Many-Server Loss Models with Non-Poisson Time-Varying Arrivals

Ward Whitt, Jingtong Zhao

Department of Industrial Engineering and Operations Research, Columbia University, New York, New York 10027

Received 26 May 2016; revised 13 March 2017; accepted 2 May 2017
DOI 10.1002/nav.21741
Published online 25 May 2017 in Wiley Online Library (wileyonlinelibrary.com).

Abstract: This article proposes an approximation for the blocking probability in a many-server loss model with a non-Poisson time-varying arrival process and flexible staffing (number of servers) and shows that it can be used to set staffing levels to stabilize the time-varying blocking probability at a target level. Because the blocking probabilities necessarily change dramatically after each staffing change, we randomize the time of each staffing change about the planned time. We apply simulation to show that (i) the blocking probabilities cannot be stabilized without some form of randomization, (ii) the new staffing algorithm with randomiation can stabilize blocking probabilities at target levels and (iii) the required staffing can be quite different when the Poisson assumption is dropped.

Keywords: nonstationary loss models; staffing a service system; capacity planning; queues with time-varying arrival rates; modified offered load; approximation; non-Poisson nonstationary arrival processes

1. INTRODUCTION

For over 100 years, the multi-server Erlang ($M/M/s/\infty$ delay and $M/M/s/0$ loss) models have been used to help set capacities in multi-server service systems, ranging from telephone exchanges to customer contact centers and hospital emergency departments [6, 16]. However, these systems often have two features that deviate significantly from these models: (i) a strongly time-varying arrival rate over each day, and (ii) non-Markov (or non-Poisson and nonexponential) stochastic variability.

There tends to be less interest in dynamic staffing in response to time-varying arrival rates for loss models than for delay models, because the staffing usually is less flexible in loss systems, for example, as with circuits in a telecommunications system. When staffing should be regarded as fixed, it is natural to consider controlling the demand instead, for example, by dynamic pricing, as has been considered in [19, 20] and references therein. However, there often is more flexibility in staffing than we might at first think. For example, loss models are natural for an ambulance base serving several hospitals as in [62], for the rooms in a hotel as in [34], and for a bike-sharing system as in [22]. In a short-time scale the available resources are fixed, but in a longer time scale adjustments can be made. For example, the number of available ambulances or bicycles may be dynamic, because transfers can be made. More generally, with tight capacity constraints, there is growing interest in reconfigurable resources; for example, the way rooms are used in hotels. This reconfiguration typically cannot be done rapidly enough to respond instantaneously to current demand, but it may be done rapidly enough to respond to anticipated demand in the near future, for example, later on the same day.

With time-varying arrival rates, it is common to assume that the arrival process is a nonhomogeneous Poisson process (NHPP, denoted M_t), but there is growing Evidence of deviation from the Poisson property from data analysis (often called over-dispersion) [3, 23, 28, 30]. For the loss systems considered here, there is extra motivation for considering non-Poisson variability in arrival processes because arrival processes are often overflows from other loss systems, which have greater variability than Poisson, as reviewed in [35], where staffing methods were considered for the stationary $G/GI/s/0$ loss model.

There is a significant body of research aimed at addressing time-varying arrival rates and non-Markov stochastic variability separately, as illustrated by [16] and [35]. First, [16] reviews staffing methods for NHPP's, while [35] reviews staffing methods for stationary non-Markov $G/G/s/0$ loss models. This article is an effort to simultaneously address *both* of these complicating features for loss models. In particular, we develop an effective time-varying staffing strategy (dynamically controlling the number of servers) to stabilize blocking probabilities at target levels in an $G_t/GI/s_t/0$ loss

Correspondence to: Ward Whitt (ww2040@columbia.edu)

model having staffing flexibility and an arrival process that is both non-Poisson and nonstationary (the G_t) as well as a nonexponential service-time distribution (the GI).

This article is directly a sequel to [36] in which we developed an effective time-varying staffing strategy to stabilize blocking probabilities at target levels in an $M_t/GI/s_t/0$ loss model having staffing flexibility and a time-varying arrival rate, with an arrival process that is an NHPP (the M_t) as well as a general service-time distribution. To meet the significant challenge presented by the extension to G_t arrival processes, we apply the modified-offered-load (MOL) method together with results for the stationary $G/G/s/0$ model, drawing on [35]. This article also relates to [21] which stabilized the performance in $G_t/GI/s_t + GI$ delay models.

Because we treat a very general model in this article, we exploit a variety of results and techniques developed over many years. Consequently, In Section 2, we also provide a survey the literature. Broad surveys of the literature on time-varying queues have recently been provided by Defraeye and van Nieuwenhuyse [9] and Schwarz et al., [63]. These are much broader than the earlier surveys in Massey [45] and Green et al., [16], which are more directly related to this article. A good account of the remarkable early work on the Erlang models by Erlang appears in [6], while C. Palm's 1943 early work on time-varying queues appears in [54].

1.1. Stochastic and Deterministic Stochastic Models

The extension to general $G_t/GI/s/0$ loss models has an important implication for modeling. The classical Erlang models and their generalizations with time-varying M_t arrival processes can be regarded as *deterministic stochastic models*, because the exponential service-time distributions are fully specified by their deterministic means, while the Poisson arrival processes are fully specified by their deterministic rates. There is no separate specification of the extent of the stochastic variability. In contrast, when we include GI service times and G_t arrival processes, the $G_t/GI/s/0$ model directly requires that the extent of the stochastic variability be considered when constructing the model. The extension to the $G_t/GI/s/0$ model is important, in part, because it invites examining the stochastic variability more closely. It is significant that more complex forms of stochastic variability have been identified in practice, for example, see [3, 7, 23, 28, 30], and they can have a big impact on performance, as we will show next.

1.2. Non-Poisson Arrivals Can Make a Big Difference

In this article, we show that the non-Poisson property can make a big difference in the staffing. To demonstrate this important conclusion, we now show the arrival-rate function and staffing functions for eight cases in Fig. 1. The arrival-rate function for all our examples is the sinusoidal arrival-rate function

$$\lambda(t) = \bar{\lambda} + \beta \sin(\gamma t), \quad t \geq 0, \tag{1.1}$$

with average arrival rate $\bar{\lambda}$, amplitude β, and cycle length (or period) $T = 2\pi/\gamma$ (or equivalently, frequency $\gamma = 2\pi/T$). However, we emphasize that the algorithm is by no means limited to sinusoidal arrival-rate functions. We fix the time units by letting the mean service time be $\mu^{-1} \equiv E[S] = 1$. In that scale, we consider both long cycles having $T = 100$ and short cycles having $T = 10$. We consider two different performance targets: higher Quality of Service (QoS) with a blocking target of $B = 0.01$ and lower QoS with a blocking target of $B = 0.1$.

Figure 1 displays the staffing functions for two arrival-rate functions, the case of long cycles ($T = 100$ on top) and the case of short cycles ($T = 10$ on bottom). In both cases, $\bar{\lambda} = 100$ (large scale) and $\beta = 25$ (moderate fluctuations). For each case, four staffing functions are shown. There are two forms of stochastic variability: (i) an NHPP arrival process and (ii) a more variable $(H_2)_t$ arrival process, constructed as a time transformation of a renewal process with $H_2 \equiv H_2(4)$ (hyperexponential, mixture of two exponential distributions) having a squared coefficient of variation (scv, variance divided by the square of the mean) $c_a^2 = 4.0$. (We explain in Section 3.1.) Figure 1 shows the $2 \times 2 \times 2 = 8$ cases 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.

The five curves in each plot of Fig. 1 are easy to distinguish: The arrival rate function is the only smooth function; the staffing functions are all integer-valued. The two $(H_2)_t$ cases appear in the thicker blue curves, which lie above the corresponding M_t dashed red curves. In each case, the staffing is necessarily higher to meet the lower blocking target $B = 0.01$ and higher for the more variable arrival process.

Figure 1 shows that the performance target makes the biggest difference; the required staffing is much higher with the target $B = 0.01$ than with the target $B = 0.1$. Second, Fig. 1 shows that the variability of the arrival process also makes a big difference; the required staffing for the more variable $(H_2)_t$ arrival process is much higher than the required staffing for the corresponding M_t arrival process when $B = 0.01$. (The difference is less with target $B = 0.1$.) (The staffing functions in Fig. 1 do not include the randomization discussed in Section 1.3.)

Consistent with previous research, for example, formulas (14) and (15) in [11], the staffing plots in Fig. 1 show that there is a time lag in the peak staffing after the peak arrival rate of about 1 mean service time, because the customers remain in the system after their arrival for their service times. This

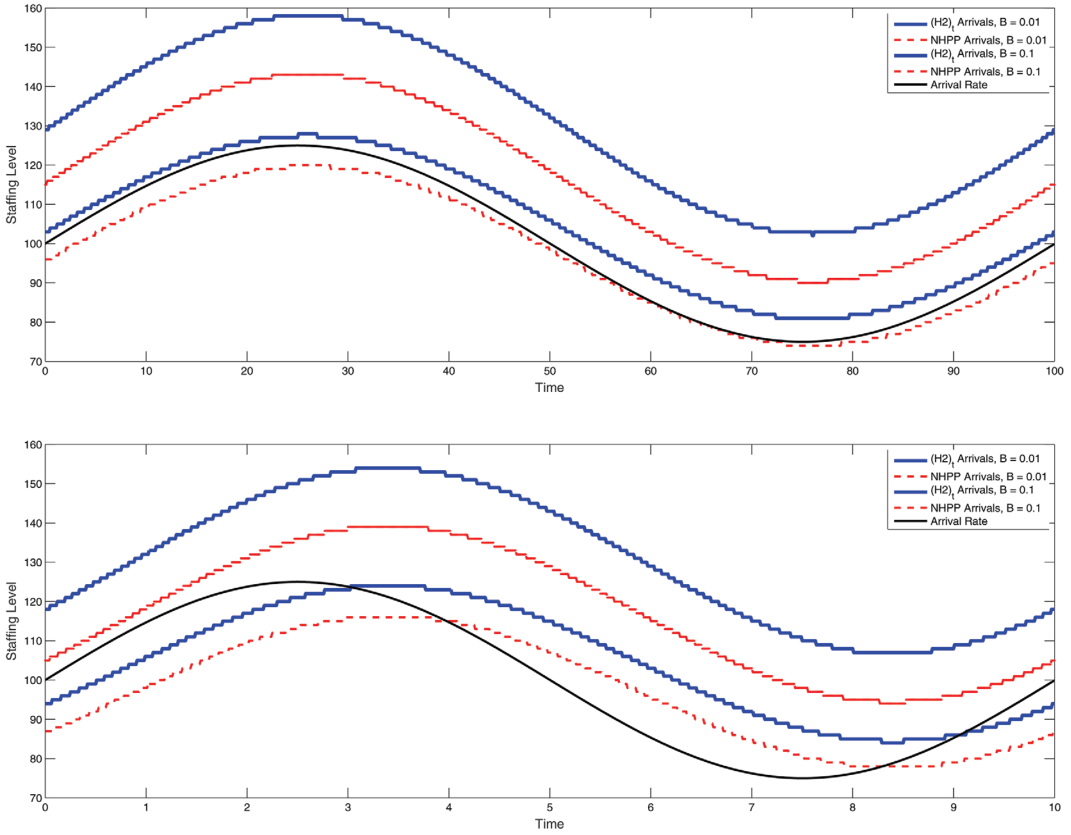

Figure 1. The MOL staffing functions and the sinusoidal arrival rate function in (1.1) for $G_t/M/s/0$ models with M_t and $(H_2)_t$ arrival processes, for $\mu = 1$, average arrival rate $\bar{\lambda} = 100$, amplitude $\beta = 25$ and two blocking probability targets $B = 0.1$ and $B = 0.01$: for long cycles $T = 100$ (top) and short cycles $T = 10$ (bottom). [Color figure can be viewed at wileyonlinelibrary.com]

time lag of about 1 is much more noticeable for short cycles than for long cycles. For long cycles, we can staff by using the pointwise stationary approximation (PSA), that is, by using the stationary model with arrival rate $\lambda(t)$ at time t (but that model is $G/GI/s/0$, which requires an approximation such as in [35]). Figure 1 shows that PSA should be reasonably effective for $T = 100$, but not for $T = 10$. Thus, we should anticipate that staffing to stabilize blocking with $T = 10$ is much more difficult than with $T = 100$.

1.3. The Need for Extra Randomization or Averaging

Stabilizing blocking probabilities in loss models turns out to be fundamentally more difficult than stabilizing delay probabilities or abandonment probabilities in delay models. Indeed, it is *not* possible to stabilize time-varying blocking probabilities by only choosing an appropriate deterministic staffing function $s(t)$, because blocking probabilities necessarily change dramatically at the time of each staffing change. First, the blocking probability decreases to 0 immediately after a staffing increase, because there necessarily is space for another arrival; second, simulations show that the blocking probability also increases sharply after each staffing decrease. That is illustrated by the plot on the left in Fig. 2 for the $(H_2)_t/M/s_t/0$ model with the sinusoidal arrival rate in (1.1) having parameters $\bar{\lambda} = 100$, $\beta = 25$, and $T = 2\pi/\gamma = 100$ (and the fixed $\mu = 1$), blocking probability target $B = 0.1$ and the staffing functions shown in Fig. 1 (top).

As in [36], we either (i) randomize the time of each staffing change or (ii) average the blocking probabilities in a small interval about the time of each fixed staffing change. Fig. 2 (right) shows simulation estimates of the blocking probability after randomization has been applied. (We elaborate in Section 4.) To interpret the left-hand plot in Fig. 2 without randomization, note that Fig. 1 shows that the staffing is nonincreasing in the middle portion, roughly over [22, 78], and is increasing outside that interval. That explains why we have the jumps up (down) in the middle (outside).

Here is how the rest of this article is organized. In Section Section 2, we develop the approximation for the time-varying blocking probability in the general $G_t/GI/s_t/0$ loss model. We provide a survey of the literature to put our research in context. In Section 3, we describe our simulation experiments and in Section 4, we present our simulation results. In Section

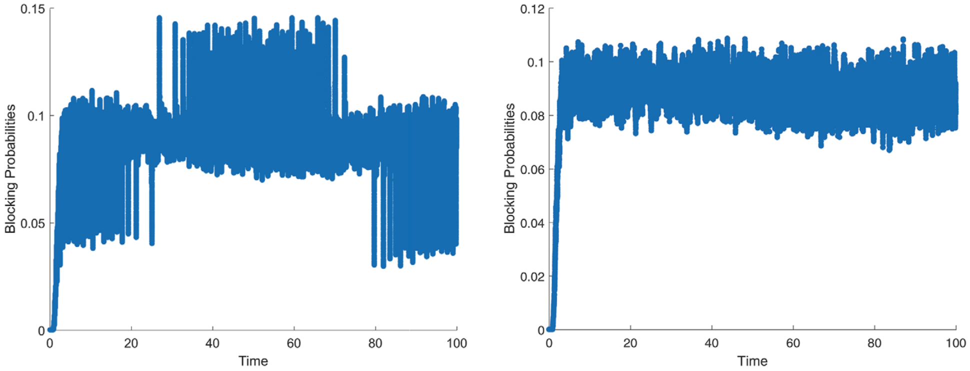

Figure 2. Simulation estimates of the blocking probabilities in the nonstationary $(H_2)_t/M/s_t/0$ model for the sinusoidal arrival rate in (1.1) having parameters $\bar{\lambda} = 100$, $\beta = 25$, and $T = 2\pi/\gamma = 100$ (and the fixed $\mu = 1$), blocking probability target $B = 0.1$ and the staffing functions shown in Fig. 1 (top): before randomization (left) and after randomization (right). [Color figure can be viewed at wileyonlinelibrary.com]

5, we show that the staffing algorithm remains effective for very high blocking targets, and explain why blocking targets usually should be lower than delay probability targets in delay models. In Section 6, we provide a simple approximation for the performance during the initial transient period when we start with an empty system. In Section 7, we discuss time congestion. Finally, in Section 8, we draw conclusions. Additional experimental results are presented in the appendix [73] available from the authors' web pages.

2. STAFFING IN TIME-VARYING MANY-SERVER LOSS MODELS

This section has nine short subsections: In Section 2.1 we describe a special composition construction for the G_t arrival process; in Section 2.2, we review the results for infinite-server queues; in Section 2.3, we review the MOL approximation; in Section 2.4, we review relevant many-server heavy-traffic (MSHT) limits; in Section 2.5, we show how we implement the MOL approach with the $G_t/GI/s/0$ model; in Section 2.6, we emphasize that the stabilization is easier than approximating performance for a system that is inappropriately staffed; in Section 2.7, we discuss empirical investigations of the square-root-staffing (SRS) formula; in Section 2.8, we show that our model can be fit to data relatively easily; and Section 2.9, we discuss two forms of blocking that arise for non-Poisson and nonstationary arrival processes.

2.1. A Special Composition Construction for the Arrival Process

Let $A(t)$ count the number of arrivals over the interval $[0, t]$ and let $\lambda(t)$ be its deterministic time-varying arrival-rate function, satisfying $0 < \lambda_{LB} \leq \lambda(t) \leq \lambda_{UB} < \infty$ for positive numbers λ_{LB} and λ_{UB}. Let $\Lambda(t)$ be the cumulative arrival-rate function, that is,

$$\Lambda(t) = \int_0^t \lambda(s)ds, \quad t \geq 0. \tag{2.1}$$

We assume that the distribution of the arrival process is approximately Gaussian, that is,

$$A(t) \approx N(\Lambda(t), c_a^2\Lambda(t)) \tag{2.2}$$

for t suitably large, where $N(m, \sigma^2)$ denoted a random variable with a normal (Gaussian) distribution with mean m and variance σ^2. We then characterize the variability of the arrival process by the asymptotic variability parameter c_a^2. In Section 2.8, we show how to exploit the structure in (2.2) to fit the parameter c_a^2 to data.

As a specific construction, we assume that our general nonstationary arrival process A can be represented as the composition of a stochastic counting process N and the cumulative arrival rate function Λ, using the composition function $\circ$, with $(x \circ y)(t) \equiv x(y(t))$, $t \geq 0$; that is,

$$A \equiv N \circ \Lambda \quad \text{or, equivalently,} \quad A(t) \equiv N(\Lambda(t)), \quad t \geq 0, \tag{2.3}$$

where N is a stochastic counting process with nondecreasing non-negative integer-valued sample paths. Recall that this is a standard construction when A is an NHPP; then N is a rate-1 Poisson process and $E[A(t)] = \Lambda(t)$, $t \geq 0$.

Our use of the special composition construction in (2.3) follows [21, 43], but this approach was proposed much earlier in [49] and then again in [14, 51]. As emphasized in Remark 2.2 of [21], this construction is restrictive. In general, we should not expect such a simple one-parameter characterization of variability. The level of variability might also be time-varying. Nevertheless, even in that case, one-parameter characterization of variability might be appropriate locally over subintervals. In any case, that extra level of complexity in the variability is challenging to capture in data analysis, for example, when the model is fit to data, as we discuss in Section 2.8.

As we think of Λ as specifying the deterministic rate of arrivals, it is natural to assume that our stochastic process N is a rate-1 stationary counting process, but we only assume that N obeys a functional central limit theorem (FCLT); that is, we introduce a sequence of processes indexed by n and obtain

$$\hat{N}(t) \equiv n^{-1/2}[N(nt) - nt] \Rightarrow c_a B(t) \quad \text{in} \quad \mathcal{D} \quad \text{as} \quad n \to \infty, \tag{2.4}$$

where $\Rightarrow$ denotes convergence in distribution and $\mathcal{D}$ is the function space of right continuous real-valued functions with left limits as in [69], while $B(t)$ is standard (mean 0 and variance 1) Brownian motion. Thus, N can be very general. It could be a renewal process with mean interarrival time 1 as well as its stationary (or equilibrium) version, as in Section V.3 of [2], which necessarily satisfy the same FCLT in (2.4) [53]. However, it need not be either of those, which means that dependence among the interarrival times is allowed (under regularity conditions implying [2.4]).

As an immediate consequence of (2.3) and (2.4), we have a FCLT for the associated sequence of arrival processes $A_n(t) \equiv N(n\Lambda(t)), t \geq 0, n \geq 1$.

THEOREM 2.1 (FCLT for the arrival process): If conditions (2.3) and (2.4) hold, then

$$\hat{A}_n(t) \equiv n^{-1/2}[A_n(t) - n\Lambda(t)] \Rightarrow c_a B(\Lambda(t)) \quad in \quad \mathcal{D} \quad as \quad n \to \infty. \tag{2.5}$$

PROOF: Starting from the assumed FCLT in (2.4), apply the continuous mapping theorem with the composition function; see Theorems 3.4.1 and 13.2.1 of [69]. □

Theorem 2.1 supports the Gaussian approximation in (2.2).

2.2. Exploiting Infinite-Server Models

As in [21, 36], we exploit the close connection between the time-varying many-server (TVMS) models and the associated TV infinite-server (TVIS) models. The TVIS models are useful because they are remarkably tractable. Key properties of the $M_t/GI/\infty$ TVIS model are reviewed and extended in [11, 47]. Assuming that the system started empty in the distant past (so we can avoid any initial conditions), the number of customers in the system at time t, $Q(t)$, has a Poisson distribution with mean

$$m(t) \equiv E[Q(t)] = \int_0^\infty \lambda(t-s)\bar{G}(s)ds = E[\lambda(t - S_e)]E[S], \quad t \geq 0, \tag{2.6}$$

where $\lambda(t)$ is the deterministic arrival rate at time t, G is the cumulative distribution function (cdf) of a service time S, $\bar{G}(s) \equiv 1 - G(s) \equiv P(S > s)$ and S_e is a random variable with the *stationary-excess* or equilibrium lifetime cdf of G, that is,

$$G_e(x) \equiv P(S_e \leq x) \equiv \frac{1}{E[S]}\int_0^x \bar{G}(s)ds, \quad x \geq 0. \tag{2.7}$$

We call $m(t)$ in (2.6) the *offered load*, because with finitely many servers, it represents the expected number of servers needed if we ignored the capacity constraints (considered the associated IS model). In the stationary case, when λ is a constant, we have the familiar formula $m \equiv m(\infty) = \lambda E[S]$. The second formula in (2.6) shows the TV analog with a random time lag by S_e. To get a rough idea of the impact of the service-time cdf, we can use the mean $E[S_e] = E[S](c_s^2 + 1)/2$; see (2) in [11].

As discussed in Section 3 of [27] and Section 4.3 of [16], if we directly approximate the TVMS model by a TVIS model, then we immediately obtain a Poisson distribution, which leads to the Gaussian approximation $Q(t) \approx N(m(t), m(t))$, where $m(t)$ is the offered load in (2.6) and $N(m, v)$ denotes a random variable with a normal distribution having mean m and variance v. (We have variance equal to the mean because of the Poisson distribution.) If we choose $s(t)$ so that $P(N(m(t), m(t)) > s(t)) = \alpha$, then we obtain the classical SRS formula

$$s(t) = m(t) + \beta^* \sqrt{m(t)}, \tag{2.8}$$

where $\beta^* \equiv P(N(0,1) > \alpha$ is a quality-of-service (QoS) parameter (not to be confused with β in (1.1).

The distribution of $Q(t)$ is more complicated in the $G_t/GI/\infty$ TVIS model, so that analysis becomes much more challenging. The success of the MOL approximation depends partly on the following result from [47].

THEOREM 2.2: (the mean for G_t [47]; see Theorem 2.1 and Remark 2.3): The formula for the mean $m(t) \equiv E[Q(t)]$ in the $M_t/GI/\infty$ model in (2.6) remains unchanged if the arrival process is changed to G_t with the same arrival-rate function.

To approximate the distribution of $Q(t)$, we now rely on the MSHT limit for the $G_t/GI/\infty$ model in [56] and references therein. For that MSHT limit, we consider a sequence of models indexed by n with arrival processes as in Theorem 2.1. Let $Q_n(t)$ be the number of busy servers in model n at time t.

THEOREM 2.3: (MSHT FCLT for the $G_t/GI/\infty$ model from [56]) If conditions (2.3) and (2.4) hold for the sequence of arrival processes in the $G_t/GI/\infty$ model, then

$$n^{-1/2}[Q_n(t) - nm(t)] \Rightarrow X(t) \quad in \quad \mathcal{D} \quad as \quad n \to \infty, \tag{2.9}$$

where $m(t)$ is the offered load in (2.6) and $X(t)$ is a mean-0 Gaussian process with time-varying variance function

$$\begin{aligned} v(t) &\equiv \int_0^\infty \lambda(t-s)V(s)ds \quad with \\ V(s) &\equiv \bar{G}(s) + (c_a^2 - 1)\bar{G}(s)^2 \\ &= m(t) + (c_a^2 - 1)\int_0^\infty \lambda(t-s)\bar{G}(s)^2 ds, \end{aligned} \tag{2.10}$$

so that the ratio of $v(t)$ in (2.10) to $m(t)$ *in* (2.6) (called the time-varying MSHT peakedness) is

$$z(t) \equiv \frac{v(t)}{m(t)} = 1 + (c_a^2 - 1)m(t)^{-1}\int_0^\infty \lambda(t-s)\bar{G}(s)^2 ds. \tag{2.11}$$

PROOF: Assumption 1 of [56] requires that the arrival counting process satisfies a FCLT. That condition is satisfied by virtue of Theorem 2.1 with asymptotic variability parameter c_a^2. □

Theorem 2.3 supports the Gaussian approximation $Q(t) \approx N(m(t), z(t)m(t))$ for $m(t)$ in (2.6) and $z(t)$ in (2.11). Because the TV MSHT peakedness formula in (2.11) is complicated, we approximate it by the MSHT peakedness formula in the associated stationary $G/GI/\infty$ model (letting $m(t)^{-1} \approx \mu/\lambda$ and $\lambda(t-s) \approx \lambda$ in (2.11)) to obtain the MSHT stationary peakedness

$$z \equiv 1 + (c_a^2 - 1)\mu\int_0^\infty \bar{G}(s)^2 ds. \tag{2.12}$$

Formula (2.12) is the MSHT limit of the *peakedness* (ratio of the variance to the mean) in the stationary $G/GI/\infty$ IS model. In general, we use the scv instead of the variance for a non-negative random variable and the peakedness instead of the variance function for a counting process, because they expose the level of variability independent of scale. An exponential random variable has scv 1 and a Poisson process has peakedness 1 for all t, for all possible mean values.

This analysis leads to our final TV approximation for the $G_t/GI/\infty$ model:

$$Q(t) \approx N(m(t), zm(t)), \tag{2.13}$$

where $m(t)$ is given in (2.6) and z is given in (2.12). From (2.13), we see that the time-varying behavior of $Q(t)$ is captured by $m(t)$ in (2.6), while the impact of non-Poisson stochastic variability in the arrival process (which also depends on G) is captured by z in (2.12). The use of the heavy-traffic approximation for z in the stationary $G/GI/s/0$ loss model and the TVMS $G_t/GI/s/\infty$ delay model is discussed and examined, respectively, in [35] and [21].

2.3. The Modified-Offered-Load Approximation

The MOL approximation applies to queues with finitely many servers. It exploits the mean number of busy servers in the associated TVIS model, which is the offered load in (2.6). The MOL approximation for delay was discussed in [8, 12, 16, 21, 27, 38, 40, 41, 74],]. Short surveys of the MOL approach to staffing are given in [70, 71].

The MOL method was originated by Jagerman [26] for the $M_t/M/s/0$ model with a fixed number of servers. Theoretical support for the MOL approximation for that model and the more general $M_t/Ph/s/0$ model were provided in [48]. Peak congestion in $M_t/GI/s/0$ models was studied using TVIS models in [50]. The time-varying performance of the nonstationary loss model with fixed staffing was also discussed in [17, 58].

For both delay and loss models, the MOL approximation is an alternative to two natural simple approximations. The first is the *pointwise-stationary-approximation* (PSA), which is the steady-state distribution of the stationary model using the instantaneous arrival rate at each time. The second approximation is the *simple stationary approximation* (SSA), which uses the stationary model with the long-run average arrival rate. The SSA approximation usually exhibits poor performance whenever the arrival rate fluctuates significantly, but for relatively short service times or, equivalently, for a slowly changing arrival-rate function, the PSA is effective, and is commonly used in practice. However, the PSA deteriorates substantially with longer service times. Figures 1–3 of [27] show the big advantage of the new infinite-server (IS) staffing scheme over PSA and SSA for multi-server delay models with longer service times. (In [27] a direct IS approximation is first proposed, but it is extended to the MOL approximation in Section 4; see the review in [16].)

At each time t, instead of the actual arrival rate $\lambda(t)$ used by PSA, the MOL method use the stationary model with arrival rate

$$\lambda_{mol}(t) \equiv \frac{m(t)}{E[S]}, \tag{2.14}$$

where $m(t)$ is the offered load in (2.6) and $E[S]$ is a mean-service time. There is a simple logic: If the IS model were stationary at time t, then the offered load would be $m(t) =$

$\lambda(t)E[S]$ by Little's law ($L = \lambda W$); the mean $m(t)$ provides a better starting point for a staffing approximation than $\lambda(t)$, because it also accounts for the service-time distribution. The MOL approximation was first used for loss systems in [26].

In [36] the MOL approach was found to be effective for stabilizing the blocking probability in $M_t/GI/s_t/0$ loss models after incorporating the randomization or averaging. Our goal here is to investigate if the MOL approach in [36] remains effective for more general $G_t/GI/s_t/0$ loss models with non-NHPP arrival processes. These parallels [21], which showed that the MOL approach remains effective for $G_t/GI/s_t/\infty$ delay models, with or without customer abandonment.

2.4. Many-Server Heavy-Traffic Limits for Time-Varying Queues

Because no explicit steady-state formulas are available for the general stationary $G/GI/s/0$ loss model, we exploit MSHT limits for loss models. The early papers on heavy-traffic limits for stationary queueing models can be traced from [69]. Particularly relevant are the MSHT limits for the stationary IS model. The seminal paper is the 1965 paper by Iglehart [24]. The early MSHT limits for more general $G/G/\infty$ models can be traced from [56, 57].

The seminal paper on MSHT limits for time-varying queues is Mandelbaum et al. [44]. MSHT limits for stationary loss models were obtained in [4, 5, 64]; as in [35], we will draw on [5].

MSHT limits for queues that switch between overloaded and underloaded regimes are contained in [37, 39]. Recent MSHT limits for complex Markovian TVIS queues and networks of queues have been obtained by [32, 33]. Recent MSHT limits and approximations for complex Markov finite-server models have been obtained by [59].

2.5. Implementing MOL with G_t Arrival Processes

For Markov models, the MOL approximation is easy to implement, because we can apply the appropriate stationary Erlang model at each time point for the required time-varying distribution of the steady-state performance. In contrast, that approach is not available for the general $G_t/GI/s_t/0$ loss model, because exact steady-state performance measures for the stationary $G/GI/s/0$ model are not available. Thus, we rely on MSHT approximations to set staffing levels for the stationary $G/GI/s/0$ model in [35]. We rely heavily on this earlier work in [35] studying staffing methods for the stationary $G/GI/s/0$ model with parmeters λ, μ, and z.

In particular, we apply (18) of [35] to construct the steady-state part of the MOL approximation. In particular, we approximate the blocking probability $B(s, \alpha, z)$ in the stationary $G/GI/s/0$ model as a function of the number s of servers, the offered load $\alpha \equiv \lambda/\mu$, and z is the MSHT limit for the peakedness z in (2.12) by

$$B \approx B(s, \alpha, z) \equiv \sqrt{\frac{z}{\alpha}} \left(\frac{\phi\left((s-\alpha)/\sqrt{\alpha z}\right)}{\Phi\left((s-\alpha)/\sqrt{\alpha z}\right)} \right), \qquad (2.15)$$

where Φ and ϕ are the cdf and pdf respectively of the standard (mean 0, variance 1) Gaussian distribution. Thus, using the MOL logic, here we use the TV blocking approximation

$$B(t) \approx B(s(t), m(t), z) \qquad (2.16)$$

for B in (2.15), where $s(t)$ is the staffing level, $m(t)$ is the offered load in (2.6) and z is the MSHT limit for the stationary peakedness in (2.12). (The MOL approximation directly justifies replacing $z(t)$ in (2.11) by z in (2.12).)

For the $M/GI/s/0$ model, $c_a^2 = 1$, so that the peakedness is $z = 1$ and thus plays no role. Table 1 of [21] gives peakedness values as a function of c_a^2 and several common service-time distributions. The peakedness in (2.12) captures a complex interaction between the arrival process and the service-time cdf. In particular, the impact of variability in the service-time distribution upon the steady-state number in system depends on the sign of $c_a^2 - 1$; see [21, 35] and references therein for more discussion.

In addition to the MSHT limit for TVIS models in Theorem 2.3, significant theoretical justification for approximation (2.15) with (2.12) comes from an early MSHT limit for the $GI/M/s/0$ model in Theorem 15 (2) on p. 226 of [5], as reviewed Section 4 of [35].

THEOREM 2.4: (MSHT FCLT for the $G/M/s/0$ model from [5]): Consider a sequence of $G/M/s/0$ models indexed by n, with $\mu_n \equiv \mu \equiv 1$, $\alpha_n = \lambda_n \equiv n\lambda$, arrival processes $A_n(t) \equiv N(\Lambda_n(t))$, $t \geq 0$, for a rate-1 process N satisfying (2.4), and $(s_n - \alpha_n)/\sqrt{n} \to \beta$ as $n \to \infty$. Then the blocking approximation in (2.15) is asymptotically correct as $n \to \infty$

PROOF: The proof for GI arrivals is given in [5]. It extends to G under assumption (2.4) by applying Section 7 of [55]. □

2.6. Stabilizing Versus Approximating

In (2.16), we developed an approximation for the TV blocking in a general $G_t/GI/s_t/0$ TVMS loss model, but we are only going to apply it to choose a TV staffing function $s(t)$ to stabilize the blocking probability. The approximation could be used more generally, but it should be used with caution, because it is less reliable more generally.

Experience indicates that it is far easier to develop staffing functions that stabilize performance at desired targets with time-varying arrivals than it is to approximate the performance with arbitrary staffing functions (even at fixed staffing levels) under which the system may alternate among critically loaded (quality-and-efficiency-driven or QED), overloaded (ED), and underloaded regimes (QD) regimes. Experience indicates that approximation (2.16) should perform well within a single MSHT regime.

In support of approximation (2.16) for stabilizing, we now state a supporting MSHT limit for the $M_t/M/s_t + M$ delay model with customer abandonment with scaling that forces the system to remain in the QED MSHT regime. For this purpose, we apply Puhalskii [61]. Our formulation here follows and corrects Theorem 12.1 in the EC of [12], which draws on [44], but see Remark 1 on p. 132 of [61]. The following result is for delay models, but it applies approximately to loss models if we let the abandonment rate θ be very large.

Hence, consider a sequence of $M_t/M/s_t+M$ delay models indexed by n with fixed service rate $\mu = 1$ and abandonment rate θ, $0 \le \theta < \infty$. Let the arrival rate functions in model n be $\lambda_n(t) \equiv n\lambda(t)$ for a fixed arrival-rate function $\lambda(t)$ as above, with $\lambda(t) = 0$ for all $t < 0$. We write $g(t) = o(t)$ if $g(t)/t \to 0$ as $t \to \infty$. Suppose that the staffing functions satisfies

$$s_n(t) = nm(t) + \sqrt{n}c(t) + o(\sqrt{n}) \quad \text{as} \quad n \to \infty, \quad (2.17)$$

where $m(t)$ is the offered load in (2.6) and $c(t)$ is an integrable function for all t, which we think of as a staffing control function. As in [61], when the staffing decreases with all servers busy, let the customers be moved to the end of the queue and let them receive a new full service when they are next assigned. Let $Q_n(t)$ be the number of customers in model n at time t.

THEOREM 2.5: (QED MSHT FCLT supporting stabilization in the $M_t/M/s_t+M$ delay model from [12, 61]): For the sequence of $M_t/M/s_t + M$ delay models specified above, if $\bar{Q}_n(0) \Rightarrow q(0)$ in $\mathbb{R}$ as $n \to \infty$, where $q(0)$ is deterministic, then

$$\bar{Q}_n(t) \equiv n^{-1}Q_n(t) \Rightarrow q(t) \quad \mathit{in} \quad \mathcal{D} \quad \mathit{as} \quad n \to \infty, \quad (2.18)$$

where $q(t)$ satisfies the ordinary differential equation $\dot{q}(t) = \lambda(t) - q(t)$, so that $q(t) = m(t)$, the OL in the $M_t/M/\infty$ IS model.

If, in addition, $\hat{Q}_n(0) \Rightarrow \hat{Q}(0)$ in $\mathbb{R}$ as $n \to \infty$, then

$$\hat{Q}_n(t) \equiv n^{-1/2}[Q_n(t) - nq(t)] \Rightarrow \hat{Q}(t) \quad \mathit{in} \quad \mathcal{D} \quad \mathit{as} \quad n \to \infty, \quad (2.19)$$

where $\hat{Q}(t)$ is a diffusion process satisfying

$$\hat{Q}(t) = \hat{Q}(0) - \int_0^t (\hat{Q}(s) \wedge c(s))ds - \theta \int_0^t (\hat{Q}(s) - c(s))^+ + \int_0^t \sqrt{\lambda(s) + q(s)}dB(s) \quad (2.20)$$

with B being standard Brownian motion. As a consequence, if $\lambda(t)$ is Lipschitz continuous and $c(t) > 0$ for all t, then

$$P(Q_n(t) \ge s_n(t)) = P(\hat{Q}_n(t) \ge c(t) + o(1)) \to P(\hat{Q}(t) \ge c(t)) > 0 \quad \text{as} \quad n \to \infty \quad (2.21)$$

for all $t > 0$. Hence, the staffing in (2.17) puts the system asymptotically in the QED MSHT regime for each $t > 0$.

PROOF: This is a simplification of Theorems 1 and 2 of [61]. In particular, in the setting there we have: $\gamma_s = \beta_s = 0$, $q_s = \kappa_s = m(s)$, $\alpha_s = \lambda(s)$ and $\delta_s = c(s)$ for all s. Theorem 1 of [61] implies that the limit in (2.18) holds with limit $q(t)$, where $q(t)$ satisfies the ordinary differential equation $\dot{q}(t) = \lambda(t) - q(t)$. However, Corollary 4 of [11] implies that the OL $m(t)$ also satisfies the same ODE. Hence, $q(t) = m(t)$, $t \ge 0$. The second limit in (2.19) follows from Theorem 2 of [61]. The Lipschitz continuity of $\lambda(t)$ ensures that the one-dimensional distribution of the diffusion process $\hat{Q}(t)$ has a continuous cdf for each t, which is required for the limit to hold for all $c(t)$ in (2.21); see Theorem 3.2.1 of [65]. □

REMARK 2.1: (the scaling): The conventional MSHT scaling uses centering at the staffing level, but instead we center at the fluid limit, which shows that the staffing control (provided by the function $c(t)$) operates on the diffusion scale. In particular, the MSHT FWLLN in (2.18) is independent of the staffing control $c(t)$, and the control appears in the limiting diffusion process in (2.19). The scaling here follows Theorem 12.1 of [12] and [61]. (The error in [12] evidently occurred because this convention was forgotten when writing down the limiting diffusion process.) We can stabilize the delay probability in (2.21) asymptotically as $n \to \infty$ at any desired target $\alpha > 0$ if we can find a function $c(t)$ so that $P(\hat{Q}(t) \ge c(t)) = \alpha$ for all t. A main contribution of [12] was to show that an iterative staffing algorithm (ISA) could be used with simulation to find the desired function $c(t)$ for the pre-limit.

Fortunately, the story simplifies in the setting of the $M_t/M/\infty$ IS model, which is related to the effectiveness of the MOL approximation.

COROLLARY 2.1: (the special case of the IS model; Corollary 5.1 of the EC to [12]): If $\theta = \mu = 1$, then $Q_n(t)$

has the same Poisson distribution for each t as in the associated $M_t/M/\infty$ model while the limit in (2.20) becomes the much more tractable linear stochastic differential equation

$$\hat{Q}(t) = \hat{Q}(0) - \int_0^t \hat{Q}(s)ds + \int_0^t \sqrt{\lambda(s) + q(s)}dB(s), \tag{2.22}$$

where again $q(t) = m(t)$, which satisfies the ordinary differential equation $\dot{m}(t) = \lambda(t) - m(t)$. The limit $\hat{Q}(t)$ in (2.22) has a Gaussian distribution.

PROOF: The Poisson distribution claim is discussed in Section 6 of [12]. The representation (2.22) follows from (2.20). The ODE characterization of $m(t)$ comes from Corollary 4 of [11], just as in Theorem 2.5. The Gaussian property is well known for linear stochastic differential equations. □

In contrast to the stabilization considered in this article, even for the $M_t/M/s_t+M$ model, general approximation can be difficult, but excellent approximations for those models have been obtained in [46] by exploiting closure approximations starting from the exact functional forward differential (ffd) equations for the Markov model. Similar approximations have been obtained for networks in [60]. These refined approximations can also be used for stabilization, but their greatest advantage seems to be for more general approximations when the system alternates between the different MSHT regimes.

These ffd approximations are evidently limited to Markov models, but Markov models can be very general if we consider time-varying phase-type (Ph_t) or Markov additive processes (MAP_t), as in [33, 52, 66]. However, these formulations introduce complex parameters that are harder to fit to data, and harder to apply to gain insight into the impact of the model on performance, as we discuss next. Nevertheless, these approaches have the advantage that they are not limited by the composition structure in Section 2.1 and the one-parameter characterization of variability exploited here.

2.7. Empirically Evaluating the Square-Root-Staffing Formula

The asymptotic staffing condition in (2.17) suggests that it should be good to use the SRS formula in (2.8). Indeed, Fig. 3 in the EC of [12] validated the SRS staffing in the $M_t/M/s_t/\infty$ delay model without customer abandonment by plotting the implied *empirical QoS*

$$\overline{\beta}^*(t) \equiv \frac{s(t) - m(t)}{\sqrt{m(t)}}, \quad 0 \le t \le T, \tag{2.23}$$

for a periodic arrival-rate function with period T. They showed that the empirical QoS in (2.23) is a constant function for each target across a wide range, after an initial transient, while Fig. 12 in the EC of [12] validated the SRS staffing in the $M_t/M/s_t + M$ delay model with customer abandonment by showing that the empirical QoS was again a constant function after an initial transient. In both these cases, it still remained to find the appropriate QoS parameter β^*, which was done by the MOL method. But the empirical QoS demonstrated the the SRS formula is appropriate. It only remained to specify the single QoS parameter.

We have investigated the empirical QoS for the associated loss model. (Let $s(t)$ be the computed staffing function before any randomization is applied.) Unlike our previous experience with those delay models, we find that the empirical QoS has more fluctuations than before. Figure 3 shows that the MOL staffing using (2.16) is roughly consistent with the SRS formula in (2.8), but has significant periodic fluctuations, more than shown for the $M_t/M/s_t + M$ models in Figs. 3 and 12 of the EC to [12].

2.8. Analyzing, Fitting, and Testing the Model

When we increase the generality of a model, there is a danger that it will become too unwieldy to be of practical value for designing and managing an actual system, such as a call center or a hospital emergency department. We clearly need to be able to solve the more general model, which itself is a big challenge, but we also need to fit the model to data, and perform statistical tests to ensure that the fitted model is reasonable.

For the TVMS $M_t/M/s/0$ loss model, we need to estimate (i) the arrival-rate function $\lambda(t)$ and (ii)) the mean service time $E[S] = \mu^{-1}$. With our approach, when we generalize to the $G_t/GI/s/0$ model, we require in addition only estimating one more function, the service-time cdf G, and one more parameter, the arrival process asymptotic variability parameter c_a^2.

For the service times, we need to verify the assumption that the service times come from a sequence of independent and identically distributed (i.i.d.) random variables, which partly can be based on engineering judgement (does the i.i.d. assumption make sense intuitively?) and partly by standard statistical tests. (Even though we do not consider dependence among the service times, it too can be analyzed in TVMS models; e.g., see [35].) We can estimate the service-time distribution by constructing the empirical cumulative distribution function (ecdf) of the service times, $\hat{G}_n(x) \equiv \sum_{k=1}^n 1_{\{S_k \le x\}}$ (where 1_A is the indicator function of the set A), and then applying it to conduct statistical tests, for example, as in [29] and the references therein.

The time-varying arrival process presents much harder statistical fitting and testing challenges. Even if we have operational data from a long-time period, we need to have some

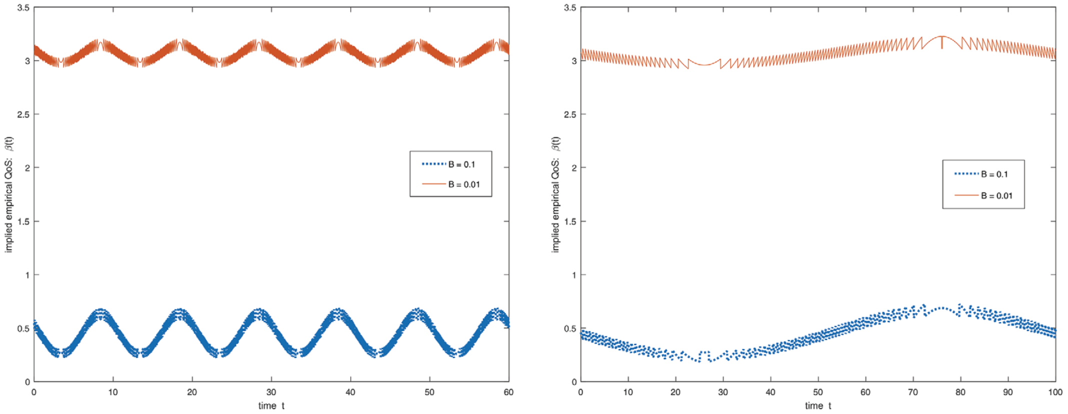

Figure 3. The implied empirical quality of service $\overline{\beta}^*(t)$ in (2.23) of the staffing function in the nonstationary $(H_2)_t/M/s_t/0$ model with the staffing algorithm for parameter pairs $(T, B) = (10, B)$ (left) and $(100, B)$ (right) and blocking targets $B = 0.01$ (top) and $B = 0.1$ (bottom) in each case. [Color figure can be viewed at wileyonlinelibrary.com]

structure. In practice, the required structure in the arrival rate often occurs naturally because the arrival rate often can be regarded as periodic with a daily or weekly cycle, as in the data analysis of an emergency department in Section 3.2.1 of [1], Section 6.3 of [16], and Section 3 of [72]. With such periodic structure, we can estimate the periodic arrival-rate function by averaging over many days.

Ways to test for the NHPP property are discussed in [30, 31] and references therein. The asymptotic variability parameter can be estimated by looking at the index of dispersion for counts (IDC), which is a normalized variance-time curve. In particular, if $A(t)$ counts the number of arrivals in the interval $[0, t]$, then the IDC is the function

$$I_c(t) \equiv \frac{Var(A(t))}{E[A(t)]}, \quad t \geq 0. \tag{2.24}$$

If $A(t)$ is an NHPP, then $I_c(t) = 1$ for all t.

Even for time-varying arrival processes, under regularity conditions such as (2.2), we can obtain the asymptotic variability parameter c_a^2 from estimates of the IDC over a suitably long time interval; that is,

$$c_a^2 = \lim_{t \to \infty} I_c(t). \tag{2.25}$$

For an NHPP, we have $c_a^2 = 1$; for more (less) variable arrival processes, we have $c_a^2 > (<)1$.

The IDC has been used to evaluate arrival processes in TVMS queues that are departure processes from other TVMS queues in Section 4 of [40]. The bottom-left plots in Figs. 6 and 7 of [40] show IDC estimates supporting an NHPP arrival process with $c_a^2 = 1$, whereas the bottom-right plots show IDC estimates supporting a G_t arrival process with $c_a^2 \approx 3.5$. (This deviation from the NHPP property is partly caused by the previous TVMS queue having H_2 service times with $c_s^2 = 4$.) Thus, we see that network structure as with re-entrant customers in [74] is likely to induce non-Markov arrival processes.

Additional discussion of ways to estimate the asymptotic variability parameter c_a^2 are contained in [21]. Ways to calculate the asymptotic variability parameter c_a^2 are discussed in Section 2 and Section 5 there; see (2.3), (2.4), (2.8), and (2.9) for connections to the central limit theorem.

2.9. Two Forms of Blocking: Call Congestion and Time Congestion

It is important to recognize that there are two natural forms of blocking: There is the blocking, B_C, experienced by arriving customers (call congestion) and there is the proportion of time all servers are busy, B_T (time congestion). By the Poisson Arrivals See Time Averages (PASTA) property, these two forms of blocking coincide for the stationary $M/GI/s/0$ loss model. However, they do not coincide for the stationary $G/GI/s/0$ loss model with non-Poisson arrivals. Moreover, the difference can be substantial, as discussed in [35] and references therein.

The delay probabilities seen at arrival and at an arbitrary time are different in stationary $G/GI/s$ multi-server delay models, but in [42] it is shown that these two delay probabilities do not differ much for large-scale $G/GI/s$ multi-server delay models. From [35], it is evident that the story changes for loss models, where the two probabilities can be very different.

When the arrival process is nonstationary, the situation is more complicated. In multiple replications, we estimate the call congestion at time t by counting (i) the number of arrivals in a small interval about t and (ii) the number of these that are blocked; we estimate the blocking (call congestion) by the ratio. In contrast, we can estimate the time congestion at time t by the proportion of all replications that find all servers busy at time t. For the most part in this article, we focus on the call congestion, using the notation $B \equiv B_C$.

Consider the $M_t/GI/s/0$ model with an NHPP arrival process, where we assume it has a smooth arrival-rate function. If we consider a small-time interval about any time t, the arrival-rate function can be regarded as approximately constant there, which is tantamount to having a homogeneous Poisson arrival process in that short time interval. Consequently, we should have B_C approximately the same as B_T at each time point t for the $M_t/GI/s/0$ model, and that is what we find.

Conversely, for the more general $G_t/GI/s/0$, the model behaves locally about any time t as a $G/GI/s/0$ model, so that we should expect B_C to differ from B_T, and that is what we find as well. We discuss the time congestion further in Section 7 and in the appendix [73].

3. THE SIMULATION EXPERIMENTS

We first describe the experimental setting considered and the staffing algorithm. Then we describe our simulation algorithm.

3.1. The Experimental Setting

As in [36] and most earlier work on queues with a time-varying arrival-rate function, we use the sinusoidal arrival rate function in (1.1) with average arrival rate $\bar{\lambda}$, amplitude β, and cycle length (or period) $T = 2\pi/\gamma$ (or equivalently, frequency $\gamma = 2\pi/T$). However, the algorithm is not limited to sinusoidal arrival-rate functions. For model parameters, throughout we assume that the average arrival rate is $\bar{\lambda} = 100$ and the mean service time is $\mu^{-1} = E[S] = 1$. For the stationary cases, $\beta = 0$, while for the nonstationary cases, we let $\beta = 25$. These choices are fixed.

We consider two cycle lengths, $T = 100$ (long) and $T = 10$ (short), so that $\gamma = 0.0628$ and $\gamma = 0.628$. We consider two blocking probability targets, $B = 0.1$ (a higher target, providing a lower QoS) and $B = 0.01$ (a lower target, providing a higher QoS). (We elaborate in Section 5, where we consider even higher targets.) It remains to specify the G_t arrival process beyond its time-varying rate $\lambda(t)$ and the GI service times beyond its mean $E[S] = \mu^{-1} = 1$.

In this article, we let the underlying process N in (2.3) and (2.4) be a renewal process, using the H_2 (hyperexponential of order 2, that is, a mixture of two exponentials) distribution, to represent higher non-Poisson variability. The H_2 renewal process can also be regarded as an interrupted Poisson process, which is a special case of a Markov modulated Poisson process; see [21] and references therein.

We characterize the variability of renewal processes by the variability of their inter-renewal times, and we characterize the variability of all non-negative random variables by their scv, using c_a^2 for the interarrival time scv and c_s^2 for the service-time scv. (For a renewal process, the asymptotic variability parameter in Theorem 2.1 coincides with the scv of an interarrival time, c_a^2, so the common notation here is justified. That would not be the case for nonrenewal processes; e.g., see [13].) All H_2 distributions have $c^2 \geq 1$. For reference, the exponential distribution (M) has scv $c^2 = 1$, while the deterministic distribution (D) has $c^2 = 0.0$. The H_2 distribution has probability density function (pdf) $f(t) = p_1 e^{-t/m_1} + p_2 e^{-t/m_2}$, where $p_1 + p_2 = 1$, so that there are three parameters. For any specified mean, we fix the scv at $c^2 = 4.0$ and we fix the third parameter by assuming balanced means ($m_1 p_1 = m_2 p_2 = 0.5$). The two component exponentials have means 4.437 and 0.563.

In summary, the arrival processes will be either M_t with $c_a^2 = 1$ or $(H_2)_t$ with $c_a^2 = 4$, while the GI service times will be one of D, M, H_2 or LN, which are characterized by their scv's, 0, 1, 4 or 4, respectively. Given this scheme, it suffices to specify the form of the G_t/GI stochastic variability and the parameter pair (T, B).

3.2. A Key Staffing Assumption: Server Switching

There are two different cases for the capacity units: They may be exchangeable (can be regarded as commodities) or may not be. Hotel rooms, hospital rooms, and bicycles for sharing usually can be regarded as exchangeable, whereas human servers cannot. Given that we can and do use dynamic staffing, with human servers we usually need to develop a shift schedule for each individual server.

With human servers, we allow server switching: When that server is scheduled to depart, we do not require that the customer in service stay in service with the same server until their service is complete (called the exhaustive service discipline [25]). Instead, we allow the service in progress to be handed off to another available server. Moreover, we do not force a customer out of service if the staffing is scheduled to decrease when all are busy. Instead, we release the first server that becomes free after the time of scheduled staffing decrease. These practical staffing issues are implemented in the simulation.

To appreciate why there is a significant difference, consider a call center with 100 busy servers, each having an exponential service time with mean 1 (time being measured in mean service times). The mean time until the first server becomes free is 0.01, because the minimum of n i.i.d. mean-1 exponential random variables has mean $1/n$, whereas the mean time until any given server becomes free is 1.0, which is 100 times longer.

3.3. The Staffing Algorithm

We obtain our deterministic staffing function by applying (2.16), which exploits (2.15), which in turn exploits (2.6) and (2.12) and assumes the system starts empty in the distant past, so that we have periodic steady-state formulas. In particular, for each t, given $m(t)$ calculated from (2.6) and z calculated from (2.12), we let $s(t)$ be the smallest possible value such that $B(t) \leq B$. This search is not difficult, because $B(s, \alpha, z))$ is monotone in s. Figure 1 shows the result of the algorithm.

We then apply the randomization algorithm from Section 3.1 of [36]. In particular, given the staffing function $s(t)$ specified above, we construct 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

$$s(t) = s_i, \quad t_{i-1} \leq t < t_i, \quad 1 \leq i \leq n. \qquad (3.1)$$

We randomize by adding a small random time shift to each of the scheduled staffing change times, using a sequence of i.i.d. Gaussian random variables $\{\epsilon_i, i \geq 1\}$ with mean 0 and variance σ^2. First, the sequence of scheduled staffing changes $\{t_i\}$ is replaced with a random sequence $\{\tilde{t}_i\}$, where

$$\tilde{t}_0 = 0 \quad \text{and} \quad \tilde{t}_i = t_i + \epsilon_i, \quad \text{for all} \quad i \geq 1. \qquad (3.2)$$

Second, we force the sequence of of randomized staffing change times $\{\tilde{t}_i\}$ to be nondecreasing and be contained in the chosen time interval by truncating $\tilde{t}_i$, that is, by replacing $\tilde{t}_i$ by $(\tilde{t}_i \vee \tilde{t}_{i-1}) \wedge t_{i+1}$ for each i successively, $1 \leq i \leq n-1$, where $a \vee b \equiv \max\{a, b\}$ and $a \wedge b \equiv \min\{a, b\}$. As a consequence, we have

$$\tilde{t}_{i-1} \leq \tilde{t}_i \leq t_{i+1} \quad \text{for all} \quad 1 \leq i \leq n-1, \qquad (3.3)$$

so that the sequence $\{\tilde{t}_i\}$ is nondecreasing. (The parameter σ should be chosen small enough so 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 (which creates a batch of arrivals at the same time). In simulations, we estimate the blocking probability by performing many independent replications. For the randomization parameter σ and the averaging parameter Δ, we primarily use the values derived in [36]: $\sigma = 0.08$ and $\Delta = 0.20$. (Fig. 1 shows how to interpret these values. They are large compared to interarrival times, but short compared to the cycle lengths.)

In the alternative 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 $\Delta \geq 0$, we look at the time-average blocking probability in the time interval $[t - \frac{\Delta}{2}, t + \frac{\Delta}{2}]$. This averaging coincides with the way blocking probabilities are measured from system or simulation data.

3.4. The Simulation Algorithm

In the simulation experiments, we first calculated the staffing levels (as illustrated by Fig. 1) and recorded the times of staffing changes and the corresponding staffing level in two matrices. And we used our randomization methods describe above to generate the new staffing change times. For each simulation run, we start the system empty at time 0, so that there is an initial warmup period before the system reaches steady-state, which we discuss in Section 6. (For the plots, we do not eliminate an initial part of each run to avoid the warmup period, which tends to be shorter for many-server loss models; see [64]. For the most part, the warmup period is easy to interpret in the plots, but it lasts longer as the service-time variability increases, as we explain in Section 6. We could elect to staff to stabilize during the warmup period too, as in Section 6 and Section EC.2 of [38], but we do not do that here, because that is not our main concern. Moreover, we do not eliminate the warmup period before doing our statistical analysis, because we are not doing any averaging over time.

The first step in generating the arrival process is to generate an H_2 renewal arrival process with rate 1, which we do by generating i.i.d. H_2 interarrival times. We then transformed time to convert this into a sinusoidal arrival rate, with our arrival rate function, using Algorithm 1 in [43]. That algorithm exploits the inverse Λ^{-1} of the cumulative arrival rate function Λ in (2.1). For the sinusoidal arrival rate function in (1.1), the cumulative arrival rate function is

$$\Lambda(t) = \bar{\lambda} t + (\beta/\gamma)(1 - \cos(\gamma t)), \quad t \geq 0. \qquad (3.4)$$

Explicit formulas for the offered load $m(t)$ are given in [10].

To estimate the blocking ($B \equiv B_C$), we recorded the total number of arrivals that were blocked in an interval of length 0.01 around each sampling time, which was taken to be 0.01. The final blocking probability at each sampling time was then calculated by dividing the total number of arrivals that were blocked in that subinterval in all replications by the total number of arrivals in that subinterval over all replications. The number of replications was 10,000 for $B = 0.1$, while it was 100,000 for $B = 0.01$. Overall, this is a challenging experiment because, for $B = 0.01$ and $T = 100$, the expected total number of arrivals over a single sinusoidal cycle in each experiment was $\bar{\lambda} \times T \times n = 10^2 \times 10^2 \times 10^5 = 10^9$. (We used a single cycle for $T = 100$, but multiple cycles for $T = 10$. The experiment required more than a full day of computer time on a 2.7 GHz personal computer.

4. THE RESULTS OF THE EXPERIMENTS

Our base case is the $(H_2)_t/M/s_t/0$ loss model with an H_2 renewal process having $c_a^2 = 4$ (and balanced means) serving as the base counting process N in (2.3), mean-1 exponential

Table 1. Simulation estimates of the blocking probability over four unit intervals each containing one staffing change, for the $(H_2)_t/M/s_t/0$ model with $\mu = 1$, $\bar{\lambda} = 100$, $\beta = 25$, and parameter pair $(T, B) = (100, 0.1)$ using the MOL staffing and randomization (left) and averaging (right)

Estimated call congestion over intervals of length 1								
Staffing change			Randomization: $\sigma = 0.08$			Averaging: $\Delta = 0.2$		
Time	From	To	Min.	Average	Max.	Min.	Average	Max.
39.7	120	119	0.084	0.098	0.118	0.084	0.091	0.102
60.3	92	91	0.078	0.094	0.109	0.084	0.091	0.101
90.2	89	90	0.076	0.094	0.109	0.079	0.087	0.094
99.9	102	107	0.085	0.098	0.110	0.081	0.089	0.096

The minimum, average and maximum values over a unit interval are shown.

service times and the sinusoidal arrival-rate function in (1.1) with $\bar{\lambda} = 100$, $\beta = 25$, and long cycles, with the parameter pairs $(T, B) = (100, 0.1)$ and $(100, 0.01)$. Afterward, we consider the more difficult cases of short cycles, having $T = 10$ (Section 4.2) and then nonexponential service-time distributions, first with long cycles (Section 4.3) and then short cycles (Section 4.4). We extend our study to consider low-variability arrival processes, using Erlang E_4 renewal processes for N in (2.3), in Section 4.5. We discuss heuristic refinements for the difficult case $(T, B) = (10, 0.01)$ in Section 4.6.

Our $(H_2)_t/M/s_t/0$ base case can be compared to previous results for the M_t arrival process in [36]. It also can be compared to previous results for non-Poisson arrival processes, because the same H_2 renewal arrival process is used for N in (2.3) in their corresponding (i) stationary $H_2/M/s/0$ loss model in the second H_2I/MI case of Table 3 in [35] and (ii) in the time-varying $(H_2)_t/GI/s_t$ delay models, with and without customer abandonment, throughout [21].

4.1. The Base Case: Exponential (*M*) Service and Long Cycles ($T = 100$)

From the second H_2I/MI case of Table 3 in [35] with offered load $\alpha = 100$ and blocking targets $B = 0.1$ and 0.01, we see that the staffing approach is good, over-staffing by only $1 - 2$ servers. (That means that the blocking probability approximations are slightly low.) From the tables in [36], we see that MOL staffing with the randomization and averaging works well in stabilizing the blocking probabilities in the $M_t/M/s_t/0$ model.

We now look at similar tables for hyperexponential arrivals, which leads to the MOL staffing in (2.15) with $z = 2.50$ instead of $z = 1.00$; see Table 1 of [21]. Table 1 shows the performance of the two averaging approaches for the base model with parameter pair $(T, B) = (100, 0.1)$, randomization parameter $\sigma = 0.08$, and averaging parameter $\Delta = 0.20$. Given the estimates obtained at all sample points, Table 1 shows the minimum, average, and maximum blocking probabilities over four separate intervals of length 1. (Thus, there are 100 estimates in each subinterval.) Table 1 measures call congestion B_C; corresponding results for the time congestion appear in the appendix.

The good performance of the cases with exponential service times can also be seen graphically. Figure 4 shows the blocking probabilities in the nonstationary $(H_2)_t/M/s_t/0$ model with parameter pairs $(100, 0.1)$ (left) and $(100, 0.01)$ (right), using randomization with $\sigma = 0.08$. The plot shows the estimate for each of the $100/0.01 = 10,000$ sampling points in the time interval. The right plot shows that even with very low blocking target, the blocking probabilities can still remain stable in long cycles with exponential service time. In Section 5, we show examples with higher blocking probability targets, in particular, for $B = 0.2$, 0.4, and 0.8. Since the system starts empty, there is an initial warmup period, which is decreasing in the blocking target; for example, we see that it is over by roughly time 5 for $B = 0.1$ and time 8 for $B = 0.01$; see 6 for further discussion.

4.1.1. Remaining Fluctuations at the Staffing Change Times

Figure 4 is actually somewhat misleading. It may give the impression that the stabilization works perfectly and that the thickness of the plots, that is, the interval $[0.08, 0.10]$ on the left and $[0.007, 0.010]$ on the right, is due only to statistical error. Recall that the plots include all the observations of all the replications, so we do observe the consequences of both statistical fluctuations (each replication yields a different outcome, due to the randomness) and systematic staffing changes (which are always at fixed times). However, a closer look shows that visible jumps remain at the time of each staffing change, but these fluctuations are far less than without the randomization (or averaging), as illustrated by Fig. 2.

These extra fluctuations are dramatically shown by Fig. 5 which displays portions of Fig. 4 over the subintervals $[21, 31]$ and $[50, 60]$. These two plots look quite different, because as can be seen from Fig. 1 the interval $[21, 31]$ is where the staffing is at its peak and so is relatively flat. There

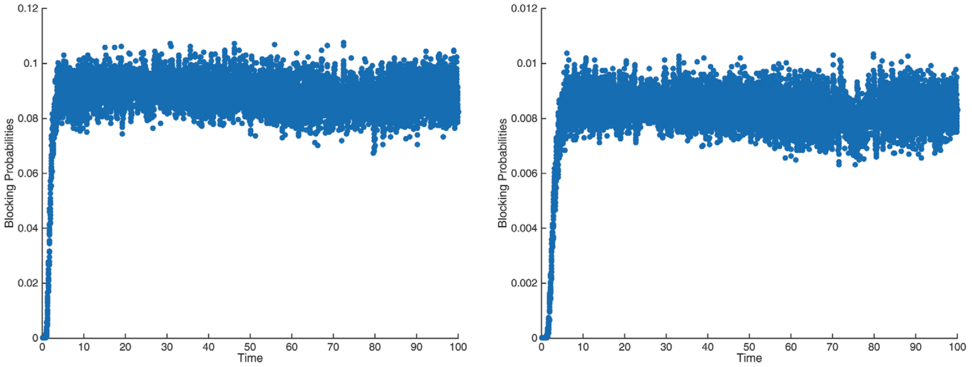

Figure 4. Simulation estimates of the blocking probabilities in the nonstationary $(H_2)_t/M/s_t/0$ model with parameter pairs $(100, 0.1)$ (left) and $(100, 0.01)$ (right) using randomization with $\sigma = 0.08$. [Color figure can be viewed at wileyonlinelibrary.com]

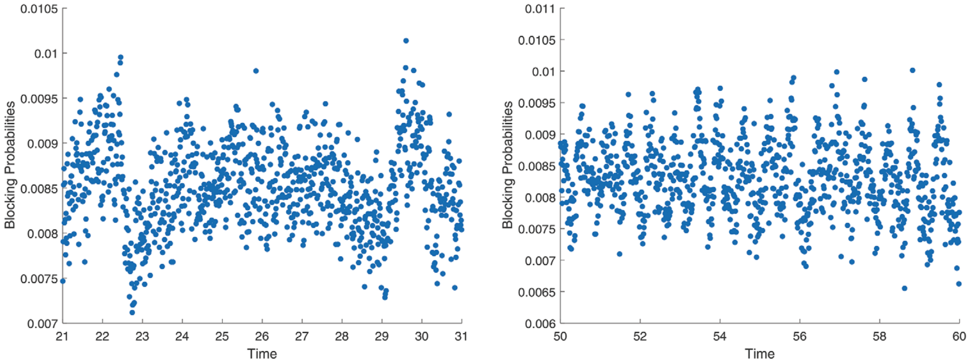

Figure 5. Simulation estimates of the blocking probabilities in the nonstationary $(H_2)_t/M/s_t/0$ model with parameter pair $(T, B) = (100, 0.01)$ having average using randomization with $\sigma = 0.08$ over the subintervals $[21, 31]$ where the staffing is relatively flat (left) and $[50, 60]$ where the staffing is decreasing (right). [Color figure can be viewed at wileyonlinelibrary.com]

are relatively few staffing changes in the interval $[21, 31]$. Indeed, the staffing increases only once from 157 to 158 at time 22.62 and decreases only once at time 29.42. These correspond to low point and high points in the plot.

In contrast, the staffing function is decreasing steadily from 131 to 114 over the interval $[50, 60]$, so that we should expect to see jumps up at the time of each staffing decrease. And that is what we see: the staffing decrease times, including just before and after the interval, are: 49.92, 50.49, 51.07, 51.64, 52.21, 52.79, 53.37, 53.95, 54.53, 55.12, 55.71, 56.31, 56.91, 57.53, 58.15, 58.79, 59.44, 60.11, 60.79. These times coincide with the peaks seen in Fig. 5. This careful analysis shows that we do not succeed in stabilizing the blocking perfectly, but it is far better than without the randomization.

4.1.2. *The Statistical Precision*

In this section, we discuss the statistical precision of our experiments. To a large extent, plots like Figs. 4 and 5 directly show the statistical precision, because estimated values at times not too close should be approximately independent. The plots show the blocking probability estimates for each of the $T/0.01$ sampling points. The plots above show that the systematic fluctuations due to staffing changes tend to dominate the statistical error.

Each estimate is based on 10^5 i.i.d. replications. Figure 4 not only shows that the blocking probabilities are successfully stabilized, but also shows that the statistical precision is about $10-15\%$, ranging from about 0.08 to 0.10 for $B = 0.1$ on the left and from about 0.007 to 0.010 for $B = 0.01$ on the right.

To see that our experimental design should provide good statistical precision, first consider the case of a stationary model with a Poisson arrival process. The sampling interval of 0.01 with $\bar{\lambda} = 100$ means that the number of arrivals in each sampling interval is a Poisson random number with mean 1, but the probability of blocking is about 0.01. Hence, the overall estimate from each replication can be considered to be approximately a Bernoulli random variable with mean 0.01. The total sample size for this sampling time is $n = 10^5$. Hence the sample mean has variance approximately equal to $(0.01)(0.99)/10^5 \approx 10^{-7}$ and standard deviation $\sqrt{10^{-7}} \approx 3.16 \times 10^{-4} = 0.00032$, which makes the relative

Table 2. Confidence intervals of the blocking probabilities in the nonstationary $(H_2)_t/M/s_t/0$ model with parameter air $(T, B) = (100, 0.01)$ with the staffing algorithm using randomization with $\sigma = 0.08$, based on 4 i.i.d. replications of the entire experiment

	$t = 20$	$t = 40$	$t = 60$	$t = 80$	$t = 100$
Replication 1	0.009210	0.008849	0.008134	0.008155	0.008091
Replication 2	0.008650	0.008448	0.008388	0.008533	0.009072
Replication 3	0.008589	0.008761	0.007724	0.008473	0.007360
Replication 4	0.009921	0.008608	0.007542	0.008537	0.007724
Mean	0.009093	0.008667	0.007947	0.008425	0.008062
Standard deviation	0.000619	0.000176	0.000384	0.000182	0.000737
Halfwidth 95% c.i.	±0.000985	±0.000280	±0.000611	±0.000290	±0.001173

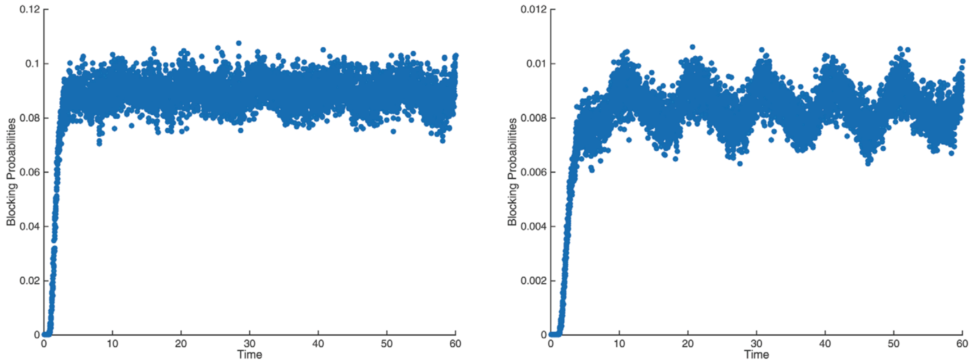

Figure 6. Simulation estimates of the blocking probabilities in the nonstationary $(H_2)_t/M/s_t/0$ model with parameter pairs $(10, 0.1)$) (left) and $(10, 0.01)$) (right) using randomization with $\sigma = 0.08$. [Color figure can be viewed at wileyonlinelibrary.com]

error only $0.00032/0.01 = 0.032$ or 3.2%, which is very good.

However, that analysis does not take account of the randomization. The randomization in the staffing is done before the experiments, so that the blocking estimates in the i.i.d. replications are only *conditionally i.i.d., given the outcome of the randomization*. Hence, to calculate correct confidence bounds in Table 2 below, we did i.i.d. replications of the entire experiment.

Unfortunately, it is very difficult to analytically calculate the exact impact of the randomization or averaging. To get a rough idea of the impact, we create a simple approximate model. Suppose that our estimate is one of three values 0.008, 0.010, or 0.012, with probabilities 1/4, 1/2, and 1/4. Then the variance would be 2×10^{-6}, so that the standard deviation would be $1.4 \times 10^{-3} = 0.0014$, which produces a relative error of 14%, which is consistent with our figures.

To estimate the actual statistical precision, we conducted 4 i.i.d. replications of the entire experiment, and estimated the blocking probability at 5 time points, choosing some near staffing changes and others not. (We use only 4 replications because the run time for each experiment is so long.) We estimate the 95% confidence intervals using the Student t distribution, yielding $\bar{x}_4 \pm 3.182\bar{s}_4/\sqrt{4}$. The results are shown in Table 2. Each case contains simulation estimates of the B_C blocking probabilities in the nonstationary $(H_2)_t/M/s_t/0$ model with parameter pair $(T, B) = (100, 0.01)$ using randomization with $\sigma = 0.08$. In the simulation program, the number of replications is 100,000. We can see that the results yield accuracy at $3 - 15\%$.

4.2. Exponential Service Times and Short Cycles ($T = 10$)

We know from [36] that cases with shorter cycles of length $T = 10$ are more challenging than the cases with long cycles of length $T = 100$. Now the arrival rate changes 10 times more quickly.

Figure 6 shows the simulation estimates of the blocking probabilities in the nonstationary $(H_2)_t/M/s_t/0$ model with parameter pairs $(10, 0.1)$ (left) and $(10, 0.01)$ (right) using randomization with $\sigma = 0.08$. These two plots show that the blocking probabilities are stabilized well when the target is $B = 0.1$, but less well for $B = 0.01$. But even with target $B = 0.01$, the range of possible values clusters quite tightly about the target; the range is about $[0.007, 0.010]$. The fluctuations are far less than without the randomization, as shown by Fig. 2 (left). The stabilization seems that it should be adequate for engineering applications, but there clearly is room for improvement. We show that more stable blocking can be obtained from heuristic refinements in Section 4.6.

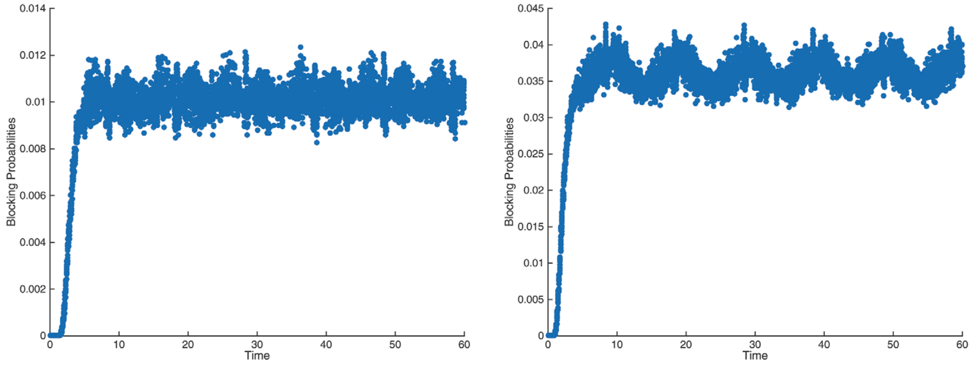

Figure 7. Simulation estimates of the blocking probabilities in the nonstationary $M_t/M/s_t/0$ model with the challenging parameter pair $(10, 0.01)$ using randomization with $\sigma = 0.08$ (left) and in the $(H_2)_t/M/s_t/0$ model with the staffing based on M_t (right). [Color figure can be viewed at wileyonlinelibrary.com]

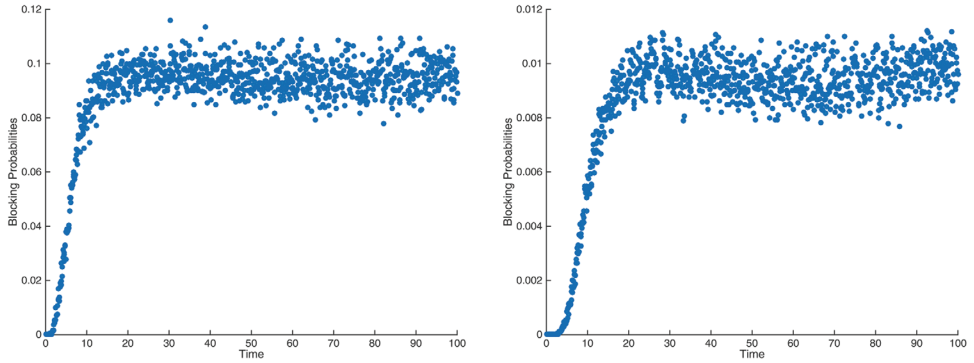

Figure 8. Simulation estimates of the blocking probabilities in the nonstationary $(H_2)_t/H_2/s_t/0$ model with parameter pairs $(100, 0.1)$ (left) and $(100, 0.01)$ (right) using randomization with $\sigma = 0.08$. [Color figure can be viewed at wileyonlinelibrary.com]

There is a fundamental difference between the fluctuations in Figs. 5 and 6. In Fig. 6, we see periodicity with period T rather than jumps at individual staffing change times. Thus, we conclude that the staffing algorithm is less effective for $(H_2)_t$ arrivals than for M_t arrivals when we have both a short period ($T = 10$) and a low blocking target ($B = 0.01$).

To put Fig. 6 in perspective, Fig. 7 makes connections to the corresponding $M_t/M/s_t/0$ model. First, the plot on the left shows the blocking probabilities for the same case with the arrival process changed to M_t. Because there are negligible fluctuations, we deduce that the degradation in stabilization we see for $T = 10$ and $B = 0.01$ is due to the $(H_2)_t$ arrival process. It remains an open problem to explain the reason. Of course, we know that the H_2 renewal process (and thus also the $(H_2)_t$ process) tends to have more short and more long interarrival times than a Poisson process.

The plot on the right show what happens if we apply that algorithm for M_t to the $(H_2)_t/M/s_t/0$ model in Fig. 6. We again see the fluctuations, but now about an inappropriately high blocking level. (Consistent with Fig. 1, the staffing is too low on the right.) Interestingly, the M_t algorithm still stabilizes the blocking probabilities for $(H_2)_t$ at this higher level, although again imperfectly (at the wrong level).

4.3. Nonexponential Service Times and Long Cycles ($T = 100$)

For the rest of our experiments, we consider nonexponential service-time cdf's. In particular, we examine cases with hyperexponential, lognormal and deterministic service times. We will see that, just as in the corresponding cases with exponential service times, the blocking probabilities can be stabilized through our randomization methods. We start with the easier case of long cycles. In this section, we will see that initial warmup period changes with the service-time cdf, tending to get longer as the service time gets more variable; see Section 6.

4.3.1. Hyperexponential Service Times

First, we look at the $(H_2)_t/H_2/s_t/0$ model. Figure 8 shows the simulation estimates of the blocking probabilities in the

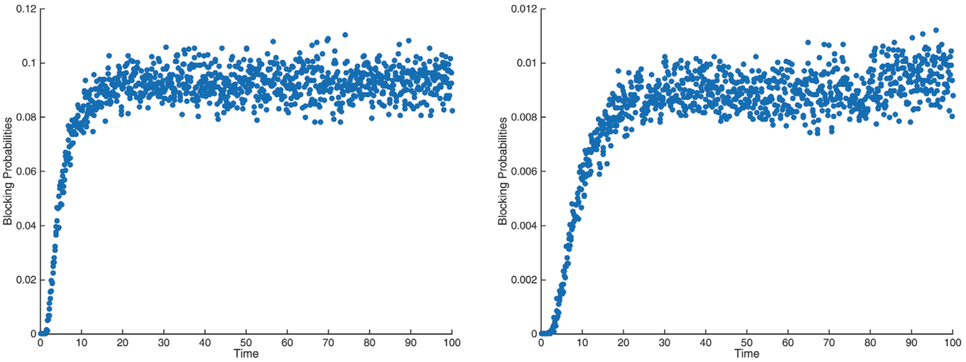

Figure 9. Simulation estimates of the blocking probabilities in the $(H_2)_t/LN/s_t/0$ model with parameter pairs $(100, 0.1)$ (left) and $(100, 0.01)$ (right) using randomization with $\sigma = 0.08$. [Color figure can be viewed at wileyonlinelibrary.com]

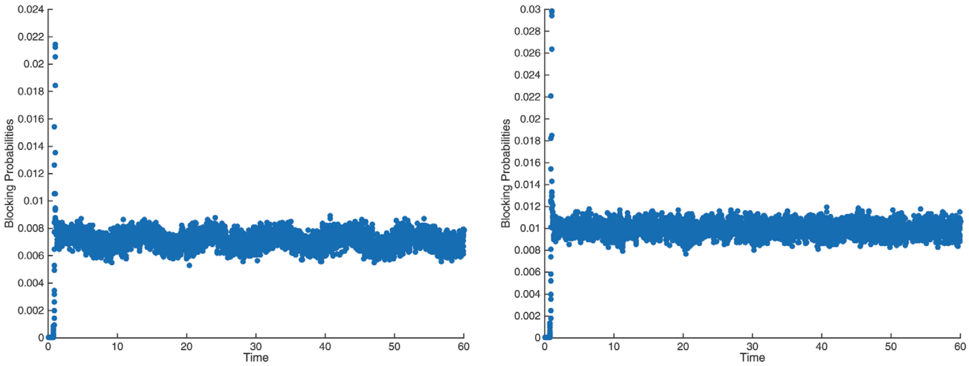

Figure 10. Simulation estimates of the blocking probabilities in the nonstationary $(H_2)_t/D/s_t/0$ model with parameter pair $(T, B) = (10, 0.01)$ using randomization with $\sigma = 0.08$ (left) and with 3 fewer servers (right). [Color figure can be viewed at wileyonlinelibrary.com]

nonstationary $(H_2)_t/H2/s_t/0$ model with parameter pairs $(100, 0.1)$ (left) and $(100, 0.01)$ (right) using randomization with $\sigma = 0.08$. We see that in both plots, the blocking probabilities are again well stabilized (after a longer warmup period than with M service).

4.3.2. *Lognormal Service Times*

We next consider the $(H_2)_t/LN/s_t/0$ cases. Figure 9 shows the simulation estimates of the B_C blocking probabilities in the nonstationary $(H_2)_t/LN/s_t/0$ model with parameter pairs $(100, 0.1)$ (left) and $(100, 0.01)$ (right) using randomization with $\sigma = 0.08$ Again, we see that the blocking probabilities in both plots are stable.

4.4. Nonexponential Service Times and Short Cycles $(T = 10)$

Next, we look at the harder cases with short cycles of $T = 10$. We consider three service-time distributions: deterministic, hyperexponential and lognormal.

4.4.1. *Deterministic Service Times*

We first consider the cases where the service times are deterministic and of length 1. Figure 10 (left) shows the simulation estimates of the blocking probabilities in the nonstationary $(H_2)_t/D/s_t/0$ model with parameter pair $(T, B) = (10, 0.01)$ using the same randomization with $\sigma = 0.08$. We see that the blocking is somewhat low, but the plot on the right shows that we can reach the target exactly by just reducing the number of servers by three at all times. In both cases, the plots are remarkably stable.

4.4.2. *Hyperexponential Service Times*

Next, we look at the $(H_2)_t/H_2/s_t/0$ model with $T = 10$. Figure 11 shows the simulation estimates of the blocking probabilities in the $(H_2)_t/H_2/s_t/0$ model with parameter pairs $(10, 0.1)$ (left) and $(10, 0.01)$ (right) using randomization with $\sigma = 0.08$. If we look at the right plot for $B = 0.01$, we again see imperfect stabilization. Moreover, explained in Section 6, there is a longer warmup period; steady state is reached at about 20. Conversely, the left plot with $B = 0.1$ is rather well stabilized. The stabilization for short cycles

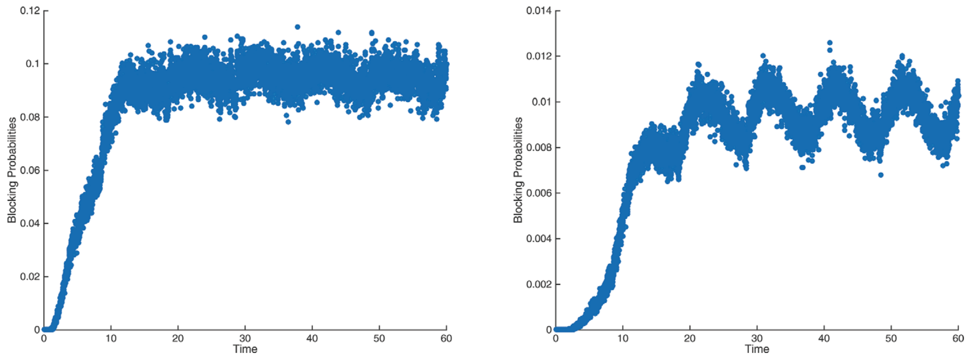

Figure 11. Simulation estimates of the blocking probabilities in the $(H_2)_t/H_2/s_t/0$ model with parameter pairs $(10, 0.1)$ (left) and $(10, 0.01)$ (right) using randomization with $\sigma = 0.08$. [Color figure can be viewed at wileyonlinelibrary.com]

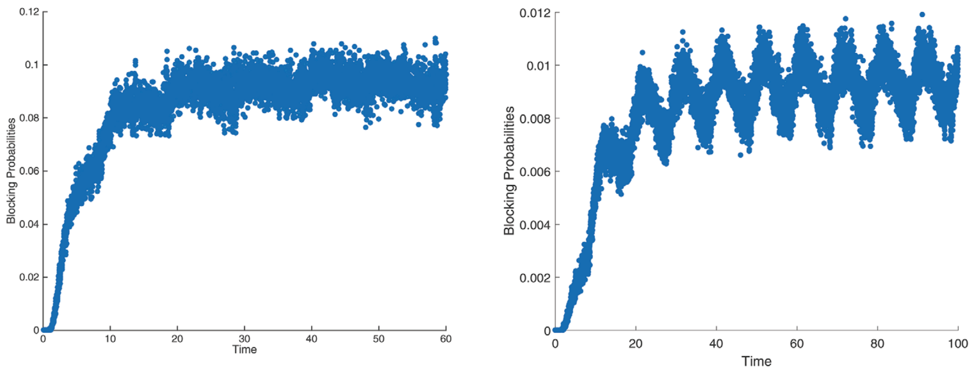

Figure 12. Simulation estimates of the blocking probabilities in the $(H_2)_t/LN/s_t/0$ model with parameter pairs $(10, 0.1)$ (left) and $(10, 0.01)$ (right) using randomization with $\sigma = 0.08$. [Color figure can be viewed at wileyonlinelibrary.com]

and low blocking targets (for the pair $(T, B) = (10, 0.01)$, we again see imperfect stabilization. Nevertheless, the result should be adequate for most engineering applications. It remains to (i) explain why the stabilization degrades in this case and (ii) to find an algorithm that does better.

4.4.3. Lognormal Service Times

Finally, we consider the $(H_2)_t/LN/s_t/0$ cases, where the service times are lognormal with the same mean 1 and scv 4. Figure 12 shows the simulation estimates of the B_C blocking probabilities in the nonstationary $(H_2)_t/LN/s_t/0$ model with parameter pairs $(10, 0.1)$ (left) and $(10, 0.01)$ (right) using randomization with $\sigma = 0.08$. Again, we see that the left plot with $B = 0.1$ is rather well stabilized, while the right plot with $B = 0.01$ shows a longer warmup period and is only imperfectly stabilized. The performance is quite similar to that of the cases with hyperexponential service times. The longer warmup period may perhaps be due to the larger variance: $Var(S_e) = 35.4$ for LN and 13.65 H_2; see (2) of [11]. Because of the longer warmup period, we show both plots over the time interval $[0, 100]$.

4.5. Low-Variability Arrival Processes

We also considered loss models with time-varying arrival processes less variable than Poisson. For that purpose, we considered $(E_4)_t$ arrivals, constructed by letting the base process N in (2.3) be an E_4 renewal process, where the times between renewals have an Erlang E_4 distribution. A mean-1 E_4 random variable can be represented as the sum of 4 i.i.d. exponential random variables each with mean 0.25. An E_4 random variable has scv $c^2 = 0.25$.

Just as for the $(H_2)_t/GI/s_t/0$ model, we conducted simulation experiments for the $(E_4)_t/GI/s_t/0$ model for M and H_2 service in the cases (T, B) with $T = 100$ and 10 and $B = 0.1$ and 0.01. Most of the results appear in the appendix. For long cycles ($T = 100$), the staffing algorithm (again using randomization with $\sigma = 0.08$) remains effective, essentially the same as before.

For short cycles ($T = 10$), the results are essentially the same as before too. Figure 13 shows the blocking probabilities in the nonstationary $(E_4)_t/M/s_t/0$ model with parameter pairs $(10, 0.1)$ (left) and $(10, 0.01)$ (right), using randomization with $\sigma = 0.08$. Just as for $(H_2)_t$ arrivals, there is noticeable periodicity in the case $(T, B) = (10, 0.01)$, but

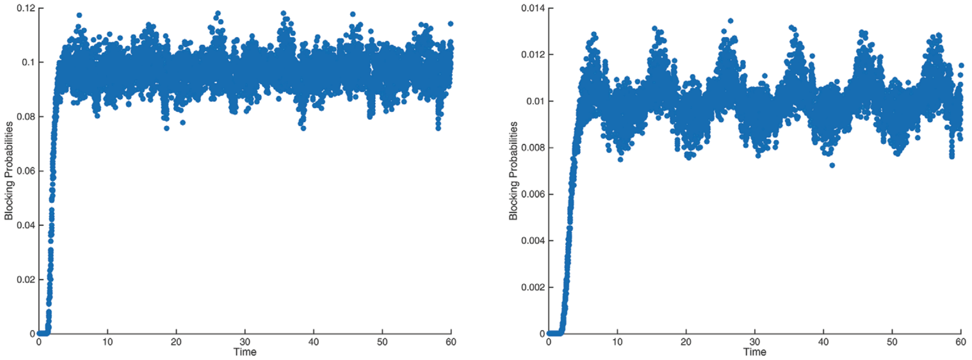

Figure 13. Simulation estimates of the blocking probabilities in the nonstationary $(E_4)_t/M/s_t/0$ model with parameter pairs $(10, 0.1)$ (left) and $(10, 0.01)$ (right) using randomization with $\sigma = 0.08$. [Color figure can be viewed at wileyonlinelibrary.com]

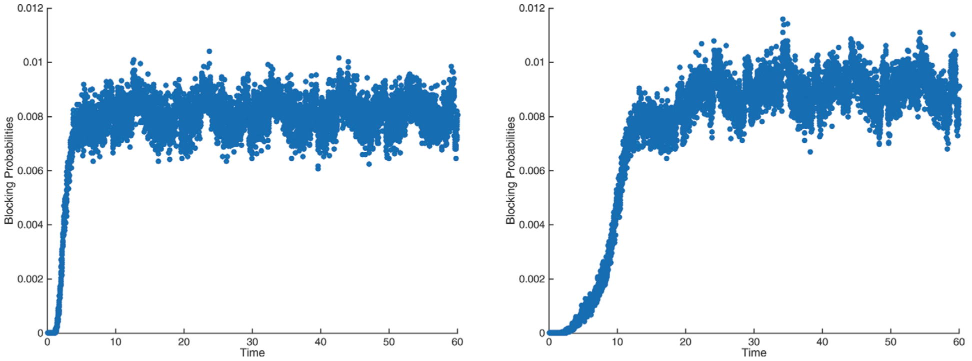

Figure 14. Simulation estimates of the blocking probabilities in the nonstationary $(H_2)_t/GI/s_t/0$ model with M service (left) and H_2 service (right), parameter pair $(T, B) = (10, 0.01)$ with the MOL staffing algorithm increased by 1 during intervals $[9.5, 12.5]$, $[19.5, 22.5]$, etc. (Left) and $[9.5, 14.0]$, $[19.5, 24.0]$, etc. (right), using randomization with $\sigma = 0.08$. [Color figure can be viewed at wileyonlinelibrary.com]

the stabilization seems adequate for engineering purposes. The results for the $(E_4)_t/H_2/s_t/0$ model, where service has been changed from M to H_2, with parameter pairs $(10, 0.1)$ and $(10, 0.01)$ look similar, except that there is again a longer warmup period, consistent with Section 6; see the appendix [73].

4.6. Heuristic Refinements for the Difficult Case $(T, B) = (10, 0.01)$

The right-hand plots in Figs. 6, 11, 12, and 13 show that cyclical fluctuations remain with short cycles $(T = 10)$ and light loading $(B = 0.01)$. That problem is not too serious because the range of values is quite small, but we now show that it is possible to improve the performance by applying simulation to iteratively search for improvements, starting from the staffing solution provided by our algorithm. That is in the spirit of the iterative simulation-based staffing algorithms in [67] and [12].

Starting from our initial staffing, we conducted a local search using additional simulation experiments to find a better staffing function. We considered several adjustments to the initial staffing function (with the steps described in the appendix [73]). Figure 14 shows the results of several iterations. In particular, Fig. 14 shows simulation estimates of the blocking probabilities in the nonstationary $(H_2)_t/GI/s_t/0$ model with M service (left) and H_2 service (right), parameter pair $(T, B) = (10, 0.01)$ with the MOL staffing function increased by 1 during intervals $[9.5, 12.5]$, $[19.5, 22.5]$, and so forth. for M service on the left and on the intervals $[9.5, 14.0]$, $[19.5, 24.0]$, etc. for H_2 service on the right, using randomization with $\sigma = 0.08$, as before. Figure 14 shows clear improvement over Figs. 6 and 11.

5. VERY HIGH BLOCKING PROBABILITY TARGETS

The examples in Section 4 were for the higher blocking probability target $B = 0.1$, which represents a lower QoS, and the lower target $B = 0.01$, which represents a higher QoS. We think that these two cases cover the range of targets of greatest interest in applications, but higher blocking could well occur in some of the applications mentioned

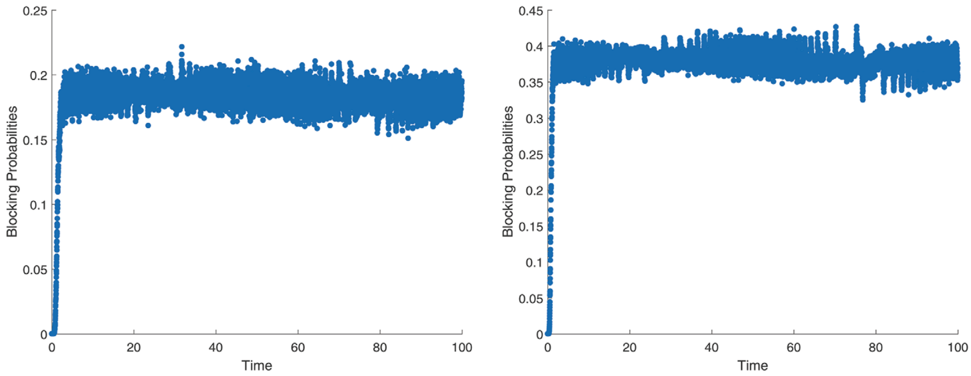

Figure 15. Simulation estimates of the blocking probabilities in the nonstationary $(H_2)_t/M/s_t/0$ model having average arrival rate $\bar{\lambda} = 100$ and relative amplitude $\beta = 25$ with the staffing algorithm for parameter pairs $(T, B) = (100, 0.2)$ (left) and $(T, B) = (100, 0.4)$ (right) using randomization with $\sigma = 0.08$. [Color figure can be viewed at wileyonlinelibrary.com]

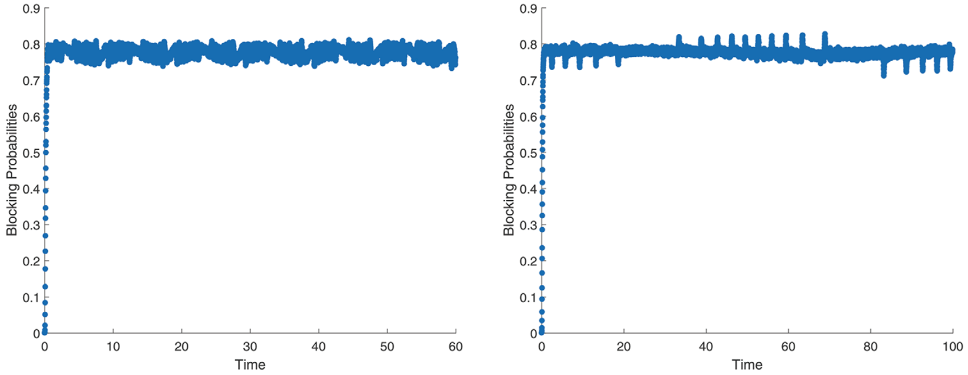

Figure 16. Simulation estimates of the blocking probabilities in the nonstationary $(H_2)_t/M/s_t/0$ model having average arrival rate $\bar{\lambda} = 100$ and relative amplitude $\beta = 25$ with the staffing algorithm for parameter pairs $(T, B) = (10, 0.8)$ (left) and $(T, B) = (100, 0.8)$ (right) using randomization with $\sigma = 0.08$. [Color figure can be viewed at wileyonlinelibrary.com]

in Section 1. Accordingly, we have also investigated the performance of our algorithm for higher blocking probability targets.

Figure 15 is an analog of Fig. 3 for the $(H_2)_t/M/s_t/0$ model with parameter pairs $(100, 0.2)$ (left) and $(100, 0.4)$ (right), while Fig. 16 shows the performance for the same model with both $T = 10$ (left) and $T = 100$ (right) for the very high blocking probability target $B = 0.8$. These figures show that the stabilization staffing algorithm remains effective in these cases with higher blocking probability targets, in fact it is even more effective, although there are visible peaks in these figures.

5.1. Relating Blocking Probabilities to Delay Probabilities

The blocking probability targets 0.1 and 0.01 are much lower than the delay probability targets considered in previous work on delay models, especially with customer abandonment from queue, as [12]. With customer abandonment, the delay probability targets might range from $\alpha = 0.1$ to 0.9, with $\alpha = 0.5$ being a reasonable value. The case $\alpha = 0.5$ often leads to staffing at the offered load, as discussed in paragraph 3 of Section 8 in [12]. (The missing case 3 in Fig. 2 referred to there appears in Fig. 3 on p. 21 of the longer version on the authors' web pages.) This simple case is discussed in Section 3.4 in that longer version as well.

The difference between blocking probabilities and delay probabilities can be understood from the MSHT limits for the stationary models. For that purpose, consider sequences of stationary Erlang $M/M/s/0$ loss and $M/M/s/\infty$ delay models indexed by n with individual service rate $\mu = 1$ and arrival rate n, where we will let $n \to \infty$. As usual, we let $s \to \infty$ along with n, so that

$$(s - n)/\sqrt{n} \to \beta > 0. \tag{5.1}$$

The limit is equivalent to the alternative form $(1-\rho)\sqrt{n} \to \beta$ in [18], where $\rho \equiv n/s$. Let Q_n be the steady-state number of customers in model n.

First, consider the delay model. Let D_n be the steady-state delay probability in model n. The basic MSHT limit for the $M/M/s/\infty$ delay model states that, under condition (5.1),

$$D_n \equiv P(Q_n \geq s) \to \alpha \equiv [1 + \beta\Phi(\beta)/\phi(\beta)]^{-1} > 0 \quad \text{as} \quad n \to \infty; \tag{5.2}$$

see Proposition 1 in [18].

Second, consider the loss model. Let B_n be the steady-state blocking probability in loss model n. The basic MSHT limit for the $M/M/s/0$ loss model states that, under condition (5.1),

$$\sqrt{n}B_n = \sqrt{n}P(Q_n \geq s) \to \gamma \equiv \phi(\beta)/\Phi(\beta) > 0 \quad \text{as} \quad n \to \infty; \tag{5.3}$$

as a consequence of Theorem 15 (2) on p. 226 of [5]. The main point is that D_n converges to a nondegenerate limit as $n \to \infty$ in (5.2), whereas

$$B_n = O(1/\sqrt{n}) \to 0 \quad \text{as} \quad n \to \infty \tag{5.4}$$

in (5.3). Thus, for larger values of n, we expect B_n to be considerably smaller than D_n. Even larger values of D_n are appropriate when we consider customer abandonment.

It is also noteworthy that we can go from one limit to the other, because the steady-state delay probability in the $M/M/s/\infty$ delay model is intimately related to the steady-state blocking probability in the $M/M/s/0$ loss model. In particular,

$$B_n \equiv B(s,n) = \frac{(1-\rho)D(n,s)}{1-\rho D(s,n)} = \frac{(1-(n/s))D(n,s)}{1-(n/s)D(s,n)} \tag{5.5}$$

for $D_n \equiv D(s,n)$ in (5.2); for example, see (2.8) on p. 8 of [68]. Elementary algebra provides a derivation of (5.2) and (5.3) from the other.

6. THE INITIAL TRANSIENT WARMUP PERIOD

In each of our simulation experiments, we started with an empty system. Thus, in each plot we see an initial transient warmup period where the blocking probability rises toward its dynamic periodic steady-state limit. This warmup period differs from model to model, but a good basis for understanding exists in the literature. In particular, the IS model is helpful once again, because convenient explicit formulas exist that expose the structure of the warmup period as a function of the IS model.

In particular, we refer to formula (20) in [11], which shows the offered load $m(t)$ in (2.6) as a function of time in an $M/GI/\infty$ system starting empty at time 0. By Theorem 2.2, this formula also applies to the associated stationary $G/GI/\infty$ model, and can be used as an approximation for many-server $G/GI/s/0$ models.

For the stationary $G/GI/\infty$ model, the mean number of busy servers at time t, starting empty at time 0, is exactly

$$m(t) \equiv E[Q(t)] = m(\infty)G_e(t) \equiv \lambda E[S]P(S_e \leq t), \quad t \geq 0, \tag{6.1}$$

where S is a service time (here assumed to be $E[S] = 1$) and $G_e(t) \equiv P(S_e \leq t)$ is the associated service-time stationary-excess cdf in (2.7). The mean

$$E[S_e] = \frac{E[S^2]}{2E[S]} = \frac{E[S](c_s^2+1)}{2} \tag{6.2}$$

is one form of the *relaxation time*, that is, a measure of the approximate time to reach steady state in the stationary $G/GI/\infty$ model starting out empty at time 0. Of course, the system gradually approaches steady state over time, and does not reach it exactly at any finite time, but this helps us understand the main effects.

The exponential distribution is the unique probability distribution on the positive halfline for which the stationary-excess distribution coincides with the original distribution. As can be seen from (6.2), the mean $E[S_e]$ gets larger as the scv c_s^2 increases. All the service-time distributions considered in this article have mean 1, but the scv's of the D, M, H_2, and LN distributions considered in this article are respectively, 0, 1, 4, and 4. Hence, $E[S_e] = 0.5$, 1.0, 2.5, and 2.5 for the D, M, H_2, and LN distributions. Moreover, the variance of S_e is 35.4 for LN, compared to 13.6 for H_2 and 1 for M.

To elaborate on the transient formula (6.1), we plot the complementary cumulative distribution functions (ccdf's) $\bar{G}_e \equiv P(S_e > x)$ for the M, H_2, and LN distributions in Fig. 17. To better expose the differences, we also plot in log scale on the right.

Figures 4 and 6 show that the warmup period is quite short for the $(H_2)_t/M/s_t/0$ model with M service. As should be expected, it is somewhat longer for the lower blocking probability target $B = 0.01$ than for the higher target $B = 0.1$. In contrast, Figs. 8–9, 11, and 12 show that the warmup period is substantially longer with the more-highly-variable H_2 and LN service times. Consistent with $E[S_e]$, we see that the warmup period is roughly about 2.5 times longer than for M service. As for M service, we see that the warmup period is somewhat longer for the lower blocking probability target.

Now, referring to Fig. 10, we see that the warmup period is shorter for the $(H_2)_t/D/s/0$ stationary model than for M_t, because there is no service-time variability at all. For the deterministic D distribution, S_e is uniformly distributed over the interval [0, 1], so that $E[S_e] = 0.5$. As discussed in [15], for deterministic service times the stationary IS model actually reaches steady-state after one service time. (For all other

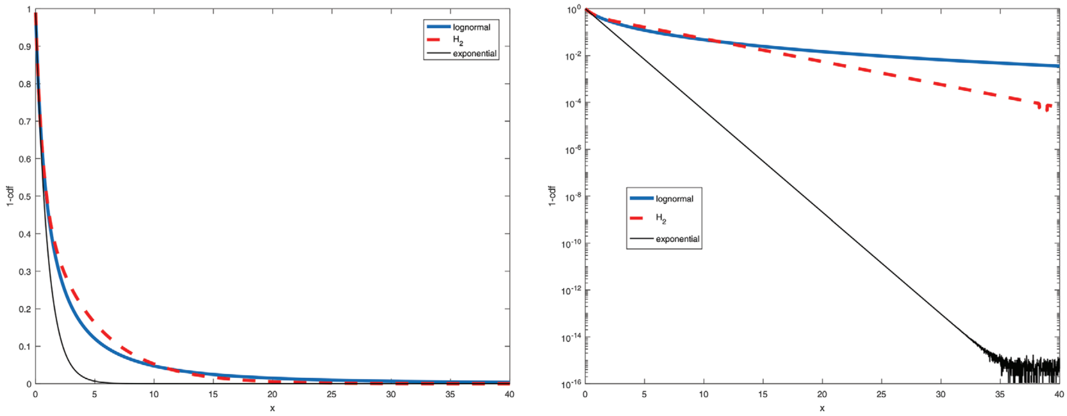

Figure 17. The ccdf $\bar{G}_e(x) \equiv P(S_e > x)$ for the mean-1 $LN(4)$, $H_2(4)$, and $M(1)$ distributions with scv in parentheses. [Color figure can be viewed at wileyonlinelibrary.com]

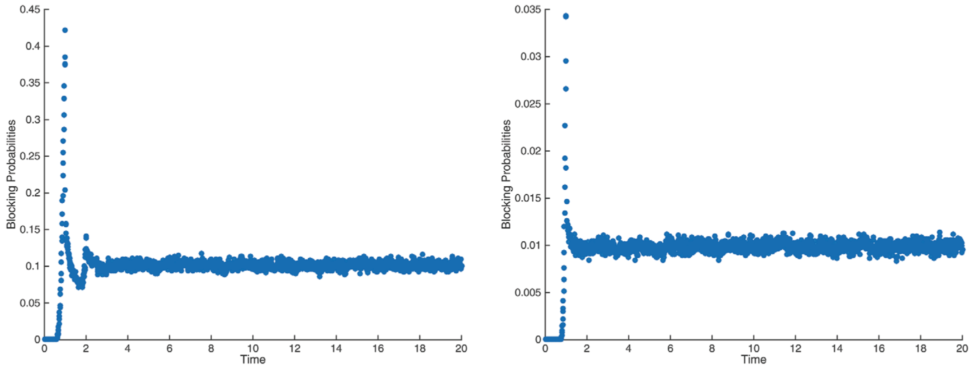

Figure 18. Simulation estimates of the blocking probabilities in the stationary $H_2/D/s/0$ model with parameter triple (100, 0, 10) having constant arrival rate $\bar{\lambda} = 100$ with the staffing algorithm for target $B = 0.1$ (left) and $B = 0.01$ (right) using 3 servers less. [Color figure can be viewed at wileyonlinelibrary.com]

models, steady state is not achieved exactly in finite time.) To help put the time-varying case in perspective, we show the blocking in the stationary $H_2/D/s/0$ stationary model starting empty in Fig. 18.

7. TIME CONGESTION

As indicated in Section 2.9, there are two natural ways to measure the blocking probability: There is the call congestion B_C used here so far, which is the proportion of arrivals that are blocked, and the time congestion B_T, which is the proportion of time that all servers are busy. With Poisson arrivals, these should coincide, but not for non-Poisson arrivals. For the stationary $G/G/s/0$ model, the time congestion B_T is discussed in Section 6 of [35]. The tables in [35] show that: (i) B_C and B_T can be quite different and (ii) approximating B_T can be challenging.

In [35], two approximations were proposed for B_T. The first approximation from Section 6.1 of [35] is $B_T \approx B_C/\max\{z, 1\}$, while the second approximation from (28) in Section 6.2 of [35] is $B_T \approx B_C/\hat{U}_s(1)$, where $\hat{U}_s(x)$ being the Laplace transform of the mean function $E[N(t)]$, where $N(t)$ is the arrival counting process. Theorem 6 of [35] shows that the second method is exact for the $GI/M/s/0$ model, whereas the use of $B_T \approx B_C/\max\{z, 1\}$ is only heuristic, motivated by the numerical results. The implication of the first method is that we approximate B_T by the approximation for B_C with a Poisson arrival process when the actual arrival process is more variable than Poisson. In our case that means that would be acting as if the staffing were the much lower value with an M_t arrival process in Fig. 1 of the main article. Figure 1 of the main article shows that the staffing could be very different. Table 3 of [35] shows that the time congestion is indeed much lower than the call congestion for H_2 arrival processes, roughly consistent with the heuristic approximation. An intuitive explanation is given there as well.

Tables 3 and 4 show the performance of the two averaging approaches for each of the two approximation methods in the

Table 3. Simulation estimates of the time congestion B_T with staffing determined by the first approximation method, letting $z = 1$, over four unit intervals each containing one staffing change, for the $H_2/M/s_t/0$ model with $\mu = 1$ and parameter pair $(T, B) = (100, 0.1)$ $(\gamma = 0.0628)$ using the MOL staffing and randomization (left) and averaging (right)

Estimated time congestion over intervals of length 1 using $z = 1$								
Staffing change			Randomization: $\sigma = 0.08$			Averaging: $\Delta = 0.2$		
Time	From	To	Min.	Average	Max.	Min.	Average	Max.
40.2	111	110	0.084	0.095	0.107	0.091	0.094	0.103
59.7	85	84	0.090	0.098	0.110	0.091	0.097	0.106
89.9	81	82	0.087	0.100	0.112	0.084	0.096	0.103
99.9	94	95	0.087	0.098	0.107	0.084	0.097	0.104

The minimum, average, and maximum values over a unit interval are shown.

Table 4. Simulation estimates of the time congestion B_T with staffing determined by the second approximation method involving $\hat{U}_s(1)$, over four unit intervals each containing one staffing change, for the $H_2/M/s_t/0$ model with $\mu = 1$ and parameter pair $(T, B) = (100, 0.1)$ $(\gamma = 0.0628)$ using the MOL staffing and randomization (left) and averaging (right). The minimum, average and maximum values over a unit interval are shown

Estimated time congestion over intervals of length 1 using $\hat{U}_s(1)$								
Staffing change			Randomization: $\sigma = 0.08$			Averaging: $\Delta = 0.2$		
Time	From	To	Min.	Average	Max.	Min.	Average	Max.
39.9	112	111	0.085	0.093	0.106	0.086	0.091	0.101
60.1	86	85	0.080	0.089	0.102	0.082	0.088	0.098
90.1	83	84	0.076	0.088	0.098	0.079	0.090	0.096
99.5	95	96	0.079	0.090	0.102	0.081	0.089	0.096

base model with parameter $\bar{\lambda} = 100$ and $\beta = 25$, as in the main article, and for the parameter pair $(T, B) = (100, 0.1)$ with randomization parameter $\sigma = 0.08$ and averaging parameter $\Delta = 0.2$. Table 3 shows the approximations based on setting $z = 1$ when $z \geq 1$, while Table 4 shows the approximations involving $\hat{U}_s(1)$. Tables 3 and 4 show that the approximate staffing algorithm for B_T is remarkably effective, just as for the stationary model in [35].

Figure 19 shows the simulation estimates of the B_T blocking probabilities in the nonstationary $(H_2)_t/M/s_t/0$ model with the difficult parameter pair $(10, 0.01)$ with the staffing algorithm of adding $\hat{U}_s(1)$ using randomization with $\sigma = 0.08$. We again see cyclical fluctuations, but the plot looks very similar to the corresponding B_C plot. This suggests that our staffing algorithm to stabilize B_T using $\hat{U}_s(1)$ is effective, performing about as well as our algorithm for B_C.

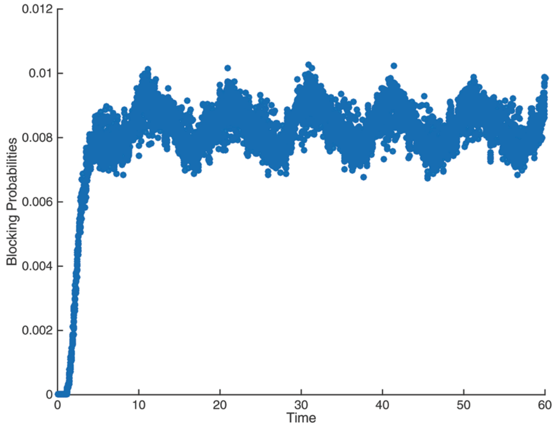

Figure 19. Simulation estimates of the B_T blocking probabilities in the nonstationary $(H_2)_t/M/s_t/0$ model using for the difficult parameter pair $(10, 0.01)$ using randomization with $\sigma = 0.08$. [Color figure can be viewed at wileyonlinelibrary.com]

8. CONCLUSION

In this article, we showed how to apply the MOL method for setting staffing levels to stabilize the blocking probability in a $G_t/GI/s_t/0$ many-server loss model with flexible staffing and non-Poisson time-varying arrivals. Figure 1 shows that the staffing levels can be very different for G_t and M_t arrivals.

In Section 2, we developed the new staffing algorithm and reviewed the extensive literature on which it is based. Five theorems stated there provide key theoretical support for the new algorithm, showing how to meet the significant challenges presented by the more general G_t arrival processes. The staffing algorithm is summarized in Section 3.3. In addition to the general approximation for the time-varying blocking probability in (2.16), an important role is played by an additional randomization about each staffing change time, which follows [36]. Figure 2 shows that the blocking oscillates wildly without this additional randomization. As in [36], we also studied an alternative averaging approach, in which we average the probabilities in a fixed interval of length Δ about each staffing change time, and once again found it to be equally effective, as illustrated by Table 1, but we concentrated on the randomization method in the rest of this article.

We used the composition construction for the arrival process in Section 2.1 to obtain the relatively simple one-parameter characterization of its stochastic variability, which plays a key role in the supporting Theorems 2.1, 2.3, and 2.4, and in fitting the models to data, as discussed in Section 2.8. We also used the composition construction to efficiently generate the arrivals, exploiting [43], which employs table lookup from the inverse of the cumulative arrival rate function in (2.1). Conducting the simulation was challenging because $n = 10^5$ replications were required to reliably estimate small blocking probabilities such as $B = 0.01$ at each time point in a periodic cycle of length 100. for $B = 0.01$, the total number of arrivals over a single sinusoidal cycle in each experiment was $\bar{\lambda} \times T \times n = 10^2 \times 10^2 \times 10^5 = 10^9$.

In Section 4, we also examined the statistical precision. The final statistical precision was about 10–15% relative error, revealed by the width of the plots such as in Fig. 4. However, analysis of the statistical precision is complicated by the remaining small-scale fluctuations about each staffing change time, as discussed in Section 4.1.1. Table 2 gives representative confidence intervals for the estimates at five time points based on four i.i.d. replications of the entire experiment.

In this article, we exposed how the performance of the staffing algorithm depends on the structure of the model. Broadly, there are three key factors: (i) scale, (ii) the QoS, and (iii) the variability. Staffing tends to get more difficult for smaller scale, higher QoS and higher variability.

Here the scale is captured by the average arrival rate $\bar{\lambda}$. We took the scale to be large (but not extraordinarily large), letting $\bar{\lambda} = 100$. We remark that there is evidence that the stabilizing methods do extend to smaller scale; for example, the case $\bar{\lambda} = 20$ was also considered in [36], the case $\bar{\lambda} = 10$ was also considered in [21, 35], and even smaller scale is considered in [8, 74]. The QoS is characterized by the blocking probability target, with lower targets corresponding to higher QoS. We consider both low QoS ($B = 0.1$) and high QoS ($B = 0.01$). We also consider the higher targets $B = 0.2, 0.4$ and 0.8 in Section 5.

Most of our attention was devoted to the impact of the variability, which takes two forms: (i) deterministic and (ii) stochastic. Consistent with intuition, stabilizing performance tends to get more difficult as the variability increases. The deterministic variability refers to the arrival-rate function. For the sinusoidal arrival rate in (1.1), the extent of that variability is determined by the cycle length T and the relative amplitude β; we kept $\beta = 25$ and considered lower variability with $T = 100$ and higher variability with $T = 10$. (This characterization of higher variability breaks down in the (uncommon in practice) limit as T gets extremely small, because as $T \rightarrow 0$, the performance approaches the stationary model with the average arrival rate. That is easy to see because the performance is determined by the cumulative arrival rate function, which then approaches a constant function.)

The stochastic variability refers to the stochastic variability in the G_t arrival process and the GI service-time distribution. These are roughly characterized by the asymptotic variability parameter c_a^2 in (2.2) and the service-time scv c_s^2. Our experience indicates, that as these increase, stabilization gets more difficult. However, our approximation for the blocking probability in (2.16), which draws on (2.6), (2.12), and (2.15) actually depends on $\lambda(t)$, c_a^2 and the full service-time cdf G in a relatively complicated way. Beyond the offered load $m(t)$ in (2.6), the performance depends on the peakedness z in (2.12).

Our simulation experiments showed that our staffing algorithm was consistently effective except when faced with the worst cases of (i) high QoS, as captured by the low target $B = 0.01$, (ii) higher deterministic variability, as captured by the shorter period $T = 10$, and (iii) higher stochastic variability. In particular, stabilization was achieved consistently except in the case $B = 0.01$ and $T = 10$. Even in this case, Fig. 7 (left) shows that there is no degradation for the $M_t/M/s_t/0$ model, while Fig. 10 shows that the blocking is too low for the $(H_2)_t/D/s_t/0$ model, but could be corrected by simply removing 3 servers at all times.

The difficult cases were for the parameter pairs $(T, B) = (10, 0.01)$, and all involve higher stochastic variability. Figure 6 treats the $(H_2)_t/M/s_t/0$ model, while Figs. 11 and 12 treat the $(H_2)_t/GI/s_t/0$ models with H_2 or LN service. In these cases, the stabilization was imperfect, but suitable for most engineering applications, clearly much better than without randomization as in Fig. 2. As explained in Section 6, there is also a long warmup period with H_2 or LN service.

There are many remaining open problems. For stabilizing, it remains to improve the performance in the difficult cases with $(T, B) = (10, 0.01)$. We have shown that it is possible to do so through the heuristic refinement discussed in Section 4.6, but it remains to develop better (more systematic and less complicated) algorithms. It remains to do more on the harder problem of approximating the unstable performance in systems with inflexible staffing, for example, as in the promising direction of [46]. For the theoretical support, it remains to establish versions of Theorems 2.4 and 2.5 to more general models. It also remains to establish more supporting theory for the time congestion, going beyond the case of M service, building on Section 6 of [35], which we have discussed in Section 2.9 and Section 7.

ACKNOWLEDGMENTS

The authors are grateful to Xu Sun for help with Section 2.6 (observing the error in Theorem 12.1 of [12]), to the referees and editors for the sympathetic constructive help, and to the National Science Foundation for financial support from NSF grants CMMI 1265070 and 1634133.

REFERENCES

[1] M. Armony, S. Israelit, A. Mandelbaum, Y. Marmor, Y. Tseytlin, and G. Yom-Tov, Patient flow in hospitals: a data-based queueing-science perspective, Stoch Syst, 5 (2015), 146–194.

[2] S. Asmussen, Applied probability and queues, second edition, Springer, New York, 2003.

[3] A.N. Avramidis, A. Deslauriers, and P. L'Ecuyer, Modeling daily arrivals to a telephone call center, Management Sci 50 (2004), 896–908.

[4] I.S. Borisov and A.A. Borovkov, Asymptotic behavior of the number of free servers for systems with refusals, Theor Probab Appl 25 (1980), 439–453.

[5] A.A. Borovkov, Stochastic processes in queueing theory, Springer, New York, 1976.

[6] E. Brockmeyer, H.L. Halstrom, and A. Jensen, The life and works of A. K. Erlang, Academy of Technical Sciences, Copenhagen, 1948.

[7] L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao, Statistical analysis of a telephone call center: A queueing-science perspective, J Am Stat Assoc (2005) 100, 36–50.

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

[9] M. Defraeye and I. van Nieuwenhuyse, Staffing and scheduling under nonstationary demand for service: A literature review, Omega 58 (2016), 4–25.

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

[11] S.G. Eick, W.A. Massey, and W. Whitt, The physics of the $M_t/G/\infty$ queue, Oper Res 41 (1993), 731–742.

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

[13] K.W. Fendick and W. Whitt, Measurements and approximations to describe the offered traffic and predict the average workload in a single-server queue, Proc IEEE 77 (1989), 171–194.

[14] I. Gerhardt and B.L. Nelson, Transforming renewal processes for simulation of nonstationary arrival processes, Informs J Comput 21 (2009), 630–640.

[15] P.W. Glynn and W. Whitt, A new view of the heavy-traffic limit for infinite-server queues, Adv Appl Probability, 23 (1991), 188–209.

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

[17] N. Grier, W.A. Massey, T. McKoy, and W. Whitt, The time-dependent Erlang loss model with retrials, Telecommun Syst 7 (1997), 253–265.

[18] S. Halfin and W. Whitt, Heavy-traffic limits for queues with many exponential servers, Oper Res 29 (1981), 567–588.

[19] R.C. Hampshire and W.A. Massey, Dynamic optimization with applications to dynamic rate queues, Tutor Oper Res 27 (2010), 208–247.

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

[21] B. He, Y. Liu, and W. Whitt, Staffing a service system with non-Poisson nonstationary arrivals, Probab Eng Inform Sci 30 (2016), 593–621.

[22] S. Henderson, E. O'Mahony, and D.B. Shmoys, (Citi) bike sharing, Cornell University, Ithaca, NY, 2016.

[23] R. Ibrahim, P. L'Ecuyer, N. Regnard, and H. Shen, "On the modeling and forecasting of call center arrivals.," in: Proceedings of the 2012 Winter Simulation Conference, Berlin, Germany, December 9–12, 2012, 256–267.

[24] D. L. Iglehart, Limit diffusion approximations for the many-server queue and the repairman problem, J Appl Probab 2 (1965), 429–441.

[25] I. Ingolfsson, E. Akhmetshina, Y. Li, and X. Wu, A survey and experimental comparison of service-level-approximation methods for nonstationary $M(t)/M/s(t)$ queueing systems with exhaustive discipline, INFORMS J Comput 19 (2007), 201–214.

[26] D.L. Jagerman, Nonstationary blocking in telephone traffic, Bell Syst Tech J 54 (1975), 625–661.

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

[28] G. Jongbloed and G. Koole, Managing uncertainty in call centers using Poisson mixtures, Appl Stoch Models Bus Ind 17 (2001), 307–318.

[29] S. Kim and W. Whitt, The power of alternative Kolmogorov-Smirnov tests based on transformations of the data, ACM Trans Model Comput Simul 25 (2015), 1–22.

[30] S.-H. Kim and W. Whitt, Are call center and hospital arrivals well modeled by nonhomogeneous Poisson processesÏ¿, Manuf Serv Oper Manage 16 (2014), 464–480.

[31] S.-H. Kim and W. Whitt, Choosing arrival process models for service systems: Tests of a nonhomogeneous Poisson process, Nav Rese Logist 61 (2014), 66–90.

[32] Y.M. Ko and J. Pender, Diffusion limmits for the $(map_t/ph_t/\infty)^N$ queueing network, Cornell University, Ithaca, NY, 2016.

[33] Y.M. Ko and J. Pender, Strong approximations for time-varying infinite-server queues with non-renewal arrival and service processes, Cornell University, Ithaca, NY, 2016.

[34] R. Levi and A. Radovanovic, Provably near-optimal LP-based policies for revenue managment in systems with reusable resources, Oper Res 58 (2010), 503–507.

[35] A. Li and W. Whitt, Approximate blocking probabilities for loss models with independence and distribution assumptions relaxed, Perform Eval 80 (2014), 82–101.

[36] A. Li, W. Whitt, and J. Zhao, Staffing to stabilize blocking in loss models with time-varying arrival rates, Probab Eng Inform Sci 30 (2016), 185–211.

[37] Y. Liu and W. Whitt, A many-server fluid limit for the $G_t/GI/s_t + GI$ queueing model experiencing periods of overloading, Oper Res Letters 40 (2012), 307–312.

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

[39] Y. Liu and W. Whitt, Many-server heavy-traffic limits for queues with time-varying parameters. Ann Appl Probab 24 (2014), 378–421.

[40] Y. Liu and W. Whitt, Stabilizing performance in networks of queues with time-varying arrival rates. Probab Eng Inform Sci, 28 (2014), 419–449.

[41] Y. Liu and W. Whitt, Stabilizing performance in a service system with time-varying arrivals and customer feedback, Eur J Oper Res 256 (2017), 473–486.

[42] Y. Liu, W. Whitt, and Y. Yao, Approximations for heavily-loaded $G/GI/n + GI$ queues, Nav Res Logist 63 (2016), 187–217.

[43] N. Ma and W. Whitt, Efficient simulation of non-Poisson non-stationary point processes to study queueing approximations, Stat Probab Lett, 102 (2016), 202–207.

[44] A. Mandelbaum, W.A. Massey, and M.I. Reiman, Strong approximations for Markovian service networks, Queue Syst 30 (1998), 149–201.

[45] W.A. Massey, The analysis of queues with time-varying rates for telecommunication models, Telecommun Syst 21 (2002), 173–204.

[46] W.A. Massey and J. Pender, Gaussian skewness approximation for dynamic rate multi-server queues with abandonment, Queue Syst 75 (2013), 243–277.

[47] W.A. Massey and W. Whitt, Networks of infinite-server queues with nonstationary Poisson input, Queue Syst 13 (1993), 183–250.

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

[49] W.A. Massey and W. Whitt, Unstable asymptotics for nonstationary queues, Math Oper Res 19 (1994), 267–291.

[50] W.A. Massey and W. Whitt, Peak congestion in multi-server service systems with slowly varying arrival rates, Queue Syst 25 (1997), 157–172.

[51] B.L. Nelson and I. Gerhardt, Modeling and simulating renewal nonstationary arrival processes to facilitate analysis, J Simul 5 (2011), 3–8.

[52] B.L. Nelson and M. Taaffe, The $Ph_t/Ph_t/\infty$ queueing system: Part I the single node, INFORMS J. Comput 16 (2004), 266–274.

[53] G. Nieuwenhuis, Equivalence of functional limit theorems for stationary point processes and their Palm distributions, Probab Relat Fields 81 (1989), 593–608.

[54] C. Palm, Intensity variations in teletraffic (in German), Ericsson Tech 44 (1943), 1–189, English translation by North Holland, Amsterdam, 1988.

[55] G. Pang, R. Talreja, and W. Whitt, Martingale proofs of many-server heavy-traffic limits for Markovian queues, Probab Surv 4 (2007), 193–267.

[56] G. Pang and W. Whitt, Two-parameter heavy-traffic limits for infinite-server queues, Queue Syst 65 (2010), 325–364.

[57] G. Pang and W. Whitt, Two-parameter heavy-traffic limits for infinite-server queues with dependent service times, Queue Syst 73 (2013), 119–146.

[58] J. Pender, Nonstationary loss queues via cumulant moment approximations, Probab Eng Inform Sci 29 (2015), 27–49.

[59] J. Pender and Y.M. Ko, Approximations for queue length distributions of time-varying many-server queues, Cornell University, Ithaca, NY, 2016.

[60] J. Pender and W.A. Massey, Approximating and stabilizing dynamic rate Jackson networks with abandonment, Probab Eng Inform Sci 31 (2017), 1–42.

[61] A. Puhalskii, On the $M_t/M_t/K_t + M_t$ queue in heavy traffic, Math Methods Oper Res (2013) 78, 119–148.

[62] M. Restrepo, S.G. Henderson, and H. Topaloglu, Erlang loss models for the static seployment of ambulances, Health Care Manage Sci 12 (2009), 67–79.

[63] J.A. Schwarz, G. Selinka, and R. Stolletz, Performance analysis of time-varying queueing systems: survey and classification, Omega 63 (2016), 170–189.

[64] R. Srikant and W. Whitt, Simulation run lengths to estimate blocking probabilities, ACM Trans Model Comput Simul 6 (1996), 7–52.

[65] D.W. Stroock and S.R.S. Varadhan, Multidimensional diffusion processes, Springer, New York, 1979.

[66] M.R. Taaffe and K.L. Ong, Approximating nonstationary $Ph(t)/M(t)/s/c$ queueing systems, Ann Oper Res 8 (1987), 103–116.

[67] R.B. Wallace and W. Whitt, A staffing algorithm for call centers with skill-based routing, Manuf Serv Oper Manage 7 (2005), 276–294.

[68] W. Whitt, The erlang B and C formulas: Problems and solutions, Columbia University, Available at: http://www.columbia.edu/~ww2040/allpapers.html, 2002, accessed on May 17, 2017.

[69] W. Whitt, Stochastic-process limits, Springer, New York, 2002.

[70] W. Whitt, What you should know about queueing models to set staffing requirements in service systems, Naval Res Logist 54 (2007), 476–484.

[71] W. Whitt, Offered load analysis for staffing, Manuf Serv Oper Manage 15 (2013), 166–169.

[72] W. Whitt and X. Zhang, A data-driven model of an Emergency Department, Oper Res Health Care 12 (2017), 1–15.

[73] W. Whitt and J. Zhou, Appendix to "Many-server loss models with time-varying arrivals," Columbia University, Available at: http://www.columbia.edu/~ww2040/allpapers.html, accessed on May 17, 2017.

[74] G. Yom-Tov and A. Mandelbaum, Erlang R: A time-varying queue with reentrant customers, in support of healthcare staffing, Manuf Serv Oper Manage 16 (2014), 283–299.