$ occurs. And their extreme values are $Q(t_Q)=E[T](a\pm\gamma)$ and $B(t_B)=\bar{F}(w)E[T](a\pm\tilde{\gamma})$. As shown in Eick et al. (1993b), much nicer formulas are obtained in the special case of exponential service times, but we do not display them.

As discussed in §4 of Eick et al. (1993a), Massey and Whitt (1997) and §4 of Green et al. (2007), for general arrival-rate functions, it is convenient to use simple approximations stemming from linear and quadratic approximations formed by Taylor series approximations. These generate simple time lags and space shifts.

THEOREM EC.2. *Consider the DIS-OL approximation for the $M_t/GI/s_t + GI$ model with quadratic arrival-rate function $\lambda(t) = a + bt + ct^2$ and delay target $w$. Then $Q(t)$ and $B(t)$ are independent Poisson random variables having quadratic means*

$$E[Q(t)] = E[T]\lambda(t - E[T_e]) + cE[T]Var[T_e],$$

$$E[B(t)] = \bar{F}(w)E[S]\lambda(t - w - E[S_e]) + c\bar{F}(w)E[S]Var[S_e]$$

The quadratic case shows a deterministic time shift which corresponds to interchanging the order of expectation in Theorem 1, and a space shift that is the variance term multiplied by a constant. We obtain expressions for linear arrival-rate functions by simply letting the quadratic coefficient be $c = 0$.

## EC.3. More about the Simulation Experiments

We now provide more details about the experiment with the $M_t/M/s_t + M$ example in §6.

### EC.3.1. Estimating Performance Measures

We first describe our estimation procedure for the following time-dependent performance measures: (i) the mean potential waiting time, $E[W(t)]$, (ii) the abandonment probability, $P_t(Ab)$, (iii) the delay probability, $P_t(Delay)$, and (iv) the mean queue length, $E[Q(t)]$.

We estimated these performance measures in a time interval $[0,T]$ with $T = 20$. First, for $E[W(t)]$, we kept track of all customer arrivals in each sample path. For a customer $n$, we keep track of the arrival time, $A_n$, and the time that the customer enters service, $E_n$. Therefore, one value for this sample path is $(t,\hat{W}(t))=(A_n,E_n-A_n)$. Of course, this customer may have already abandoned by time $E_n$. Since we are interested in the *potential waiting time*, assuming infinite patience, we keep track of the time that the customer would enter service even after it abandons; i.e., our procedure includes the behavior of virtual customers.

We divide the interval $[0,20]$ into subintervals or "bins." The bin size for $E[W(t)]$ is 0.1. For the performance measures $P_t(Ab)$ and $P_t(Delay)$, we used a larger bin size because there are less sample points. In particular, we used a bin size of 0.25. At each sample path, for the interval $[0,0.25]$ (the procedures remain the same for other intervals $[0.25,0.5],\ldots,[19.75,20]$), we obtained an estimate by computing the ratios $A(0,0.25)/N(0,0.25)$ and $D(0,025)/N(0,0.25)$, where $A(t,t+\Delta t)$ is the total number of abandonments among arrivals in interval $[t,t+\Delta t]$, $D(t,t+\Delta t)$ is the number of delays encountered by arrivals in interval $[t,t+\Delta]$, and $N(t,t+\Delta t)$ is the total number of arrivals in $[t,t+\Delta t]$. For the mean queue length, $E[Q(t)]$, we used a bin size 0.05. In particular, we sampled the queue length once every 0.05 units of time.

### EC.3.2. Confidence Intervals

We estimated the above performance measures by generating independent replications. For the example in §6 we used 5000 independent sample paths To verify the statistical precision of our estimates, we also provide the 95% confidence intervals. Let $z_{0.05}\equiv 1.96$ be the 95% percentile of the standard normal distribution, the confidence interval is $[\bar{\mu}-1.96\,S_n/\sqrt{n},\bar{\mu}+1.96\,S_n/\sqrt{n}]$, where $\bar{\mu}\equiv(1/n)\sum_{i=1}^{n}X_i$ is the sample mean and $S^2\equiv 1/(n-1)\sum_{k=1}^{n}(X_i-\bar{\mu})^2$ is the sample variance. (Since the sample size is so large, the $t$ distribution associated with unknown variance is approximately normal.) Let $\epsilon_{P(AB)}$ and $\epsilon_{E[W]}$ be the half length of the confidence intervals. These are shown in the Table EC.3.2 for different targets $\alpha^*$ and $w^*$. Since the performance is stabilized, these apply approximately at each $t$. The values in Table EC.3.2 are averages over $t\in[0,T]$. The plotted performance functions appear even smoother, because there is dependence among successive values.

| $\alpha^*$ | 0.2 | 0.15 | 0.1 | 0.05 | 0.02 | 0.01 | 0.005 |
|---|---|---|---|---|---|---|---|
| $\bar{\mu}_{P(Ab)}$ | 0.20 | 0.156 | 0.0899 | 0.0547 | 0.0202 | 0.0098 | 0.0051 |
| $\epsilon_{P(Ab)}$ | 0.0045 | 0.0042 | 0.0034 | 0.0029 | 0.0012 | 0.0008 | 0.00056 |
| $w^* = F^{-1}(\alpha^*)$ | 0.45 | 0.033 | 0.21 | 0.11 | 0.041 | 0.020 | 0.010 |
| $\bar{\mu}_{E[W(t)]}$ | 0.449 | 0.343 | 0.194 | 0.129 | 0.0388 | 0.0183 | 0.0093 |
| $\epsilon_{E[W(t)]}$ | 0.0186 | 0.0162 | 0.0119 | 0.0099 | 0.0054 | 0.0037 | 0.0027 |

**Table EC.1** Confidence intervals for estimates of $E[W(t)]$ and $P_t(Ab)$ at different targets $\alpha^*$ and $w^*$ based on 5000 independent replications.

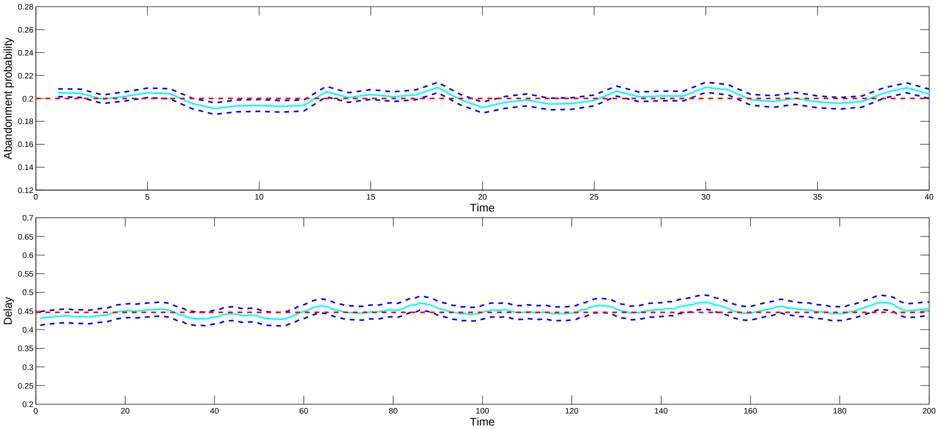


**Figure EC.1** Time-dependent confidence interval for $E[W(t)]$ and $P_t(Ab)$ with $\alpha^* = 0.2$.

To illustrate the consequence, in Figure EC.1 we also plot the variance envelope (i.e., confidence interval) for $\alpha = 0.2$ . The red dashed curves are the targets, the solid light blue curves are means (i.e., mid points of the confidence intervals) and the dashed blue curves are the confidence intervals.

## EC.4. Non-Exponential Service and Patience Distributions

The DIS-MOL approximation also performs well for the more general $M_t/GI/s_t + GI$ model with non-exponential service and patience distributions. We illustrate by displaying results from a simulation experiment for the $M_t/H_2/s_t + E_2$ model.

Let $A$ and $S$ be the generic abandon and service times. Here we fix their means $E[A] \equiv 1/\theta$ and $E[S] \equiv 1/\mu$ with $\mu \equiv 1$ and $\theta \equiv 0.5$. For the Erlang distribution, we consider Erlang-2 ($E_2$), the sum of two i.i.d. exponential random variables, which has squared coefficient of variation (SCV) $c_X^2 \equiv Var(X)/E[X]^2 = 1/2$. For the Hyper-exponential ($H_2$) distribution, $X$ is a mixture of two

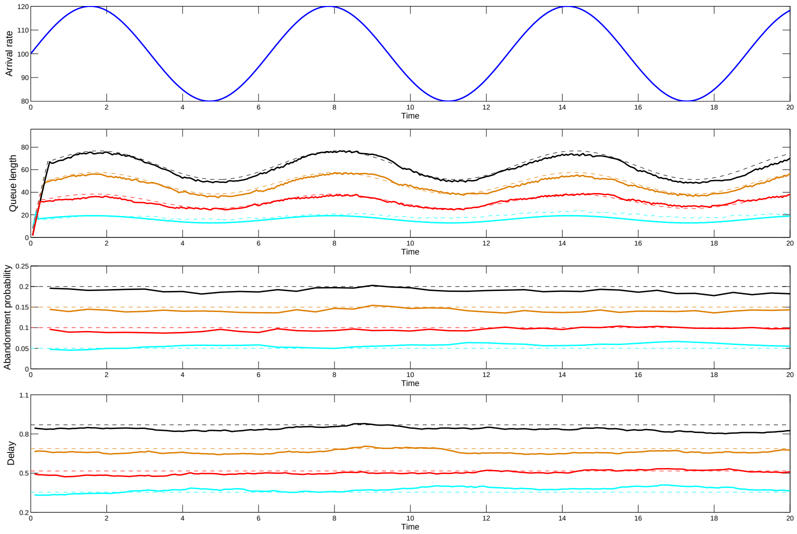


**Figure EC.2** Simulation estimates of expected queue length, abandonment probability and expected delay for the $M_t/H_2/s_t + E_2$ under the DIS-OL staffing, with the system heavily loaded $(0.05 \le \alpha \le 0.20)$.

independent exponential random variables, i.e., $X$ has complementary cdf $1 - G(x) = p \cdot e^{-\lambda_1 x} + (1-p) \cdot e^{-\lambda_2 x}$. We choose $p = 0.5(1 - \sqrt{0.6})$, $\lambda_1 = 2p/E[X]$, $\lambda_2 = 2(1-p)/E[X]$, which yields $c_X^2 = 4$ (and balanced means).

We use the same sinusoidal arrival rate function in (18) with the same parameters: $a = 100$, $b = 20$, $c = 1$. Just as for the Markovian example in §6, here we plot the simulation estimations of

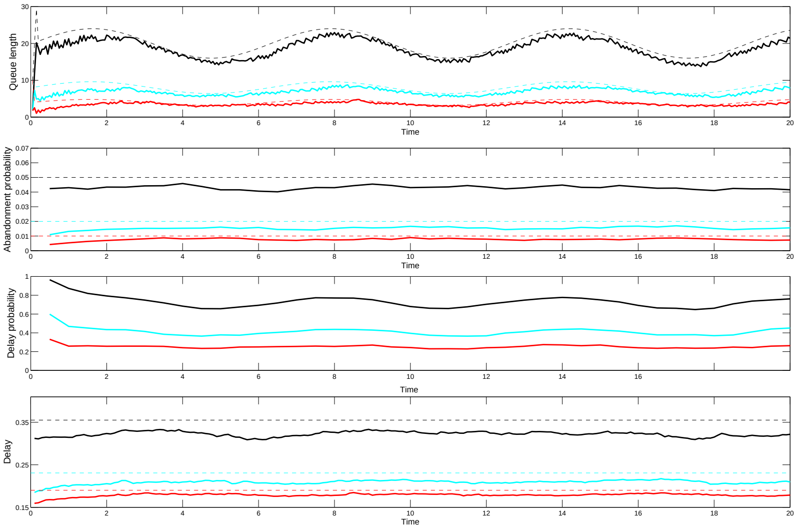


**Figure EC.3** Simulation estimates of expected queue length, abandonment probability and expected delay for the $M_t/H_2/s_t + E_2$ under the DIS-MOL staffing, with the system lightly loaded $(0.01 \le \alpha \le 0.05)$.

the expected queue lengths, abandonment probabilities and expected delays for different abandon probability targets $\alpha$ in a 20-hour day for the $M_t/D/s_t + H2$. Just as before, our estimates are based on 5000 independent replications. Estimates of the confidence intervals appear in Table EC.4.

| $\alpha^*$ | 0.2 | 0.15 | 0.1 | 0.05 | 0.02 | 0.01 | 0.005 |
|---|---|---|---|---|---|---|---|
| $\bar{\mu}_{P(Ab)}$ | 0.1904 | 0.1408 | 0.0931 | 0.0506 | 0.0191 | 0.0093 | 0.0041 |
| $\epsilon_{P(Ab)}$ | 0.0024 | 0.0022 | 0.0020 | 0.0015 | 0.0011 | 0.0009 | 0.0007 |
| $w^* = F^{-1}(\alpha^*)$ | 0.825 | 0.684 | 0.532 | 0.356 | 0.215 | 0.149 | 0.104 |
| $\bar{\mu}_{E[W(t)]}$ | 0.79 | 0.64 | 0.49 | 0.32 | 0.201 | 0.138 | 0.093 |
| $\epsilon_{E[W(t)]}$ | 0.0245 | 0.0221 | 0.0191 | 0.0153 | 0.0127 | 0.0103 | 0.0092 |

**Table EC.2** Confidence intervals for estimates of $E[W(t)]$ and $P_t(Ab)$ of the $M_t/H_2/s_t + E_2$ model, at different targets $\alpha^*$ and $w^*$ based on 5000 independent replications.

## EC.5. A New Square-Root Staffing (SRS) Formula

At the end of §6, we introduced the new SRS formula (21) for the $M_t/GI/s_t + GI$ model based on the DIS OL with abandonment target $\alpha$. Formula (21) differs from (2) by using the DIS OL $m_\alpha(t)$ instead of the standard OL $m_0(t)$ in (1).

We now verify that the DIS-MOL approximation is indeed consistent with the SRS staffing in (21); i.e., we show that the DIS-MOL staffing function $s_\alpha^{MOL}(t)$ has the form of the SRS formula (21). Following Feldman et al. (2008), for an abandonment probability target $\alpha$, we let $D_\alpha(t) \equiv s_t^{MOL} - m_\alpha(t)$ be the difference between the DIS-MOL staffing and the OL functions, and let $\beta_\alpha(t) \equiv D_\alpha(t)/\sqrt{m_\alpha(t)}$ be the *implicit QoS function* for DIS-MOL. In Figure EC.4, we plot the implicit QoS function $\beta_\alpha(t)$ for different targets $\alpha$ for the Markovian example with sinusoidal arrival rate in §6. Figure EC.4 shows that, just as in Feldman et al. (2008), except for an initial transience, $\beta_\alpha(t)$ is independent of time and decreasing in $\alpha$.

Figure EC.4 shows that there exists a QoS function $\beta_\alpha$, which is positive, strictly decreasing in $\alpha$ with the other parameters fixed, that makes the SRS function perform as well as DIS-MOL in stabilizing the abandonment probability and the expected waiting time. For all $\alpha \geq 0.1$, we found that $\beta_\alpha \approx 0$, consistent with Figure 3 and our observation more generally, that the DIS-MOL staffing always exceeds the DIS staffing. However, it remains to determine an explicit formula for

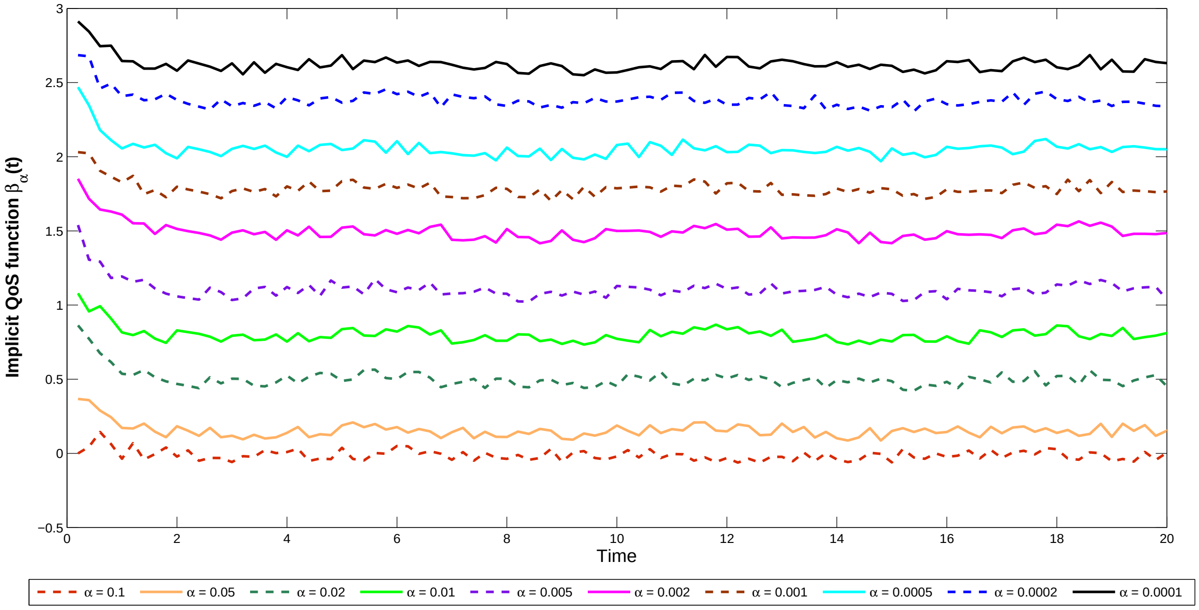


**Figure EC.4** The implicit QoS function $\beta_\alpha(t) \equiv (s_\alpha(t) - m_\alpha(t))/\sqrt{m_\alpha(t)}$ for different abandonment probability target $\alpha$, in the $M_t/M/s_t + M$ example, where $a = 100$, $b = 20$, $c = 1$, $\theta = 0.5$, $\mu = 1$.

the appropriate function $\beta_\alpha$ (even though it could be obtained numerically from plots like those in Figure EC.4).

We did statistical fitting to determine an approximation for the QoS function $\beta_\alpha$ to use in (21) for the $M_t/M/s_t + M$ model with a sinusoidal arrival rate function like (18) as a function the delay target $\alpha$ and the model parameters $(\lambda(t), \mu, \theta)$. First, we found that the dependence upon $\alpha$ tends to be linear in $-\log(\alpha)$. Second, we found that it sufficed to consider the impact of $\lambda(t)$ via its time average $\bar{\lambda}$. These fitting experiments lead us to propose the following linear approximation (in $-\log(\alpha)$) as an engineering solution:

$$\beta(\alpha, \bar{\lambda}, \theta, \mu) \approx K \cdot (-\log(\alpha)) + C(\bar{\lambda}/\mu, \theta/\mu), \tag{EC.1}$$
$$\textit{where} \quad C(\bar{\lambda}/\mu, \theta/\mu) \equiv -K \log(\sqrt{\bar{\lambda}/\mu}) + H(\theta/\mu),$$

where constant $K = 0.43$, $\bar{\lambda}$ is the average arrival rate, and $H(x)$ is the linear function $H(x) = 0.2822x - 0.0133$. Figure EC.5 shows both $\beta_\alpha$ obtained from the DIS-MOL approximation and its linear approximation (EC.1), with different $\bar{\lambda}/\mu$ and $\theta/\mu$. We fix $\mu = 1$. Since the arrival is

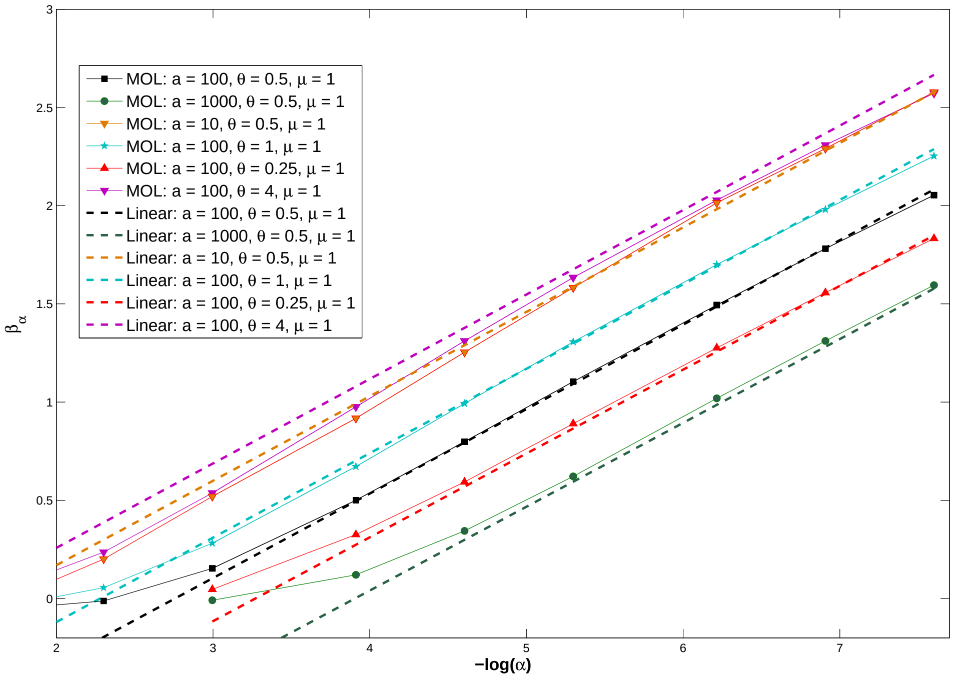


**Figure EC.5** Comparison of the linear approximation in (EC.1) for $\beta_\alpha$ to the value of $\beta_\alpha$ obtained from the DIS-MOL approximation.

sinusoidal, $a = \bar{\lambda}$ is the average arrival rate. This approximation is simple because (i) it is linear in $-\log(\alpha)$ when the other parameters are fixed, (ii) and the dependence on $\bar{\lambda}/\mu$ and $\theta/\mu$ is separable. Figure EC.5 shows that linear approximation (EC.1) is good when $a$ is not too small ($a > 10$) so that we are in the many-server regime, and when $\theta/\mu$ is not too big ($\theta/\mu < 4$), where the model begins to behave as a loss model.

## EC.6. Discretization Issues and Real-World Constraints

In this section we consider several discretization issues and real-world constraints for staffing functions. First, our basic method produces continuous staffing levels, but we have to staff at integer levels. To be conservative, we could always round up, but to obtain greater accuracy, we round to the nearest integer. That does not affect the performance much.

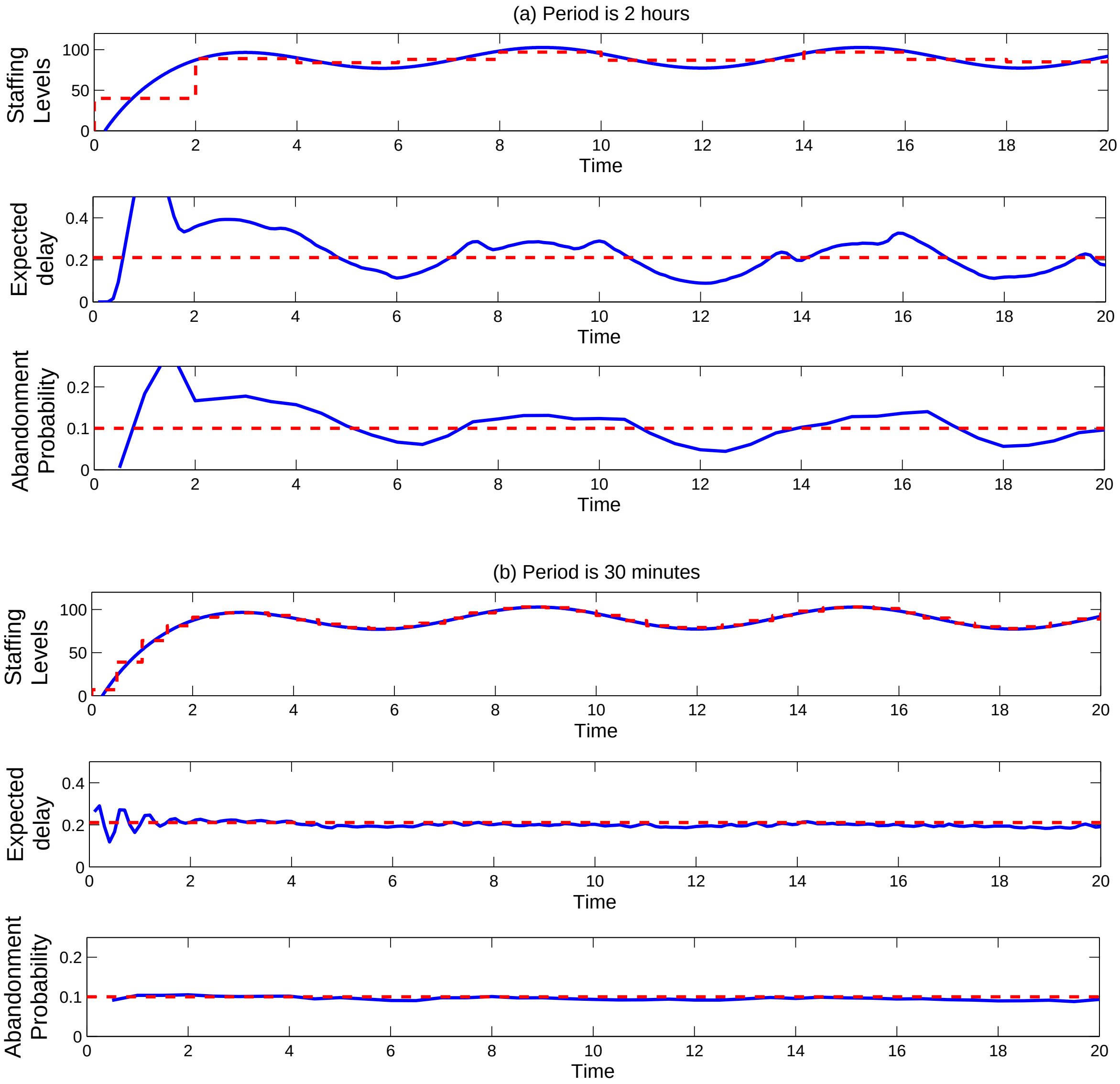
When the staffing schedule calls for decreasing the number of servers, we impose a real system constraint: We do not release any server until that server completes the service in process. However, we make an additional assumption that greatly helps achieve the required staffing level: We assume that server assignments can be switched when a server is scheduled to leave. In other words, if a specific server is scheduled to depart, then any other available server can complete the service. That means that we must only wait until the next service completion by *any* server when all servers are busy and the staffing level is scheduled to decrease by one.

This realistic feature makes the actual staffing level always lie above the integer-valued scheduled staffing level, as shown in Figure EC.24. As a consequence, the system will be slightly overstaffed when the staffing function $s(t)$ decreases. However, this effect tends to be insignificant when there are a large number of servers, provided we allow server switching as assumed above. (The number of servers ranges from 70 to 120 in the Markovian example considered in §6.) If the service-time distribution is exponential and all servers are busy when the number of servers is called to decrease, the lack-of-memory property of the exponential distribution implies that the time lag between the actual jump-down time of $s(t)$ and the scheduled one is $1/n\mu$, if there are currently $n$ servers in the system.

We can propose methods to counter this effect. In particular, during periods of planned staffing decrease, we can deliberately set the staffing level a bit lower to anticipate for that lag in releasing servers. One direct approach is to re-schedule a new staffing function which takes into consideration these time lags since they can easily be quantified. In §6 we don't have to modify the DIS and DIS-MOL staffing functions for the $M_t/M/s_t + M$ example with sinusoidal arrival since this effect is insignificant, as shown in Figure 4 and 5.

It is also interesting to consider another realistic constraint for the staffing function of the DIS-MOL approximation: Many real service systems have staffing periods with constant staffing levels. For example, it is common to only change staffing levels every half hour or every hour. We now consider this case. It is evident that the smaller the staffing period is, the less accuracy we should lose because of the requirement that the staffing function remain constant over each staffing interval.

In Figure EC.6 (and Figures EC.25 and EC.26 in the longer version), we plot the abandonment



(a) Period is 2 hours

(b) Period is 30 minutes

**Figure EC.6** Simulation estimates for the $M_t/M/s_t+M$ example with sinusoidal arrival under DIS-MOL staffing with fixed staffing period $d=30$ minutes and 2 hours, $\alpha=0.1$.

probabilities, expected delays with abandonment probability target $\alpha=0.1$, 0.05 and 0.02, respectively for two cases: fixed staffing periods of 2 hours and 30 minutes. In particular, we change the staffing once every 30 minutes (or 2 hours) by setting the constant number of servers in each period. equal to the midpoint of the proposed continuous staffing function. When the staffing period is long (2 hours), fluctuations brought by the crudeness of the discreteness of the staffing functions are unavoidable. However, these performance measures are still quite stable except during the initial transient periods, which is caused by the steep rise of the DIS-MOL staffing function.