# STAFFING TO STABILIZE BLOCKING IN LOSS MODELS WITH TIME-VARYING ARRIVAL RATES

Andrew Li, Ward Whitt and Jingtong Zhao

Operations Research Center, M.I.T.
77 Mass Ave, Bldg E40-130, Cambridge, MA 02139-4307; aali@mit.edu

Department of Industrial Engineering and Operations Research, Columbia University,
New York, NY, 10027; ww2040@columbia.edu
Department of Industrial Engineering and Operations Research, Columbia University,
New York, NY, 10027; jz2477@columbia.edu

October 6, 2015

**Abstract**

The modified-offered-load (MOL) approximation can be used to choose a staffing function (the time-varying number of servers) to stabilize delay probabilities at target levels in multi-server delay models with time-varying arrival rates, with or without customer abandonment. In contrast, as we confirm with simulations, it is not possible to stabilize blocking probabilities to the same extent in corresponding loss models, without extra waiting space, because these probabilities necessarily change dramatically after each staffing change. Nevertheless, blocking probabilities can be stabilized provided that we either randomize the times of staffing changes or average the blocking probabilities over a suitably small time interval. We develop systematic procedures and study how to choose the averaging parameters.

*Keywords:* nonstationary stochastic models; time-varying arrival rates; loss models; loss models with time-varying staffing; stabilizing performance with time-varying arrival rates;

*Short Title:* Staffing to Stabilize Blocking

*Contact Author:* Ward Whitt, ww2040@columbia.edu

# 1 Introduction

Since service systems typically have arrival rates that vary strongly over the day, there is considerable interest in developing effective time-varying staffing strategies (dynamically controlling the number of servers) to stabilize performance at target levels in face of time-varying demand. For shorter service times common in many telephone call centers, it is possible to set staffing levels to stabilize performance at target levels by using stationary models in a nonstationary way, e.g., via variants of the pointwise stationary approximation, as reviewed in [5]. However, for longer service times, the pointwise-stationary approximation is not effective. Nevertheless, as shown in [4, 9, 11, 15, 16, 17, 27], it is possible to apply modified-offered-load approaches to stabilize delay probabilities and abandonment probabilities at target levels in multi-server delay models with time-varying rates, with or without customer abandonment. Other methods have been proposed in [2, 4, 22, 23].

Thus it is natural to ask if the same conclusions can be made for multi-server loss models, without any extra waiting space. Of course, in many loss systems, such as communication systems, there is limited capability for dynamically changing the service capacity (e.g., number of circuits) in real time or even near real time. When staffing should be regarded as fixed, it is natural to consider controlling the demand instead, e.g., by dynamic pricing, as has been considered in [7, 8] and references therein. The dynamic pricing approaches can exploit the literature on the time-varying performance of the nonstationary loss model with fixed staffing [6, 10, 18, 21].

However, staffing often can be considered flexible in loss systems. For example, an ambulance base serving several hospitals may have a fixed number of ambulances, but the number of ambulances may be changed by scheduling transfers from one ambulance base to another at certain specified times. Moreover, these ambulances can only be sent out if the supporting medical personnel are available. Thus, there is a dynamic staffing question for the medical personnel in this setting, which translates to the ambulances. Similarly, a hotel has a fixed number of rooms, but the number of rooms that are available for customers at any given time is likely to be variable, because some are being renovated or cleaned, which requires various service personnel. Thus, there is a dynamic staffing question for the service personnel in this setting, which translates to the available hotel rooms.

Given that we do consider dynamic staffing, we need to carefully specify what happens when the service capacity is scheduled to decrease when all servers are busy, as discussed in the introduction

of [14]. Do we require that customers in service stay in service with the same server until their service is complete? Our analysis here applies to the case in which we allow the service in progress to be handed off to another available server. Even with such server-assignment switching, there are issues: Do we alter the prescribed staffing function to avoid forcing a customer out of service? Here in our simulations we release the first server that becomes free after the time of scheduled staffing decrease, and perform service switching at that instant.

## 2 A First Look

It may be surprising, but it is not possible to stabilize blocking probabilities in loss models with time-varying arrival rates as well as the delay probabilities have been stabilized in corresponding delay models. To quickly illustrate the difficulty, we show a simple example in Figure 1.

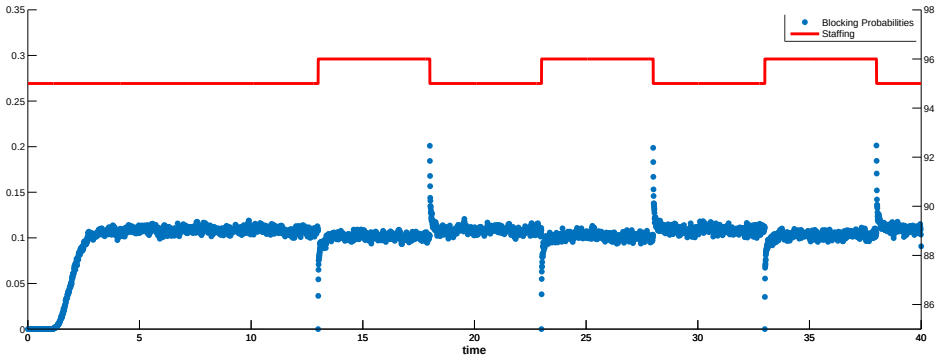


Figure 1: Simulation estimates of the time-varying blocking probabilities (left axis) with time-varying staffing (right axis) in an Erlang loss model with arrival rate $\lambda = 100$ and service rate $\mu = 1$, starting empty, in response to six staffing changes.

The model in Figure 1 is the stationary Erlang $(M/M/s/0)$ loss model with arrival rate $\lambda = 100$, individual mean service time $\mu^{-1} = 1$ and thus offered load $\alpha \equiv \lambda/\mu = 100$. With a blocking probability target of $B \equiv B(s, \alpha) = 0.1$, the target staffing level is $s = 96$. Figure 1 shows the consequence of switching the staffing level between 95 and 96 three times over the time interval $[0, 40]$. At the left in Figure 1, we see the usual startup effect associated with starting the system empty. It is evident that steady-state is achieved here in about three mean service times. However, there are dramatic changes in the blocking probabilities at the time of each staffing change, decreasing sharply at each staffing increase, and increasing sharply at each staffing decrease.

In retrospect, it is evident that the blocking probability function should behave this way. It is evident that the blocking probability necessarily drops to 0 immediately after each staffing increase.

(It can be proved that the right limit of the blocking probability function has to be 0.) The first arrival after that time of increase necessarily will not be blocked because there is new capacity that will be available. Nevertheless, from a practical point of view, we see that it makes sense to seek staffing functions that appropriately stabilize blocking probabilities in face of time-varying arrival rates when the flexibility exists.

# 3 Two Resolutions to the Problem

In this paper we propose two resolutions to the difficulty above. First, we show that the blocking probability can indeed be stabilized if we appropriately randomize the times of these staffing changes, and measure the blocking probability at each time by its average value (thus estimating the expected value with respect to the randomization, as opposed to conditional on any given realization).

Second, we show that the blocking probabilities associated with deterministic staffing rules can be stabilized if we look at appropriate average blocking probabilities, averaging in a short interval about each staffing change. The initial randomized timing might be preferred to prevent customers from discovering the staffing policy and trying to take advantage of it. For example, customers might deliberately schedule their arrivals immediately after a known staffing increase. If there were multiple customers competing for access, then a fixed change schedule could induce even more complicated behavior (which we do not study here).

In the first randomization approach we randomize according to a mean-0 Gaussian random variable, centering at the time of the scheduled deterministic staff change. With that simple Gaussian approach, we are left with only one parameter: the standard deviation $\sigma$. With the second averaging approach we average the blocking probabilities over a fixed interval of width $\Delta$, again centered at the time of the scheduled staff change. We conclude that both methods can be effective, provided that $\Delta$ and $\sigma$ are chosen properly.

## 3.1 Randomized Staffing Change Times

To describe our proposed randomization algorithm in detail, suppose that we have a non-stationary loss model over a time interval $[0, \tau]$ and an integer-valued staffing function $s(t)$ chosen to stabilize the blocking probability; we indicate how $s(t)$ can be chosen in Section 4. Naturally, $s(t)$ should be piecewise-constant and right-continuous, implying that there exists a sequence of staffing values $\{s_i, i \geq 1\}$ and a strictly increasing sequence of staffing change times $\{t_i : 0 \leq i \leq n, t_0 = 0, t_n = \tau\}$

such that

$$s(t) = s_i, \;\; t_{i-1} \le t < t_i, \;\; 1 \le i \le n. \tag{3.1}$$

We randomize by adding a small random time shift to each of the scheduled staffing change times. More specifically, let $\{\epsilon_i, i \ge 1\}$ be a sequence of i.i.d. Gaussian random variables with mean 0 and variance $\sigma^2$. In the first step, the sequence of scheduled staffing changes $\{t_i\}$ is replaced with a random sequence $\{\tilde{t}_i\}$, where

$$\tilde{t}_0 = 0 \quad \text{and} \quad \tilde{t}_i = t_i + \epsilon_i, \quad \text{for all} \quad i \ge 1. \tag{3.2}$$

We are not done because the sequence of randomized staffing change times $\{\tilde{t}_i\}$ constructed in (3.2) may fail to be nondecreasing or be contained in the time interval $[0, \tau]$. In the second step we remedy that deficiency. We do so by truncating $\tilde{t}_i$, i.e., by replacing $\tilde{t}_i$ by $(\tilde{t}_i \vee \tilde{t}_{i-1}) \wedge t_{i+1}$ for each $i$ successively, $1 \le i \le n-1$, where $a \vee b \equiv \max\{a, b\}$ and $a \wedge b \equiv \min\{a, b\}$. As a consequence, we have

$$\tilde{t}_{i-1} \le \tilde{t}_i \le t_{i+1} \quad \text{for all} \quad 1 \le i \le n-1, \tag{3.3}$$

so that the sequence $\{\tilde{t}_i\}$ is nondecreasing. (The parameter $\sigma$ should be chosen small enough so that that truncation rarely occurs.) We can then make the sequence $\{\tilde{t}_i\}$ strictly increasing by including only the last from each group of tied elements.

This procedure produces a random staffing function

$$\tilde{s}(t) = s_i, \;\; \tilde{t}_{i-1} \le t < \tilde{t}_i, \;\; 1 \le i \le n. \tag{3.4}$$

We estimate the blocking probability by performing many independent replications. Assuming that the randomization is successful, we can use previously developed performance approximations to analytically determine the performance, ignoring these complications.

### 3.2 Average Blocking Over Time Intervals

In the second averaging approach, we consider the blocking probability in intervals of fixed length, rather than the instantaneous blocking probability at a given moment. Given an interval length $\Delta \ge 0$, let $B^{\Delta}(t)$ be the time-average average blocking probability in the interval of length $\Delta$ centered at $t$, i.e. the probability that an arrival in the time interval $[t - \frac{\Delta}{2}, t + \frac{\Delta}{2}]$ is blocked. Let $B^0(t)$ be the instantaneous blocking probability, which we saw in the previous section can exhibit wild jumps when staffing changes occur. Fortunately, for appropriate interval lengths $\Delta$, we expect the averaging to smooth out these discontinuities.

As motivation for this approach, note that it coincides with the way blocking probabilities are measured from system or simulation data. Since the instantaneous blocking probability at any time generally cannot be measured directly, it is estimated by the proportion of the arrivals that occur in a small interval around that time that are blocked. This procedure is exactly the same as what we are proposing. In the model, $B^{\Delta}(t)$ is the time-average blocking probability over the interval $[t - \Delta/2, t + \Delta/2]$.

### 3.3 The Relevant Time Scale for $\sigma$ and $\Delta$

From the perspective of customer performance, usually the most relevant time scale is a single mean service time, which we have taken to be 1. However, there are two other time scales that are relevant for the choice of $\sigma$ and $\Delta$. The first is determined by the rate at which the system state tends to change and the second, closely related to the first, is the rate at which staffing levels change.

The time scale at which the state (number in system) changes typically is of order equal to a mean interarrival time. For many-server queues, that tends to be much shorter than the time scale of individual service times. With mean service time 1 and arrival rates of order 100, arrivals and departures both tend to occur at about rate 100. Thus the time scale for state changes is about 0.01.

Of course, the rate of staffing changes is related to the rate of state changes, but it also depends on the rate at which the time-varying arrival rate changes. Typically, the rate of staffing changes is significantly less than the rate of state changes. Fortunately, it often can be estimated directly. For periodic arrival rate functions, the number of staffing changes per cycle can be estimated by $\nu \equiv 2(s_U - s_L)$, where $s_U$ ($s_L$) is the maximum (minimum) staffing level in the cycle. The rate of staffing changes is then approximately $\nu/T$, where $T$ is the mean cycle length.

We estimate that the relevant values of $\sigma$ and $\Delta$ should be larger than one mean interarrival time but less than one mean service time, and about the same order as the average time between successive staffing changes.

### 3.4 Application of the Two Methods to the Example in Figure 1

We first apply the two methods to the example reported in Figure 1. We applied the two methods for a range of $\sigma$ and $\Delta$, measuring the blocking probability every 0.001 time units over each of 10,000 independent replications. Thus each estimated blocking probability has approximately mean 0.1,

variance $0.1(0.9)/10000 = 0.000009$ and standard deviation 0.003. Results from the randomization and averaging methods are given in Tables 1 and 2. The unusual extreme values in Figure 1 are seen in the mimimum (maximum) for the staffing increase (decrease) with $\sigma = \Delta = 0.00$; the improvements are seen in the other elements of those columns (highlighted in bold).

We conclude, first, that even a small amount of randomization or averaging helps a lot and, second, that both methods are effective and roughly comparable in their performance if we let $\Delta \approx 2.5\sigma$; e.g., we might select $\sigma = 0.08$ and $\Delta = 0.2$, for which the minimum and maximum pairs are $(0.088, 0.129)$ and $(0.086, 0.127)$, respectively. These values of $\sigma$ and $\Delta$ lie between the mean interarrival time 0.01 and the mean service time 1.0, without being near either one.

Table 1: The randomization method as a function of the standard deviation $\sigma$ for the model in Figure 1: the minimum, average and maximum estimated blocking probability from simulations over a unit interval centered at the time of the staffing change. The steady-state performance of the stationary model with each staffing level is given below for comparison.

| staff change from 95 to 96 at time 13 | | | | staff change from 96 to 95 at time 18 | | |
|---|---|---|---|---|---|---|
| std. dev. | min. | average | max. | min. | average | max. |
| 0.00 | **0.0087** | 0.1022 | 0.1154 | 0.0961 | 0.1106 | **0.2012** |
| 0.02 | **0.0698** | 0.1018 | 0.1162 | 0.0960 | 0.1120 | **0.1557** |
| 0.04 | **0.0782** | 0.1014 | 0.1175 | 0.0955 | 0.1103 | **0.1379** |
| 0.06 | **0.0846** | 0.1020 | 0.1184 | 0.0979 | 0.1119 | **0.1320** |
| 0.08 | **0.0879** | 0.1018 | 0.1152 | 0.0973 | 0.1114 | **0.1293** |
| 0.10 | **0.0881** | 0.1006 | 0.1141 | 0.0972 | 0.1113 | **0.1291** |
| 0.15 | **0.0937** | 0.1006 | 0.1135 | 0.0937 | 0.1109 | **0.1257** |
| 0.20 | **0.0959** | 0.1013 | 0.1160 | 0.0959 | 0.1092 | **0.1213** |
| steady state: 96 servers at time 10 | | | | steady state: 95 servers at time 16 | | |
| | min. | average | max. | min. | average | max. |
| | 0.0936 | 0.1005 | 0.1100 | 0.0979 | 0.1072 | 0.1186 |

Table 2: The averaging method as a function of the interval length $\Delta$ for the model in Figure 1: the minimum, average and maximum estimated blocking probability from simulations over a unit interval centered at the time of the staffing change. The steady-state performance of the stationary model with each staffing level is given below for comparison.

| staff change from 95 to 96 at time 13 | | | | staff change from 96 to 95 at time 18 | | |
|---|---|---|---|---|---|---|
| $\Delta$ | min. | average | max. | min. | average | max. |
| 0.00 | **0.0087** | 0.1005 | 0.1154 | 0.0961 | 0.1106 | **0.2012** |
| 0.04 | **0.0588** | 0.1005 | 0.1124 | 0.0958 | 0.1092 | **0.1641** |
| 0.10 | **0.0753** | 0.1005 | 0.1115 | 0.0983 | 0.1092 | **0.1382** |
| 0.20 | **0.0855** | 0.1005 | 0.1109 | 0.0997 | 0.1092 | **0.1271** |
| 0.40 | **0.0909** | 0.1005 | 0.1099 | 0.1006 | 0.1092 | **0.1199** |
| steady state: 96 servers at time 10 | | | | steady state: 95 servers at time 16 | | |
| | min. | average | max. | min. | average | max. |
| | 0.0936 | 0.1005 | 0.1100 | 0.0979 | 0.1072 | 0.1186 |

## 4 Stabilizing with a Time-Varying Arrival Rate

We now introduce a specific MOL approximation to specify the staffing function $s(t)$ in order to stabilize the blocking probability. We will be brief because it combines two well studied components: (i) an accurate heavy-traffic approximation for stationary loss systems [13, 25] and (ii) the modified-offered-load approximation for non-stationary systems [10, 18] and its application to stabilize performance [4, 5, 9, 11, 15, 16, 17, 27]. We refer to [13] and [5] for more background.

For the general stationary loss system $G/G/s/0$, an approximation for the blocking probability $B$ based on a heavy-traffic limit by Borovkov has been shown to be accurate for general arrival and service distributions, even with independence assumptions relaxed [13, 25]. These are based on the offered load and peakedness (the mean and the ratio of the variance to the mean, respectively, of the number of busy servers in an associated infinite-server model). Let $B(s, \alpha, z)$ be the approximate stationary blocking probability as a function of the offered load $\alpha \equiv \lambda/\mu$ and the peakedness $z$. As in [13], we would use the heavy-traffic limit of the peakedness, just as in [13], but we do not dwell on that here because we will focus on the special case of Poisson arrival processes, where $z = 1$.

We will use the approximation

$$B \approx B(s, \alpha, z) \equiv \sqrt{\frac{z}{\alpha}} \left( \frac{\phi\left((s-\alpha)/\sqrt{\alpha z}\right)}{\Phi\left((s-\alpha)/\sqrt{\alpha z}\right)} \right), \tag{4.1}$$

where $\Phi$ and $\phi$ are the cdf and pdf respectively of the standard (mean 0, variance 1) Gaussian distribution. (We remark that this Gaussian heavy-traffic approximation could be replaced by the Hayward approximation $B(s, \alpha, z) = B(s/z, \alpha/z, 1) \equiv B(s/z, \alpha/z)$, where $B(s, \alpha)$ is the exact Erlang blocking formula for the $M/M/s/0$ model [13], but that is somewhat more computationally demanding.)

To approximate the blocking probability in a non-stationary system with a given staffing function $s(t)$, we make the standard MOL adjustment to (4.1). Let $B(t)$ denote the blocking probability at time $t$. Then

$$B(t) \approx B(s(t), m(t), z), \tag{4.2}$$

where $m(t)$ is the time-varying mean number of busy servers in the infinite-server model with the same arrival and service processes. This procedure for non-Markovian models follows §5 of [11] and [9].

Now if we have a target blocking probability $B^*$, we simply set the staffing function $s(t)$ by inverting the previous approximation and rounding to the nearest integer. More specifically, at a

given time $t$, let $s$ be the (possibly non-integral) value that satisfies $B(s, m(t), z) = B^*$. Then we set $s(t) = \text{int}(s)$, where $\text{int}(s)$ is the integer closest to $s$.

# 5 Simulation Experiments

## 5.1 The Model

To evaluate our stabilization approaches, we conducted a series of simulation experiments for the time-varying $M_t/GI/s_t/0$ system with a nonhomogeneous Poisson process as an arrival process (the $M_t$) with the sinusoidal arrival rate function

$$\lambda(t) = \bar{\lambda} + \beta \sin(\gamma t), \quad t \geq 0, \tag{5.1}$$

with average arrival rate $\bar{\lambda}$, amplitude $\beta$, and cycle length (or period) $T$ (or equivalently, frequency $\gamma = 2\pi/T$). We let the mean service time be 1 time unit. We started by considering the Markovian $M_t/M/s_t/0$ special case and then considered variations of that model. We specify the model by the four-tuple $(\bar{\lambda}, \beta, T, B)$, where $B$ is the blocking probability target.

For the $M_t/M/s_t/0$ model, the heavy-traffic peakedness $z$ equals 1, and formula (15) in [3] gives the time-varying offered load:

$$m(t) = \bar{\lambda} + \frac{\beta}{1+\gamma^2}(\sin \gamma t - \gamma \cos \gamma t). \tag{5.2}$$

and formula (18) in [3] gives the difference between the maximum and minimum offered load:

$$m_U - m_L = \frac{2\beta}{\sqrt{1+\gamma^2}}. \tag{5.3}$$

We considered two values of the scale, $\bar{\lambda} = 100$ (large) and 20 (moderate), and two cycle lengths, $T = 100$ (long) and $T = 10$ (short), so that $\gamma = 0.0628$ and $\gamma = 0.628$. We let the amplitudes be $\beta = 25$ for $\bar{\lambda} = 100$ and $\beta = 5$ for $\bar{\lambda} = 20$. We consider two blocking probability targets: $B = 0.1$ (heavier loading) and $B = 0.01$ (lighter loading).

## 5.2 The Staffing Functions

We start by showing the arrival rates and the MOL staffing functions based on (5.2) and (4.2) for blocking probability targets $B = 0.1$ and $B = 0.01$ in the four cases. First, Figure 2 shows the cases with large scale $\bar{\lambda} = 100$ over a single periodic cycle. The staffing shown is appropriate for dynamic periodic steady state, as if the system started empty in the distant past. Consistent with previous research, the staffing plots in Figure 2 show that the pointwise stationary approximation

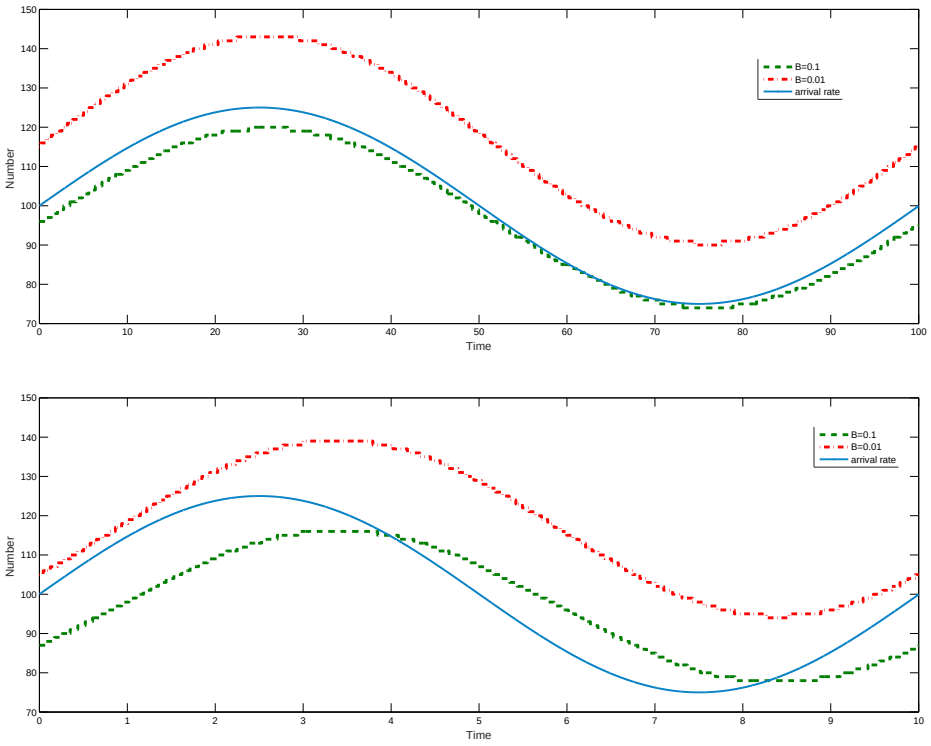


Figure 2: The MOL staffing functions and the sinusoidal arrival rate function in (5.1) for the $M_t/M/s/0$ model with $\mu = 1$, average arrival rate $\bar{\lambda} = 100$, amplitude $\beta = 25$ and two blocking probability targets $B = 0.1$ and $B = 0.01$: for long cycles $T = 100$ (top) and short cycles $T = 10$ (bottom).

(PSA) is effective for $T = 100$, but not for $T = 10$. A significant time lag and space shift is seen for $T = 10$. For $T = 100$, the staffing at time $t$ depends primarily on the arrival rate at time $t$. Figure 3 shows the corresponding cases with scale $\bar{\lambda} = 20$.

In order to judge what would be appropriate averaging parameters $\sigma$ and $\Delta$ for the $M_t/M/s_t/0$ model with the sinusoidal arrival rate function in (5.1) as a function of the model parameters $(\bar{\lambda}, \beta, T, B)$, using the MOL staffing before any averaging, we calculate the minimum and average distance between successive staffing changes in Table 3. We also show the minimum distance between successive pairs of staffing changes. We see that the minimum interval between successive pairs of staffing changes is nearly two times the minimum for a single change. We also see that the average is in between those two.

In Table 3 we also show our final choice of the averaging parameters $\sigma$ and $\Delta$. As should be expected, it turns out that there is greater freedom of choice when the arrival rate changes slowly, as in the case $T = 100$, than there is when the arrival rate changes rapidly, as for $T = 10$. Indeed,

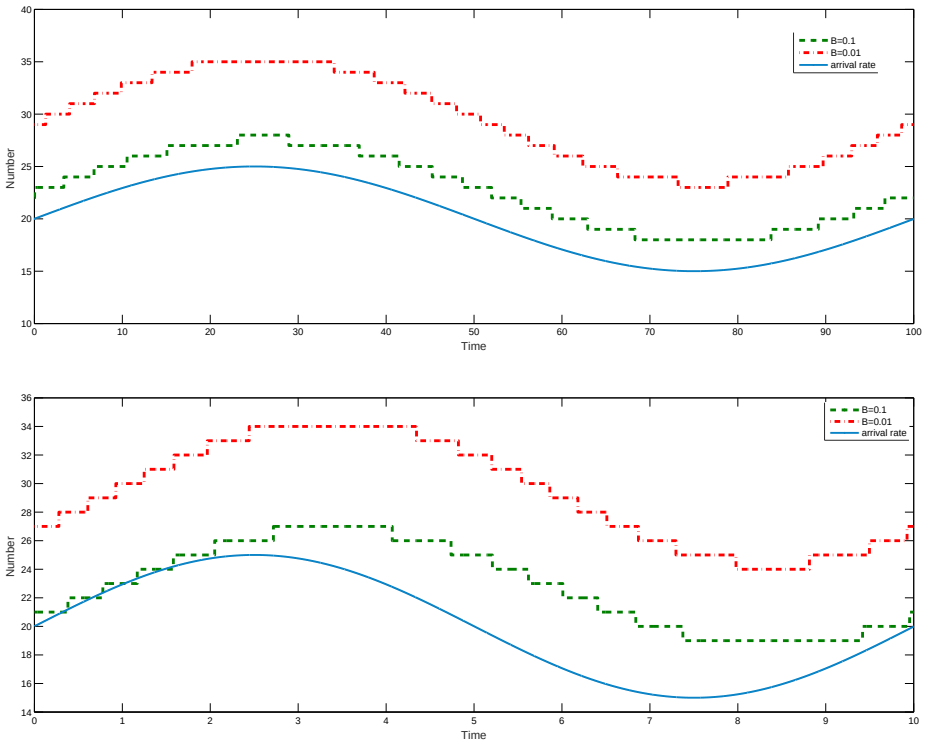


Figure 3: The MOL staffing functions and the sinusoidal arrival rate function in (5.1) for the $M_t/M/s/0$ model with $\mu = 1$, average arrival rate $\bar{\lambda} = 20$, amplitude $\beta = 5$ and two blocking probability targets $B = 0.1$ and $B = 0.01$: for long cycles $T = 100$ (top) and short cycles $T = 10$ (bottom).

we find that there is a relatively narrow range for the parameters for $T = 10$. Based on our results to be shown, we give one good parameter for $T = 10$ and a range for $T = 100$. In both cases, it is appropriate to use the minimum distance between successive staffing changes as a guide for what are appropriate parameters. (Recall that a mean interarrival time is 0.01 for $\bar{\lambda} = 100$ and 0.05 for $\bar{\lambda} = 20$.)

We remark that we can estimate the time betweeen successive staffing changes from (5.3) without performing the detailed calculations in Table 3. For $\bar{\lambda} = 100$ and $\beta = 25$, the distance in (5.3) equals 42.3 for $T = 10$ and 49.9 for $T = 100$. Thus, a rough estimate of the number of staffing changes per cycle is 84.6 for $T = 10$ and 99.8 for $T = 100$. The mean time between staffing changes should thus be about 0.12 for $T = 10$ and 1.0 for $T = 100$. That is consistent with the cases with $\bar{\lambda} = 100$ in Table 3. Since the mean interarrival time is 0.01 for $\bar{\lambda} = 100$, we infer that the "averaging halfwidths" $\Delta/2$ and $2\sigma$ should be of order 0.1 in the difficult case with $\bar{\lambda} = 100$ and $T = 10$, because 0.10 is just below 0.12, the mean time between successive staffing changes. Indeed,

Table 3: The distance between successive staffing changes for the $M_t/M/s_t/0$ model with the sinusoidal arrival rate function in (5.1) as a function of the model parameters $(\bar{\lambda}, \beta, T, B)$ using the MOL staffing before any averaging. The minimum and average distance is shown along with the minimum of two successive changes.

| distance between successive staffing changes | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| model parameters | | | | distance measure | | | ave param. | |
| $\bar{\lambda}$ | $\beta$ | $T$ | target $B$ | minimum | average | min of two | $\sigma$ | $\Delta$ |
| 100 | 25 | 10 | 0.001 | 0.067 | 0.106 | 0.135 | 0.08 | 0.20 |
| 100 | 25 | 10 | 0.01 | 0.071 | 0.109 | 0.142 | 0.08 | 0.20 |
| 100 | 25 | 10 | 0.1 | 0.082 | 0.133 | 0.164 | 0.08 | 0.20 |
| 20 | 5 | 10 | 0.001 | 0.290 | 0.517 | 0.583 | 0.32 | 0.80 |
| 20 | 5 | 10 | 0.01 | 0.319 | 0.561 | 0.641 | 0.32 | 0.80 |
| 20 | 5 | 10 | 0.1 | 0.391 | 0.723 | 0.789 | 0.32 | 0.80 |
| 100 | 25 | 100 | 0.001 | 0.571 | 0.926 | 1.142 | $0.08 - 0.96$ | $0.20 - 2.40$ |
| 100 | 25 | 100 | 0.01 | 0.604 | 0.979 | 1.208 | $0.08 - 0.96$ | $0.20 - 2.40$ |
| 100 | 25 | 100 | 0.1 | 0.700 | 1.142 | 1.400 | $0.08 - 0.96$ | $0.20 - 2.40$ |
| 20 | 5 | 100 | 0.001 | 2.461 | 4.016 | 4.940 | $0.32 - 3.84$ | $0.80 - 9.60$ |
| 20 | 5 | 100 | 0.01 | 2.702 | 4.367 | 5.425 | $0.32 - 3.84$ | $0.80 - 9.60$ |
| 20 | 5 | 100 | 0.1 | 3.302 | 5.376 | 6.670 | $0.32 - 3.84$ | $0.80 - 9.60$ |

our experiments show that too close to the lower value 0.01 is not good, but increasing much above the higher value 0.10 is not good either.

### 5.3 Simulation Methodology

In all models we consider the arrival process is a nonhomogeneous Poisson process $(M_t)$ with intensity function $\lambda(t)$. We generated arrivals over a time interval $[0, T]$ in the usual way by thinning. We first choose a constant $\lambda_U$ such that $0 \le \lambda(t) \le \lambda_U$ for all $t \le T$. We then generated potential arrivals using a Poisson process with rate $\lambda_U$ by generating the successive interarrival times as i.i.d. exponential random variables with mean $1/\lambda_U$. Finally, each potential arrival is treated as an actual arrival with probability $\lambda(t)/\lambda_U$.

We specify the staffing levels and the times of successive staffing changes without any averaging by using the MOL approach, as indicated in §4 and §5.2. To estimate the blocking probabilities using the $\sigma$ parameters, in each replication we randomize the staffing times, getting new staffing times, as indicated in §3.1. Thus, with both methods we start with the staffing levels and staffing change times for each replication.

For both methods, we keep the time of every arrival and departure in a matrix, and the number of people in the system at that time in another matrix. We then updated the two matrices depending on whether an arrival would happen first or a departure would happen first. At the end of each replication, we used these matrices to decide whether the system was full at each sampling time.

We constructed a matrix and incremented the value by 1 if the system was full at the sampling time (which was 0.001). The final blocking probability at each sampling time was then calculated by dividing the value by the number of replications. To estimate the blocking probabilities using the $\Delta$ parameters, in the final step we recorded the total number of arrivals that were blocked in an interval of length $\Delta$ around each sampling time. The final blocking probability at each sampling time was then calculated by dividing the value by the total number of arrivals in the interval.

For estimating the blocking probabilities with the parameter $\sigma$, we are estimating the probability that the system is full at the designated time, which is often referred to as the *time congestion*; see §1.2 and §6 of [13]. That alternative approach is justified because we have a Poisson arrival process; see [20] and Proposition A.1 on p. 567 of [19]. For non-Poisson arrival processes, the time congestion and call congestion can be very different; see [13]. We performed separate simulation experiments to verify that the time congestion and the call congestion indeed coincide for the models in this paper.

For estimating the blocking probabilities, we used a discrete time grid of 0.001 (the sampling time), but in order to make smaller-sized plots, we reduced the time grid to 0.01 for $T = 10$ and 0.1 for $T = 100$. We discuss the statistical precision of our estimators in §5.4.1.

## 5.4 Long Cycles ($T = 100$) and Heavy Loading ($B = 0.1$)

### 5.4.1 The Corresponding Stationary Models

We start by considering the long cycles with $T = 100$ and the higher blocking probability target $B = 0.1$. All our simulations are based on multiple i.i.d. replications of the model over the fixed time interval $[0, 100]$, starting empty. We start by showing the statistical precision of our estimators and the impact of a single server in the performance of the associated $(\bar{\lambda}, \beta, T, B) = (100, 0, 100, 0.1)$ and $(20, 0, 100, 0.1)$ stationary models. Figure 4 shows simulation estimates of the blocking probabilities in the stationary $M/M/s/0$ model for arrival rate $\lambda = 100$ (left) and $\lambda = 20$ (right) with four different staffing levels yielding blocking above and below 0.10. The plot for $\lambda = 100$ is cut in the middle so that the width of the estimates can be seen. Figure 4 shows how much the blocking probabilities should change when we change the staffing level by a single server. We see that a single server changes the blocking probability by about 10% (20%) of the target $B = 0.1$ when the scale is $\lambda = 100$ ($\lambda = 20$).

Figure 4 also shows the statistical precision of all our experiments. It can be seen from the thickness of the plot for each fixed staffing level. Given that our estimates are based on $n$ i.i.d.

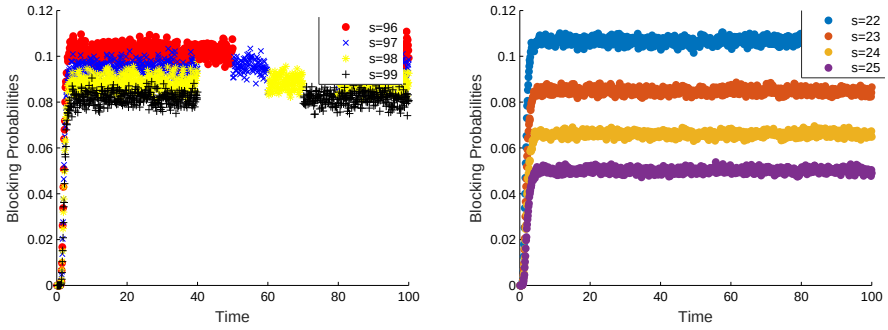


Figure 4: Simulation estimates of the blocking probabilities in the stationary $M/M/s/0$ model for arrival rate $\lambda = 100$ (left) and $\lambda = 20$ (right) with four different staffing levels yielding blocking above and below 0.10, showing the impact of a single server.

replications, we can apply the binomial distribution to estimate the variance of each estimate of the blocking probability when it is $p$. The variance is $p(1-p)/n$. For $p = 0.1$, the variance is approximately $0.1/n$, so that 5 standard deviations is about $1.5/\sqrt{n}$. For $n = 10,000$ used for $\lambda = 100$, that is 0.015. For $n = 50,000$ used for $\lambda = 20$, that is about 0.007 (explaining the thinner plots).

### 5.4.2 Without Any Averaging

We now consider the staffing algorithm without any form of averaging, i.e., the analog of Figure 1 for the $M_t/M/s_t/0$ model. Figure 5 shows the corresponding performance of the $M_t/M/s_t/0$ model with parameter triples $(100, 25, 100)$ (left) and $(20, 5, 100)$ (right), using the staffing algorithm without any form of averaging. Figure 5 clearly shows that the algorithm is not effective without either form of averaging. The blocking probability varies from about 0 to 0.18 (which is about right on average). We see excursion up (down) from the target when the staffing is decreasing (increasing).

### 5.4.3 The Two Forms of Averaging: Large Scale

We now consider the performance with the two forms of averaging for the first $(\bar{\lambda}, \beta, T, B) = (100, 25, 100, 0.1)$ case. Figure 6 shows the performance using randomization with $\sigma = 0.08$ (left) and averaging with $\Delta = 0.2$ (right). This choice is based on our experience in Tables 1 and 2. Note that $\Delta$ here is about one half the minimum distance between successive staffing changes. As before, we see excursions up (down) from the target when the staffing is decreasing (increasing), but these are very minor.

The performance gets smoother and closer to the target as we increase the interval length. We

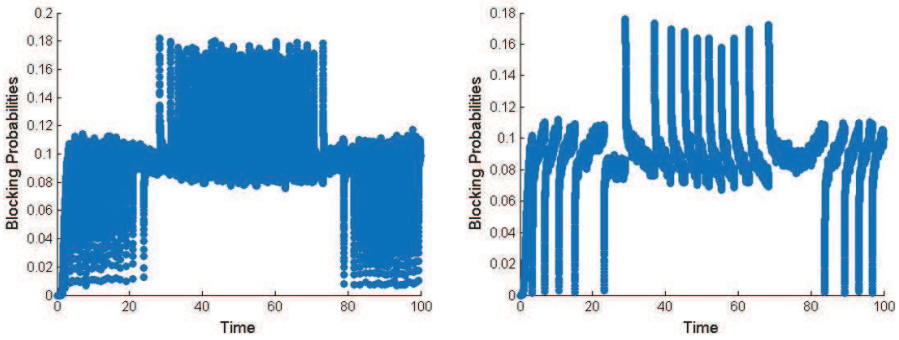


Figure 5: Simulation estimates of the blocking probabilities in the nonstationary $M_t/M/s_t/0$ model with the staffing algorithm without any averaging, for the sinusoidal arrival rate in (5.1) and parameter triple $(100, 25, 100, 0.1)$ (left) and $(20, 5, 100, 0.1)$ (right).

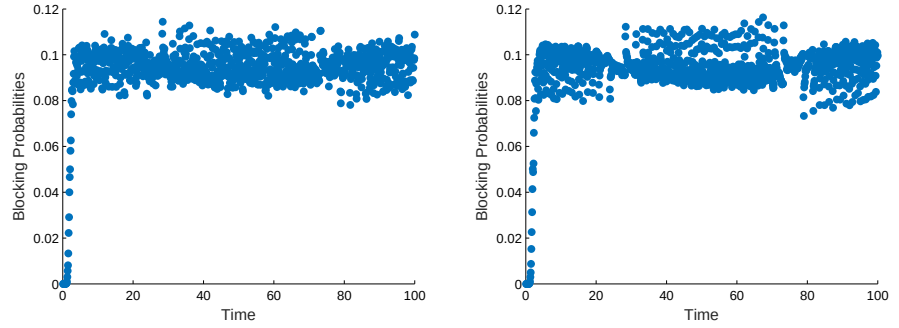


Figure 6: Simulation estimates of the blocking probabilities in the nonstationary $M_t/M/s_t/0$ model with parameter triple $(100, 25, 100)$ having average arrival rate $\bar{\lambda} = 100$ with the staffing algorithm for target $B = 0.1$ using randomization with $\sigma = 0.08$ (left) and averaging with $\Delta = 0.2$ (right).

illustrate the impact of the parameters $\sigma$ and $\Delta$ by showing the performance for two values of each in Figure 7. Table 4 gives a different view of the performance of the two averaging approaches

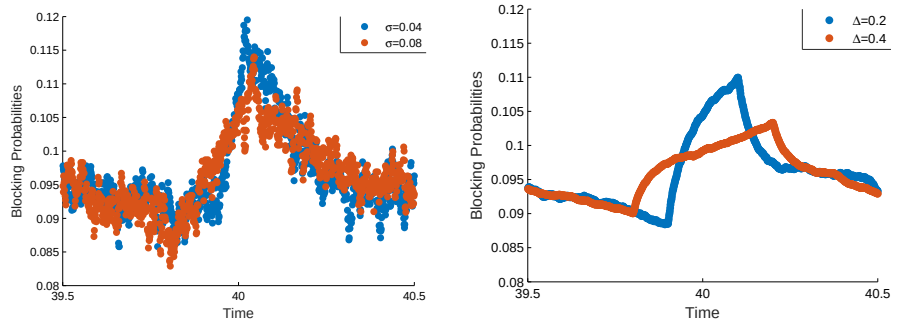


Figure 7: A careful examination of the time interval $[39.5, 40.5]$ in the setting of Figure 6: randomization for $\sigma = 0.04$ and $0.08$ (left) and averaging for $\Delta = 0.2$ and $0.4$ (right)

for the base model with parameter four-tuple $(\bar{\lambda}, \beta, T, B) = (100, 25, 100, 0.1)$ and randomization

parameter $\sigma = 0.08$ and averaging parameter $\Delta = 0.2$. Here we carefully look at the minimum, average and maximum blocking probabilities over four separate intervals of length 1.

Table 4: Simulation estimates of the blocking probabilities over four unit intervals each containing one staffing change, for the $M_t/M/s_t/0$ model with $\mu = 1$, $\lambda(t)$ in (5.1) with parameter four-tuple $(\bar{\lambda}, \beta, T, B) = (100, 25, 100, 0.1)$ $(\gamma = 0.0628)$ using the MOL staffing and randomization (left) and averaging (right). The minimum, average and maximum values over a unit interval are shown.

| estimated blocking probabilities over intervals of length 1 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| staffing change | | | randomization: $\sigma = 0.08$ | | | averaging: $\Delta = 0.2$ | | |
| time | from | to | min. | average | max. | min. | average | max. |
| 40.0 | 112 | 111 | 0.082 | 0.095 | 0.110 | 0.089 | 0.096 | 0.112 |
| 60.2 | 85 | 84 | 0.082 | 0.097 | 0.114 | 0.087 | 0.096 | 0.112 |
| 90.2 | 82 | 83 | 0.081 | 0.094 | 0.107 | 0.082 | 0.096 | 0.105 |
| 100.3 | 95 | 96 | 0.079 | 0.096 | 0.106 | 0.085 | 0.097 | 0.105 |

Next, Figure 8 shows that the parameters $\sigma$ and $\delta$ can be made larger without penalty; they are increased from Figure 6 by a factor of 12 to $(0.96, 2.4)$. Indeed, given that stabilization is achieved for $\Delta = 0.20$, as shown in Figure 6, higher values of $\Delta$ can only smooth out the estimate, giving less fluctuation. But we do not achieve that benefit for randomizing the staffing times over wider intervals. Nevertheless, since the cycles are so long, the larger value of $\sigma$ does not hurt. Indeed, the performance in Figure 8 is even better. We might nevertheless want more localized control of the blocking.

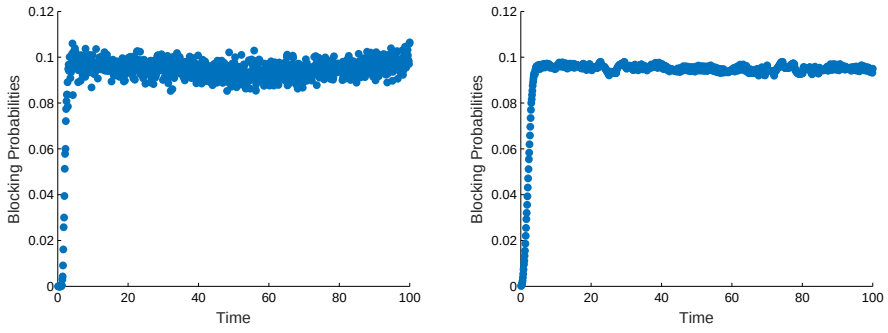


Figure 8: Simulation estimates of the blocking probabilities in the nonstationary $M_t/M/s_t/0$ model with parameter triple $(100, 25, 100)$ having average arrival rate $\bar{\lambda} = 100$ with the staffing algorithm for target $B = 0.1$ using randomization with $\sigma = 0.96$ (left) and averaging with $\Delta = 2.4$ (right).

### 5.4.4 The Two Forms of Averaging: Smaller Scale

We now shift to the smaller scale with $\bar{\lambda} = 20$ (and $\beta$ reduced proportionally). We find that the randomization and averaging interval lengths need to increase as the scale decreases. Paralleling Figure 6, Figure 9 shows the performance for the smaller scale model with parameter four-tuple $(\bar{\lambda}, \beta, T, B) = (20, 5, 100, 0.1)$ and randomization parameter $\sigma = 0.32$ and averaging parameter $\Delta = 0.8$. These parameters are 4 times larger than for parameter four-tuple $(100, 25, 100, 0.1)$ in Table 4. The average performance in Figure 9 falls somewhat below the target $B = 0.1$, but Figure

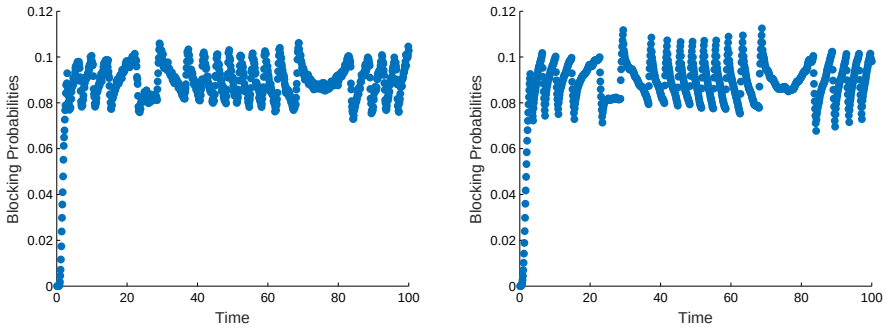


Figure 9: Simulation estimates of the blocking probabilities in the nonstationary $M_t/M/s_t/0$ model with parameter four-tuple $(20, 5, 100, 0.1)$ having average arrival rate $\bar{\lambda} = 20$ using randomization with $\sigma = 0.32$ (left) and averaging with $\Delta = 0.8$ (right)

4 shows that is consistent with the greater importance of a single server with the smaller scale, based on $\bar{\lambda} = 20$.

Paralleling Table 4, Table 5 shows the performance of the two averaging approaches for the smaller scale model.

Table 5: Simulation estimates of the blocking probabilities over four unit intervals each containing one staffing change, for the $M_t/M/s_t/0$ model with $\mu = 1$, $\lambda(t)$ in (5.1) with parameter four-tuple $(\bar{\lambda}, \beta, T, B) = (20, 5, 100, 0.1)$ $(\gamma = 0.0628)$ using the MOL staffing with randomization (left) and averaging (right). The minimum, average and maximum values over a unit interval are shown.

| estimated blocking probabilities over intervals of length 5 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| staffing change | | | randomization: $\sigma = 0.32$ | | | averaging: $\Delta = 0.8$ | | |
| time | from | to | min | average | max | min | average | max |
| 41.485 | 26 | 25 | 0.079 | 0.091 | 0.105 | 0.080 | 0.091 | 0.108 |
| 58.892 | 21 | 20 | 0.075 | 0.087 | 0.104 | 0.077 | 0.087 | 0.109 |
| 89.149 | 19 | 20 | 0.075 | 0.090 | 0.102 | 0.070 | 0.090 | 0.100 |
| 100.079 | 22 | 23 | 0.076 | 0.090 | 0.100 | 0.074 | 0.090 | 0.101 |

Next, Figure 10 shows that the parameters $\sigma$ and $\delta$ can be made larger without penalty; they are increased from Figure 9 by a factor of 12 to $(3.84, 9.6)$. Indeed, given that stabilization is achieved for $\Delta = 0.8$, as shown in Figure 9, higher values of $\Delta$ can only smooth out the estimate,

giving less fluctuation. But we do not achieve that benefit for randomizing the staffing times over wider intervals. Nevertheless, since the cycles are so long, the larger value of $\sigma$ does not hurt.

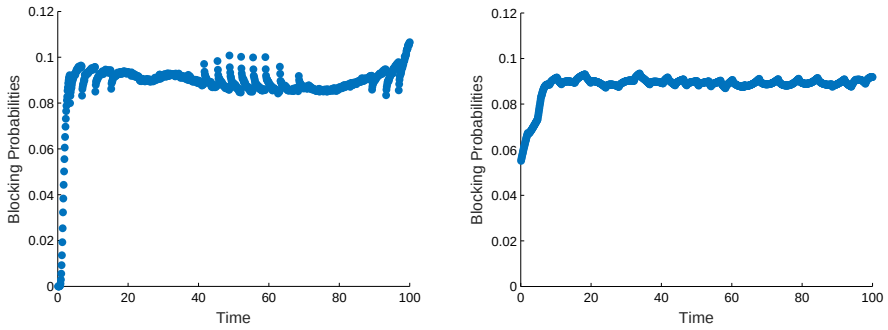


Figure 10: Simulation estimates of the blocking probabilities in the nonstationary $M_t/M/s_t/0$ model with parameter triple $(20, 5, 100)$ having average arrival rate $\bar{\lambda} = 20$ with the staffing algorithm for target $B = 0.1$ using randomization with $\sigma = 3.84$ (left) and averaging with $\Delta = 9.6$ (right).

### 5.5 Long Cycles ($T = 100$) and lighter Loading ($B = 0.01$)

In this section we repeat the experiments just done in §5.4 for the lower blocking probability target $B = 0.01$ instead of $B = 0.1$, i.e., for the models $(100, 25, 100, 0.01)$ and $(20, 5, 100, 0.01)$. To produce this lower blocking probability, the staffing has to be significantly higher, but we find that the averaging parameters $\sigma$ and $\Delta$ can be the same as before.

#### 5.5.1 The Corresponding Stationary Models with Target $B = 0.01$

Corresponding to Figure 4 for target $B = 0.1$, Figure 11 shows four staffing levels in the stationary case for the lower blocking probability target $B = 0.01$. From these figures we see that one server matters relatively more (compared to the blocking probabilities, which are ten times smaller) at these higher staffing levels.

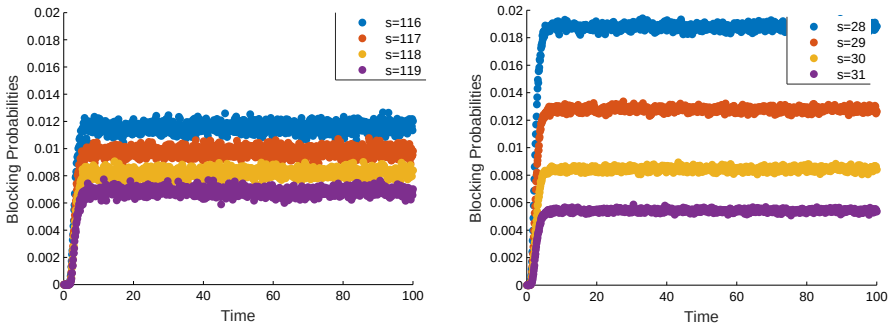


Figure 11: Simulation estimates of the blocking probabilities in the stationary $M/M/s/0$ model for arrival rate $\lambda = 100$ (left) and $\lambda = 20$ (right) with four different staffing levels yielding blocking above and below 0.01, showing the impact of a single server.

#### 5.5.2 Without Any Averaging for $B = 0.01$

We now consider the time-varying arrival rate again, but with lower blocking probability target $B = 0.01$. Figure 12 shows the performance of the direct staffing algorithm without any form of averaging for the lower blocking probability target $B = 0.01$. The figure looks just like Figure 5 except the realized blocking probabilities are now ten times smaller.

#### 5.5.3 The Two Forms of Averaging with Large Scale and Target $B = 0.01$

We now consider the performance with the averaging. Paralleling Figure 6, Figure 13 shows the performance of the base model with the staffing algorithm for target $B = 0.01$ using randomization with $\sigma = 0.08$ (left) and averaging with $\Delta = 0.2$ (right), the same as in Figure 6.

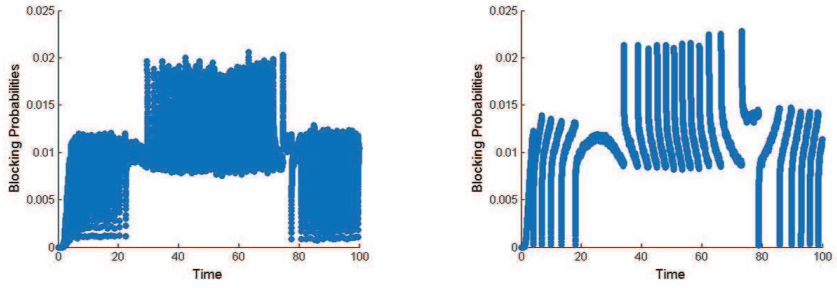


Figure 12: Simulation estimates of the blocking probabilities in the nonstationary $M_t/M/s_t/0$ model with the staffing algorithm without any averaging, for the sinusoidal arrival rate in (5.1) and parameter triple $(100, 25, 100, 0.01)$ (left) and $(20, 5, 100, 0.01)$ (right).

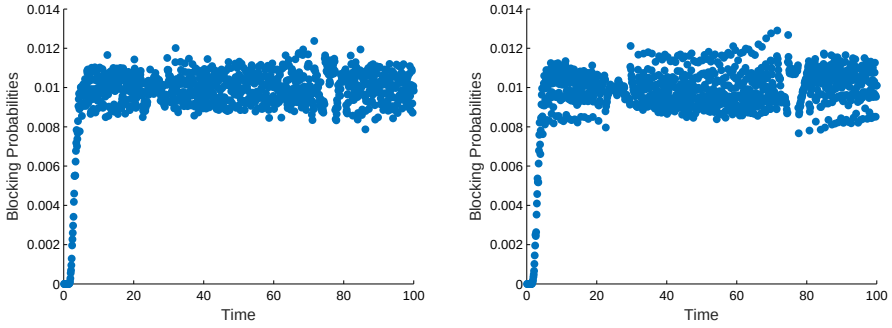


Figure 13: Simulation estimates of the blocking probabilities in the nonstationary $M_t/M/s_t/0$ model with parameter four-tuple $(\bar{\lambda}, \beta, T, B) = (100, 25, 100, 0.01)$ using randomization with $\sigma = 0.08$ (left) and averaging with $\Delta = 0.2$ (right).

We examine the two subintervals $[39.5, 40.5]$ and $[99.5, 100.5]$ in Figure 13 more carefully in Figures 14 and 15. Paralleling Table 4, we also report the minimum, average and maximum values over subintervals where there is at least one staffing change in Table 6.

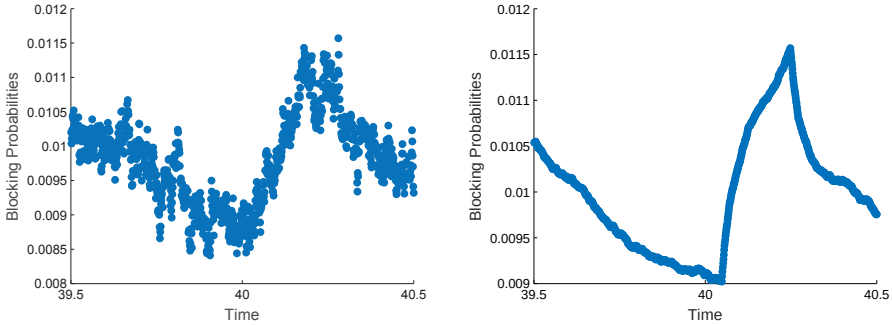


Figure 14: A careful examination of the time interval $[39.5, 40.5]$ for the case $(100, 25, 100, 0.01)$ in the setting of Figure 13: randomization for $\sigma = 0.08$ (left) and averaging for $\Delta = 0.2$ (right)

Next, Paralleling Figure 8, Figure 16 shows that the parameters $\sigma$ and $\delta$ also can be made larger

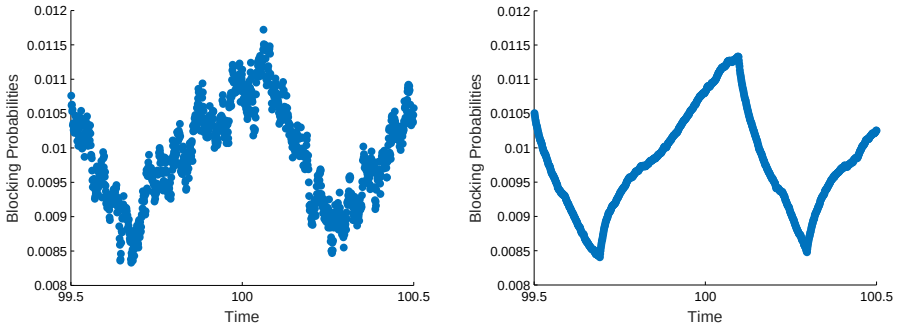


Figure 15: A careful examination of the time interval $[99.5, 100.5]$ for the case $(100, 25, 100, 0.01)$ in the setting of Figure 13: randomization for $\sigma = 0.08$ (left) and averaging for $\Delta = 0.2$ (right)

Table 6: Simulation estimates of the blocking probabilities over four intervals of length 1 each containing at least one staffing change, for the $M_t/M/s_t/0$ model with $\mu = 1$, $\lambda(t)$ in (5.1) with parameter four-tuple $(\bar{\lambda}, \beta, T, B) = (100, 25, 100, 0.01)$ using the MOL staffing with randomization (left) and averaging (right). The minimum, average and maximum values over a unit interval are shown.

| estimated blocking probabilities over intervals of length 1 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| staffing change | | | randomization: $\sigma = 0.08$ | | | averaging: $\Delta = 0.2$ | | |
| time | from | to | min | average | max | min | average | max |
| 40.148 | 134 | 133 | 0.0084 | 0.0098 | 0.0116 | 0.0090 | 0.0100 | 0.0116 |
| 60.201 | 103 | 102 | 0.0084 | 0.0101 | 0.0119 | 0.0089 | 0.0103 | 0.0121 |
| 89.617 | 99 | 100 | 0.0082 | 0.0099 | 0.0113 | 0.0083 | 0.0100 | 0.0113 |
| 90.396 | 100 | 101 | same | same | same | same | same | same |
| 99.592 | 114 | 115 | 0.0083 | 0.0099 | 0.0117 | 0.0084 | 0.0098 | 0.0113 |
| 100.197 | 115 | 116 | same | same | same | same | same | same |

without penalty for blocking probability target $B = 0.01$; just as before, they are increased from $(0.08, 0.20)$ in Figure 13 by a factor of 12 to $(0.96, 2.40)$. Indeed, given that stabilization is achieved for $\Delta = 0.20$, as shown in Figure 13, higher values of $\Delta$ can only smooth out the estimate, giving less fluctuation. The larger (relatively) fluctuation at time 80 is consistent with Figure 13. But we do not achieve that benefit for randomizing the staffing times over wider intervals. Nevertheless, since the cycles are so long, the larger value of $\sigma$ does not hurt.

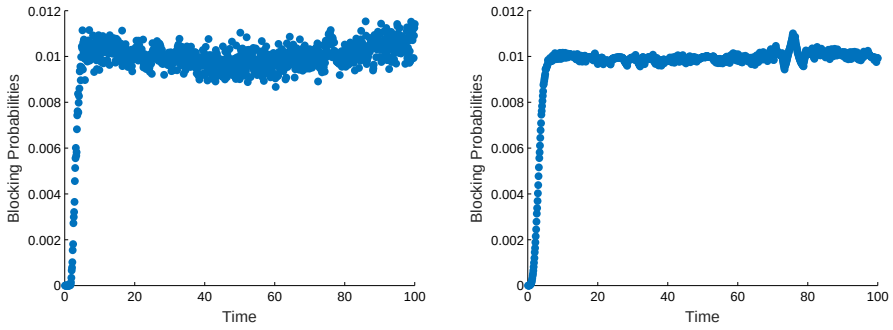


Figure 16: Simulation estimates of the blocking probabilities in the nonstationary $M_t/M/s_t/0$ model with parameter triple $(100, 25, 100)$ having average arrival rate $\bar{\lambda} = 100$ with the staffing algorithm for target $B = 0.01$ using randomization with $\sigma = 0.96$ (left) and averaging with $\Delta = 2.4$ (right).

### 5.5.4 The Two Forms of Averaging with Smaller Scale and Target $B = 0.01$

We now shift to the smaller scale with $\bar{\lambda} = 20$ (and $\beta$ reduced proportionally). We find that the randomization and averaging interval lengths need to increase as the scale decreases. Paralleling Figures 6, 9 and 13, Figures 17 and 18 and Table 7 show the performance for the smaller scale model with parameter four-tuple $(\bar{\lambda}, \beta, T, B) = (20, 5, 100, 0.01)$ and randomization parameter $\sigma = 0.32$ and averaging parameter $\Delta = 0.8$, the same as for blocking target $B = 0.1$.

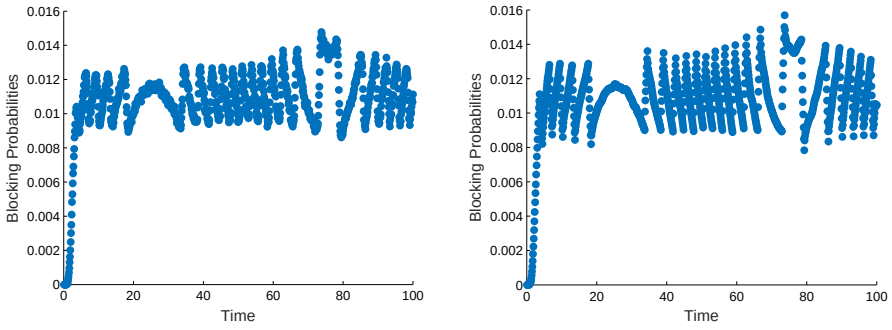


Figure 17: Simulation estimates of the blocking probabilities in the nonstationary $M_t/M/s_t/0$ model with parameter four-tuple $(20, 5, 100, 0.01)$ using randomization with $\sigma = 0.32$ (left) and averaging with $\Delta = 0.8$ (right)

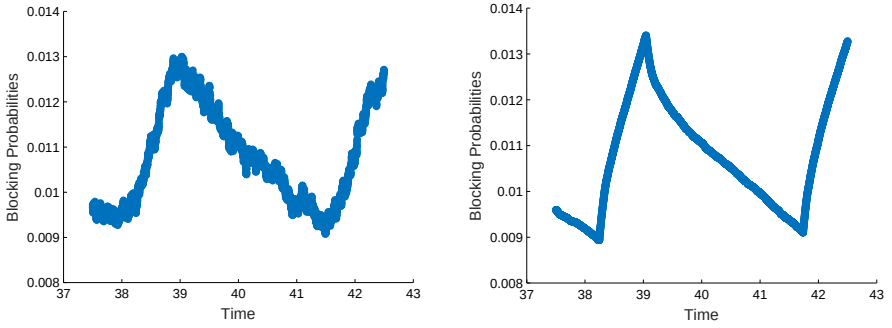


Figure 18: A careful examination of the time interval $[37.5, 42.5]$ for the case $(20, 5, 100, 0.01)$ in the setting of Figure 17: randomization for $\sigma = 0.32$ (left) and averaging for $\Delta = 0.8$ (right)

Next, Paralleling Figures 13 and 16, Figure 20 shows that the parameters $\sigma$ and $\delta$ can be made larger without penalty; they are increased from $(0.32, 0.80)$ in Figure 17 by a factor of 12 to $(3.84, 9.60)$. Indeed, given that stabilization is achieved for $\Delta = 0.8$, as shown in Figure 17, higher values of $\Delta$ can only smooth out the estimate, giving less fluctuation. As before, the relatively large fluctuation at time 80 is consistent with Figure 17 for the smaller averaging parameters. But we do not achieve that benefit for randomizing the staffing times over wider intervals. Nevertheless,

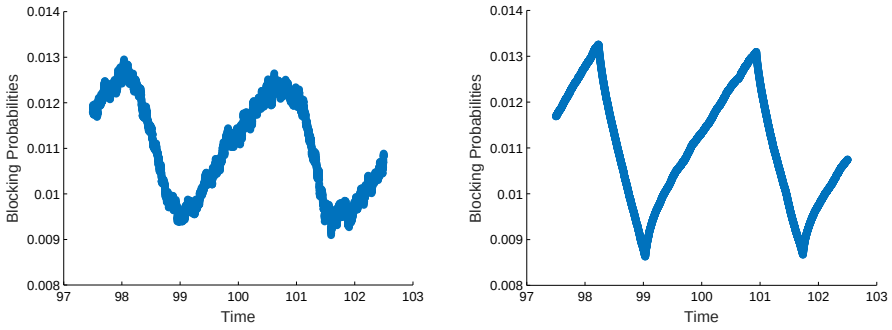


Figure 19: A careful examination of the time interval $[97.5, 102.5]$ for the case $(20, 5, 100)$ in the setting of Figure 17: randomization for $\sigma = 0.32$ (left) and averaging for $\Delta = 0.8$ (right)

Table 7: Simulation estimates of the blocking probabilities over four intervals of length 5 each containing at least one staffing change, for the $M_t/M/s_t/0$ model with $\mu = 1$, $\lambda(t)$ in (5.1) with parameter four-tuple $(\bar{\lambda}, \beta, T, B) = (20, 5, 100, 0.01)$ using the MOL staffing with randomization (left) and averaging (right). The minimum, average and maximum values over a unit interval are shown.

| estimated blocking probabilities over intervals of length 5 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| staffing change | | | randomization: $\sigma = 0.32$ | | | averaging: $\Delta = 0.8$ | | |
| time | from | to | min | average | max | min | average | max |
| 38.6450 | 34 | 33 | 0.0091 | 0.0108 | 0.0130 | 0.0089 | 0.0108 | 0.0134 |
| 42.1380 | 33 | 32 | same | same | same | same | same | same |
| 59.1260 | 27 | 26 | 0.0092 | 0.0109 | 0.0133 | 0.0090 | 0.0109 | 0.0141 |
| 62.3710 | 26 | 25 | same | same | same | same | same | same |
| 89.7040 | 25 | 26 | 0.0090 | 0.0115 | 0.0135 | 0.0085 | 0.0115 | 0.0135 |
| 98.6320 | 28 | 29 | 0.0091 | 0.0110 | 0.0130 | 0.0086 | 0.0110 | 0.0133 |
| 101.3350 | 29 | 30 | same | same | same | same | same | same |

since the cycles are so long, the larger value of $\sigma$ does not hurt.

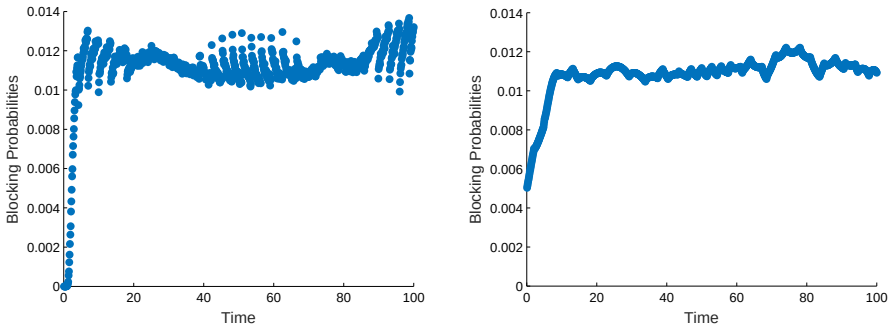


Figure 20: Simulation estimates of the blocking probabilities in the nonstationary $M_t/M/s_t/0$ model with parameter triple $(20, 5, 100)$ having average arrival rate $\bar{\lambda} = 20$ with the staffing algorithm for target $B = 0.01$ using randomization with $\sigma = 3.84$ (left) and averaging with $\Delta = 9.60$ (right).

### 5.6 Short Cycles: $T = 10$

We now consider the more challenging case of shorter cycles of length $T = 10$ instead of $T = 100$. Now the arrival rate changes 10 times more quickly. Table 3 shows that the corresponding times between staffing changes are now much less. Evidently, the parameters $\sigma$ and $\Delta$ need to be reduced. We find that the initial smaller parameters used for $T = 100$ continue to work for $T = 10$, but we no longer have the freedom to increase these parameters. Indeed, the range of good averaging parameters is significantly less.

#### 5.6.1 Larger Scale

We first show the good performance for the nonstationary $M_t/M/s_t/0$ model with parameter triple $(100, 25, 10)$ with blocking probability targets $B = 0.1$ and $B = 0.01$. Figures 21 and 22 show plots of the time-varying blocking with the same parameters $\sigma = 0.08$ and $\Delta = 0.20$ as before in Figures 6 and 13.

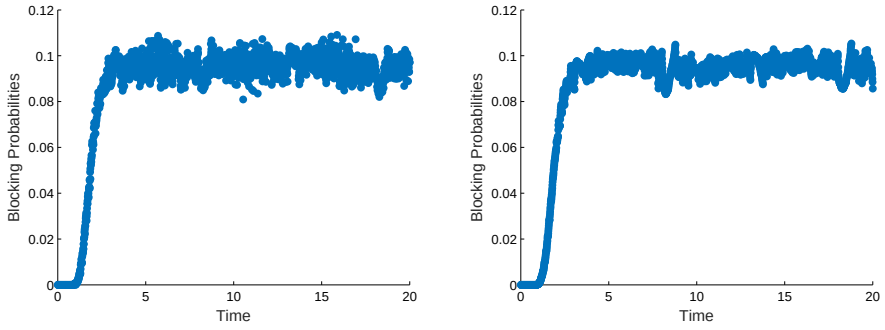


Figure 21: Simulation estimates of the blocking probabilities in the nonstationary $M_t/M/s_t/0$ model with parameter triple $(100, 25, 10)$ having average arrival rate $\bar{\lambda} = 100$ with the staffing algorithm for target $B = 0.1$ using randomization with averaging parameters $\sigma = 0.08$ (left) and $\Delta = 0.20$ (right).

We now illustrate the performance degradation for much smaller or large parameters in the case with target $B = 0.01$. Figure 23 shows the performance degradation if $\sigma$ is reduced, while Figure 24 shows the performance degradation if $\Delta$ is reduced.

As noted before, there is no difficulty if we increase $\Delta$ above 0.20, but there is a problem if we increase $\sigma$ above 0.08. Figure 25 shows the performance degradation with increased $\sigma$.

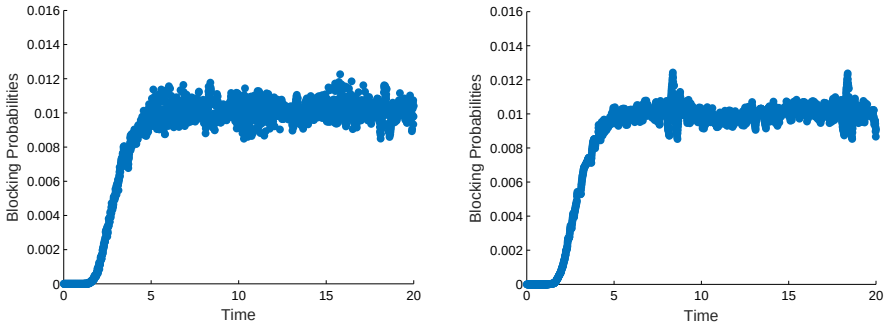


Figure 22: Simulation estimates of the blocking probabilities in the nonstationary $M_t/M/s_t/0$ model with parameter triple $(100, 25, 10)$ having average arrival rate $\bar{\lambda} = 100$ with the staffing algorithm for target $B = 0.01$ using randomization with averaging parameters $\sigma = 0.08$ (left) and $\Delta = 0.20$ (right).

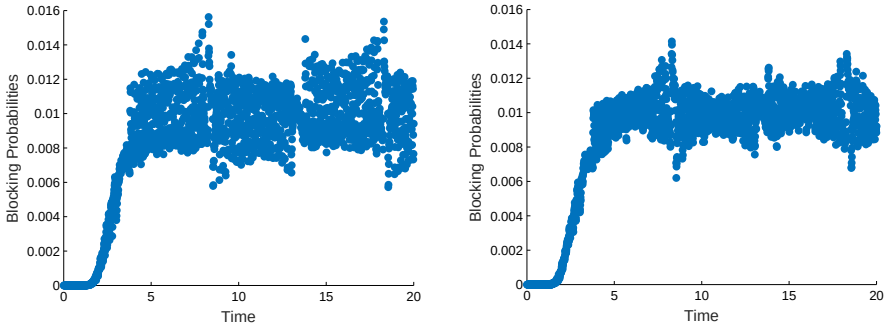


Figure 23: Simulation estimates of the blocking probabilities in the nonstationary $M_t/M/s_t/0$ model with parameter triple $(100, 25, 10)$ having average arrival rate $\bar{\lambda} = 100$ with the staffing algorithm for target $B = 0.01$ using randomization with too small averaging parameters $\sigma = 0.01$ (left) and $\sigma = 0.02$ (right).

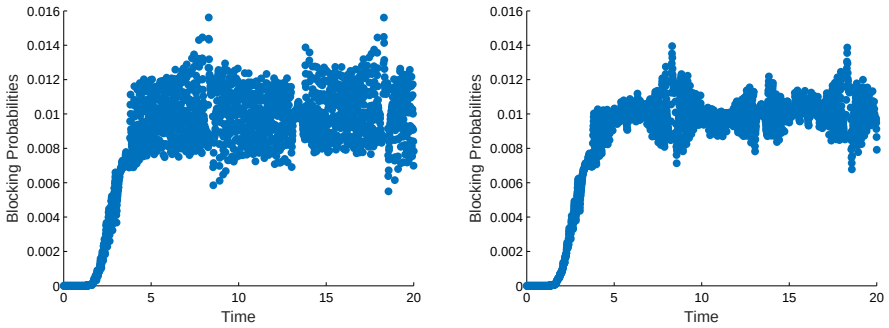


Figure 24: Simulation estimates of the blocking probabilities in the nonstationary $M_t/M/s_t/0$ model with parameter triple $(100, 25, 10)$ having average arrival rate $\bar{\lambda} = 100$ with the staffing algorithm for target $B = 0.01$ using randomization with too small averaging parameters $\Delta = 0.04$ (left) and $\Delta = 0.08$ (right).

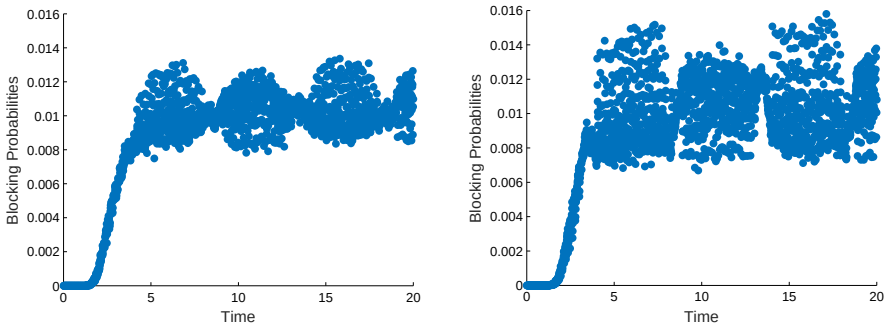


Figure 25: Simulation estimates of the blocking probabilities in the nonstationary $M_t/M/s_t/0$ model with parameter triple $(100, 25, 10)$ having average arrival rate $\bar{\lambda} = 100$ with the staffing algorithm for target $B = 0.01$ using randomization with too large averaging parameters $\sigma = 0.20$ (left) and $\sigma = 2.0$ (right).

### 5.6.2 Smaller Scale

We now show the performance for the smaller scale cases with parameter triple $(20, 5, 10)$ and blocking probability targets $B = 0.1$ and $B = 0.01$. We use $\sigma = 0.32$ and $\Delta = 0.80$, just as we did for $T = 100$ before. Figures 26 and 27 show the performance.

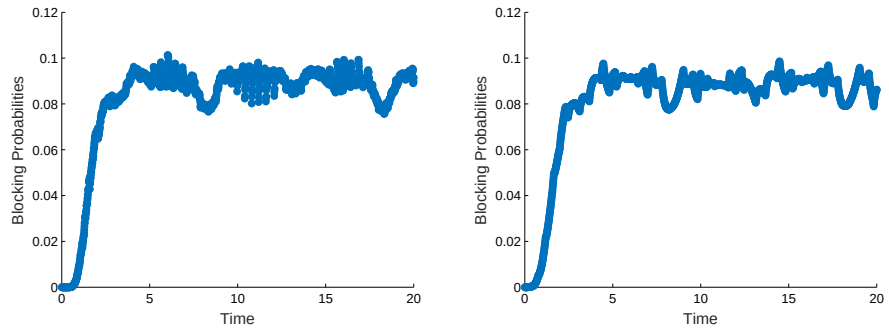


Figure 26: Simulation estimates of the blocking probabilities in the nonstationary $M_t/M/s_t/0$ model with parameter triple $(20, 5, 10)$ having average arrival rate $\bar{\lambda} = 20$ with the staffing algorithm for target $B = 0.1$ using randomization with averaging parameters $\sigma = 0.32$ (left) and $\Delta = 0.80$ (right).

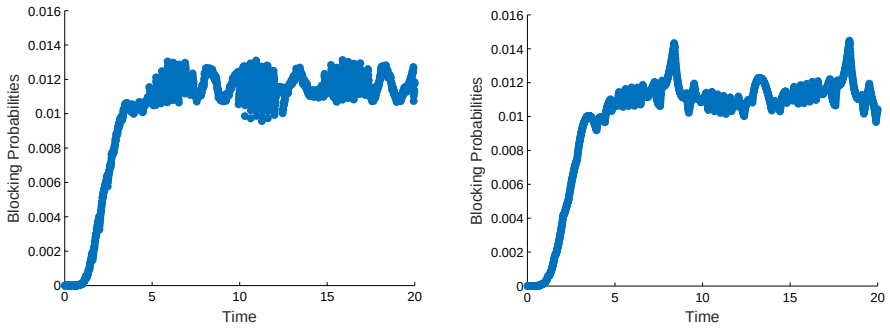


Figure 27: Simulation estimates of the blocking probabilities in the nonstationary $M_t/M/s_t/0$ model with parameter triple $(20, 5, 10)$ having average arrival rate $\bar{\lambda} = 20$ with the staffing algorithm for target $B = 0.01$ using randomization with averaging parameters $\sigma = 0.32$ (left) and $\Delta = 0.80$ (right).

# 6 Experiments for Non-Exponential Service-Time Distributions

Just as in [13], the staffing method here extends to non-Markov models, as we confirmed in several experiments with non-exponential service-time distributions. To illustrate, we discuss the cases of hyperexponential and deterministic service times. In this section we consider the same model except that we change the service-time distribution from exponential ($M$) to hyperexponential ($H_2$) and deterministic ($D$), keeping the mean at 1. We know that the stationary $M/GI/s/0$ has the insensitivity property, implying that the steady-state number in system and the blocking probability depend on the service-time distribution only through its mean. However, as shown by [1], this insensitivity property is *not* inherited by the nonstationary $M_t/GI/s/0$ model.

First, we consider $H_2(1, 4, bm)$ service times, which are mixtures of two exponential distributions, having mean 1, scv $c^2 = 4$ and balanced means; see (3.7) on p. 137 in [24]. The two means are 4.437 and 0.563 with probability 0.1127 on the first. The staffing formula is given in [3] with corrections on p. 506 of [12]. Figure 28 shows the performance of our algorithm for the nonstationary $M_t/H_2(1, 4, bm)/s_t/0$ model with parameter triple $(100, 25, 10)$ having average arrival rate $\bar{\lambda} = 100$ with the staffing algorithm for target $B = 0.1$ using randomization with $\sigma = 0.08$ (left) and averaging with $\Delta = 0.2$ (right). We should expect that the approach to steady state with $H_2$

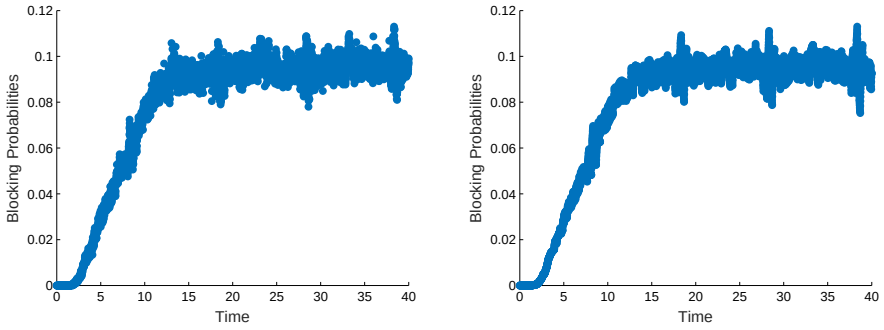


Figure 28: Simulation estimates of the blocking probabilities in the nonstationary $M_t/H_2/s_t/0$ model with parameter triple $(100, 25, 10)$ having average arrival rate $\bar{\lambda} = 100$ with the staffing algorithm for target $B = 0.1$ using randomization with $\sigma = 0.08$ (left) and averaging with $\Delta = 0.2$ (right).

service takes about 4 times longer than for $M$ service because one of the component mean service times is 4.437. That expectation is confirmed by Figures 21 and 28.

We next consider the $M_t/D/s_t/0$ model with the same arrival rate function and the same staffing algorithm in two cases: a long cycle with $T = 100$ and a short cycle with $T = 10$. Paralleling previous figures, Figure 29 examines the performance of the $M_t/D/s_t/0$ model with parameter triple

$(\bar{\lambda}, \beta, T) = (100, 25, 10)$ ($\gamma = 0.628$) using the MOL staffing with target 0.10 and randomization with $\sigma = 0.08$ (left) and averaging with $\Delta = 0.2$ (right). Figure 29 shows that the time-varying blocking probability is again stabilized after an initial transient that is over at about time 3. We see that the initial transient is quite different, but after it is over, the blocking is stable, just as before.

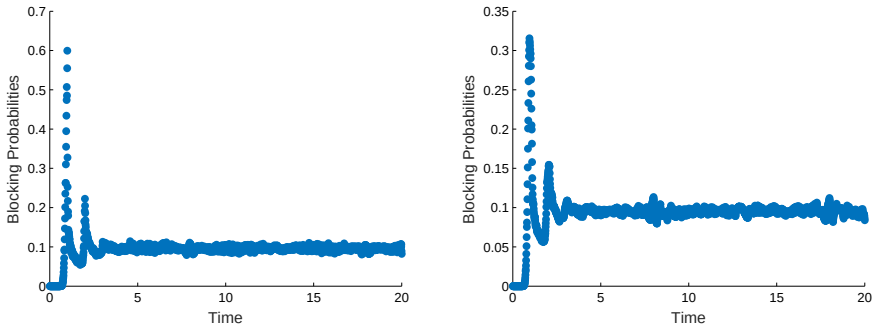


Figure 29: Simulation estimates of the blocking probabilities in the nonstationary $M_t/D/s_t/0$ model with parameter triple $(100, 25, 10)$ having average arrival rate $\bar{\lambda} = 100$ with the staffing algorithm for target $B = 0.1$ using randomization with $\sigma = 0.08$ (left) and averaging with $\Delta = 0.2$ (right).

Paralleling previous tables, Table 8 examines the performance in the case of the $M_t/D/s_t/0$ model with parameter triple $(\bar{\lambda}, \beta, T) = (100, 25, 100)$ ($\gamma = 0.0628$) using the MOL staffing with target 0.10 and randomization with $\sigma = 0.08$ (left) and averaging with $\Delta = 0.2$ (right).

Table 8: Simulation estimates of the blocking probabilities over four unit intervals each containing one staffing change, for the $M_t/D/s_t/0$ model with $\mu = 1$, $\lambda(t)$ in (5.1) with parameter triple $(\bar{\lambda}, \beta, T) = (100, 25, 100)$ ($\gamma = 0.0628$) using the MOL staffing with target 0.10 and randomization with $\sigma = 0.08$ (left) and averaging with $\Delta = 0.2$ (right). The minimum, average and maximum values over a unit interval are shown.
with

| estimated blocking probabilities over intervals of length 1 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| staffing change | | | randomization: $\sigma = 0.08$ | | | averaging: $\Delta = 0.2$ | | |
| time | from | to | min | average | max | min | average | max |
| 39.5260 | 112 | 111 | 0.0820 | 0.0975 | 0.1137 | 0.0856 | 0.0960 | 0.1090 |
| 40.4311 | 111 | 110 | same | same | same | same | same | same |
| 59.7200 | 85 | 84 | 0.0838 | 0.0948 | 0.1079 | 0.0849 | 0.0929 | 0.1058 |
| 89.6330 | 82 | 83 | 0.0818 | 0.0962 | 0.1130 | 0.0790 | 0.0944 | 0.1025 |
| 99.5230 | 95 | 96 | 0.0817 | 0.0955 | 0.1097 | 0.0819 | 0.0932 | 0.1025 |
| 100.2210 | 96 | 97 | same | same | same | same | same | same |

The service-time distribution can make a significant difference in the staffing with the $M_t/GI/s_t/0$ model having mean service time 1. Figure 30 shows that it makes a big difference with parameter four-tuple $(\bar{\lambda}, \beta, T, B) = (100, 25, 10, 0.1)$ having a short cycle with $T = 10$, but it makes hardly difference at all for the corresponding case with $T = 100$. Because the PSA performs well for

long cycles, there is almost insensitivity for long cycles. The maximum difference is 7 servers with $T = 10$, but only a single server for $T = 100$.

As observed in [1, 26], counter to conventional queueing wisdom, decreasing the variability of the service-time distribution from $M$ to $D$ tends to *increase* the maximum required staffing. It also decreases the minimum required staffing. The lower variability in the service-time distribution tends to amplify the impact of the time-varying arrival rate on congestion. Similarly, the higher variability of the $H_2$ service-time distribution tends to reduce the impact of the time-varying arrival rate on congestion. The greater service-time variability tends to "smooth" or average the impact of the fluctuations in the deterministic arrival-rate function.

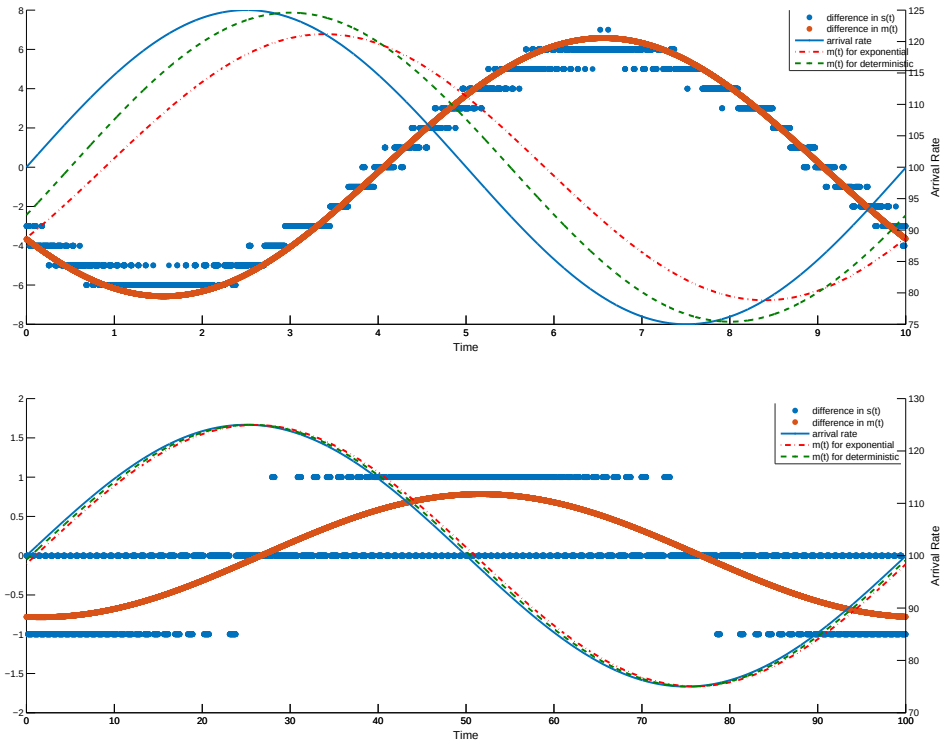


Figure 30: The difference between the staffing functions for the $M_t/M/s_t/0$ and $M_t/D/s_t/0$ models with the sinusoidal arrival rate in (5.1) and mean-1 exponential and deterministic service-time distributions and parameter four-tuples $(\bar{\lambda}, \beta, T, B) = (100, 25, T, 0.1)$ for $T = 10$ (top) and $T = 100$ (bottom).

## 7 Conclusions

In this paper we first used simulation to show that it is not possible to dynamically set staffing levels (specify the number of servers) to stabilize the blocking probability in face of time-varying

demand the same way as for delay models. However, we showed that stabilization can be achieved by the modified-offered-load (MOL) approach, just as for delay models, using one of two forms of averaging, either averaging the times of staffing changes using a mean-zero Gaussian distribution with standard deviation $\sigma$ or averaging the blocking probabilities over suitably time small intervals of width $\Delta$. We conducted extensive simulation experiments to study how to set the two averaging parameters $\sigma$ and $\Delta$. We found that these averaging parameters tend not to depend strongly on the target, by showing that the same parameters work for target blocking probabilities $B = 0.1$ and 0.01. We showed that the parameters need to increase as the scale decreases by considering average arrival rates 100 and 20. We showed that there is more freedom in the choice of the parameters when the arrival rate changes relatively slowly, as when a sinusoidal cycle is $T = 100$, than when the arrival rate function changes relatively rapidly, as when a sinusoidal cycle is $T = 10$. Table 3 shows the parameter ranges for our main examples. These show that the two parameters should be related approximately by $\Delta = 2.5\sigma$. Overall, stabilization is achieved when one of these forms of averaging is combined with the MOL approach.

**Acknowledgment** The first and third authors conducted this research as undergraduates at Columbia University. The second author received support from NSF grants CMMI 1066372 and 1265070.

## References

[1] Davis, J. L., Massey, W. A. and Whitt, W. (1995). Sensitivity to the service-time distribution in the nonstationary Erlang loss model. *Management Sci* 41(6):1107–1116.

[2] Defraeye, M. and van Nieuwenhuyse, I. (2013). Controlling excessive waiting times in small service systems with time-varying demand: an extension of the ISA algorithm. *Decision Support Systems* 54(4):1558–1567.

[3] Eick, S. G., Massey, W. A. and Whitt, W. (1993). $M_t/G/\infty$ queues with sinusoidal arrival rates. *Management Sci* 39:241–252.

[4] Feldman, Z., Mandelbaum, A., Massey, W. A. and Whitt, W. (2008). Staffing of time-varying queues to achieve time-stable performance. *Management Sci* 54(2):324–338.

[5] Green, L. V., Kolesar, P. J. and Whitt, W. (2007). Coping with time-varying demand when setting staffing requirements for a service system. *Production Oper Management* 16:13–29.

[6] Grier, N., Massey, W. A., McKoy, T. and Whitt, W. (1997). The time-dependent Erlang loss model with retrials. *Telecommunications Systems* 7:253–265.

[7] Hampshire, R. C. and Massey, W. A. (2010). Dynamic optimization with applications to dynamic rate queues. *Tutorials in Operations Research* 27:208–247.

[8] Hampshire, R. C., Massey, W. A. and Wang, Q. (2009). Dynamic pricing to control loss systems. *Prob Eng Inf Sci* 23:357–383.

[9] He, B., Liu, Y. and Whitt, W. (2015). Staffing a service system with non-Poisson nonstationary arrrivals. Columbia University, http://www.columbia.edu/∼ww2040/allpapers.html.

[10] Jagerman, D. L. (1975). Nonstationary blocking in telephone traffic. *Bell System Tech J* 54:625–661.

[11] Jennings, O. B., Mandelbaum, A., Massey, W. A. and Whitt, W. (1996). Server staffing to meet time-varying demand. *Management Sci* 42:1383–1394.

[12] Kim, S.-H. and Whitt, W. (2013). Estimating waiting times with the time-varying little's law. *Probability in the Engineering and Informational Sciences* 27:471–506.

[13] Li, A. and Whitt, W. (2014). Approximate blocking probabilities for loss models with independence and distribution assumptions relaxed. *Performance Evaluation* 80:82–101.

[14] Liu, Y. and Whitt, W. (2012). The $G_t/GI/s_t+GI$ many-server fluid queue. *Queueing Systems* 71:405–444.

[15] Liu, Y. and Whitt, W. (2012). Stabilizing customer abandonment in many-server queues with time-varying arrivals. *Oper Res* 60:1551–1564.

[16] Liu, Y. and Whitt, W. (2014). Stabilizing performance in a service system with time-varying arrivals and customer feedback. Columbia University, http://www.columbia.edu/ ww2040.

[17] Liu, Y. and Whitt, W. (2014). Stabilizing performance in networks of queues with time-varying arrival rates. *Probability in the Engineering and Informational Sciences* 28:419–449.

[18] Massey, W. A. and Whitt, W. (1994). An analysis of the modified offered load approximation for the nonstationary Erlang loss model. *Ann Appl Probab* 4:1145–1160.

[19] Massey, W. A. and Whitt, W. (1994). A stochastic model to capture space and time dynamics in wireless communication systems. *Prob in the Engineering and Informational Sciences* 8:541–569.

[20] Melamed, B. and Whitt, W. (1990). On arrivals that see time averages. *Operations Research* 38(1):156–172.

[21] Pender, J. (2015). Nonstationary loss queues via cumulant moment approximations. *Probability in the Engineering and Information Sciences* 29:27–49.

[22] Pender, J. and Massey, W. A. (2014). Approximating and stabilizing dynamic rate Jackson networks with abandonment. Cornell University, http://people.orie.cornell.edu/jpender/.

[23] Stolletz, R. (2008). Approximation of the nonstationary $M(t)/M(t)/c(t)$ queue using stationary models: the stationary backlog-carryover approach. *European Journal of Operations Research* 190(2):478–493.

[24] Whitt, W. (1982). Approximating a point process by a renewal process: two basic methods. *Oper Res* 30:125–147.

[25] Whitt, W. (1984). Heavy-traffic approximations for service systems with blocking. *AT&T Bell Lab Technical Journal* 63:689–708.

[26] Wolfe, R. W. (1977). The effect of service time regularity on system performance. In Chandy, K. M. and Reiser, M. (eds.), *Computer Performance*. Amsterdam: North-Holland, pp. 297–304.

[27] Yom-Tov, G. and Mandelbaum, A. (2014). Erlang R: a time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing and Service Oper Management* 16(2):283–299.