STAFFING A SERVICE SYSTEM WITH NON-POISSON NONSTATIONARY ARRIVALS

Beixiang He, Yunan Liu and Ward Whitt

Department of Industrial and Systems Engineering, North Carolina State University,
Raleigh NC 27695, bhe@ncsu.edu
Department of Industrial and Systems Engineering, North Carolina State University,
Raleigh NC 27695, yliu48@ncsu.edu
Department of Industrial Engineering and Operations Research, Columbia University,
New York, NY, 10027; ww2040@columbia.edu

March 21, 2016

Abstract

Motivated by non-Poisson stochastic variability found in service system arrival data, we extend established service system staffing algorithms using the square-root staffing formula to allow for non-Poisson arrival processes. We develop a general model of the non-Poisson nonstationary arrival process that includes as a special case the nonstationary Cox process (a modification of a Poisson process in which the rate itself is a nonstationary stochastic process), which has been advocated in the literature. We characterize the impact of the non-Poisson stochastic variability upon the staffing through the heavy-traffic limit of the peakedness (ratio of the variance to the mean in an associated stationary infinite-server queueing model), which depends on the arrival process through its central limit theorem behavior. We provide simple formulas to quantify the performance impact of the non-Poisson arrivals upon the staffing decisions, in order to achieve the desired service level. We conduct simulation experiments with non-stationary Markov modulated Poisson arrival processes with sinusoidal arrival rate functions to demonstrate that the staffing algorithm is effective in stabilizing the time-varying probability of delay at designated targets.

Keywords: queues with time-varying arrival rates; nonstationary queues; Cox processes; capacity planning; setting staffing levels; non-Poisson nonstationary point processes

Short Title: Staffing a Service System

Contact Author: Ward Whitt, ww2040@columbia.edu

1 Introduction

From analysis of service system data, e.g., [1, 2, 6, 18, 24, 25], there is consensus that (i) the arrival rate typically varies significantly over the day in almost all service systems and (ii) the service-time distribution typically is not nearly exponential, usually fitting a lognormal distribution far better than an exponential distribution. The situation is less clear for the stochastic properties of the arrival processes.

1.1 Non-Poisson Properties of Arrival Data

Statistical analysis of arrival data from intervals within a single day in call centers and hospital emergency departments, where arrivals primarily occur exogenously based on individual choice, are mostly consistent with the commonly assumed nonhomogeneous Poisson process (NHPP) [6, 24, 25], but analysis of data from multiple days, even restricted to the same hour and the same day of the week, show significant over-dispersion, inconsistent with the Poisson property with a deterministic arrival rate; see §§1.4 and 4-7 of [24]. Similarly, in [23] appointment-generated arrivals to an endocrinology clinic were found to be consistent with an NHPP within each day, but the daily totals over multiple days show significant under-dispersion, as expected because arrivals are controlled via appointment systems.

Indeed, several authors have found significant non-Poisson properties in service system arrival processes; see [4, 20, 22, 57]. In response, it has been suggested that the arrival process ought to be a nonstationary Cox process (doubly stochastic Poisson process), which is a Poisson process where the arrival rate itself is a nonstationary stochastic process; see [4, 5, 20, 57]. Hence, we develop an arrival process model that encompasses those suggestions and develop a staffing algorithm to stabilize performance for that model. Of particular promise for engineering applications, we also show how to apply the algorithm to set staffing levels from arrival data, without creating a complete arrival process model, by estimating the index of dispersion for counts (IDC) of the arrival process, as in [8, 13].

1.2 Two Forms of Scale: Spatial and Temporal

Experience indicates that an effective staffing algorithm should depend on two forms of scale: spatial and temporal; see [18]. By "spatial," we mean size, i.e., the typical number of servers. The size of a queueing system has a significant influence on performance. For example, the typical traffic intensity (server utilization) tends to be significantly greater in a queueing system with many servers, under normal loading; see [49]. In addition, the average waiting time before starting service in a queueing system tends to be significantly less (greater) than the average service time in a queueing system with many (few) servers, under normal loading. In this paper we are primarily concerned with larger sizes. Accordingly, most of our examples have about 100 servers, but we also consider a few examples with $4 - 20$ servers.

By "temporal scale," we mean the relevant time scale. For most service systems, the relevant time scale from the perspective of the performance experienced by customers is the expected response time, the expected time from arrival until completing service. The response time can be complicated if the service delivery is divided into temporarily separated pieces as in healthcare and web chat; then it may be useful to use a more general network model, as in [32, 33, 55]. For the multi-server queueing models considered here, where there is a single uninterrupted service time, the response time is the waiting time plus the service time. For larger systems, *the relevant time scale tends to be of order equal to the mean service time*, because the waiting time tends to be small compared to the mean service time.

As the system scale increases by increasing the number of servers and the arrival rate, but leaving the service-time and patience-time cdf's fixed, individual customer experience of service remains unchanged, but the relevant scale in the arrival process becomes many interarrival times instead of only one, because there typically are many (of order equal to the expected number of busy servers) interarrival times during one mean service time. Thus, in a many-server queue, we should expect the arrival process to influence its performance primarily through its long-time behavior (as viewed through its mean interarrival time), i.e., through its central limit theorem (CLT). This point of view is also advanced by [57]. Our staffing algorithm builds on this asymptotic view.

1.3 The Relevant Time Scale for Staffing: Short and Long Service Times

The relevant time scale (the mean service time) is important for interpreting the variation in the deterministic time-varying arrival-rate function. Even if an arrival-rate function changes dramatically over a day, it can be considered approximately constant at each time t if it changes relatively little over an interval of several mean service times. For example, in some telephone call centers, e.g., as in [6], the average service time may be about 3 minutes. Then, even if the arrival rate function varies significantly over the day, if the arrival rate function does not change too much over each half hour, it may be roughly appropriate to staff by using a pointwise stationary approximation (PSA); i.e., by using a stationary model with the arrival rate prevailing at that time. (See §3 of

[24] for an examination of when it is appropriate to assume a constant rate over a subinterval.)

With this PSA view, at each time we have a steady-state view. With short service times, if the arrival data are consistent within a day, but over-dispersed over many days, then it may suffice to staff according to a mixture of Poisson distributions, as in [22], or even a mixture of deterministic fluid approximations, as in [52], if the uncertainty is large. The key to relatively simple analysis without these special approaches (with short service times) is forecasting that successfully eliminates most of the uncertainty about the rate, as in [4, 20, 44] and references there.

In contrast, in this paper we are primarily interested in the more difficult case of longer service times, where the arrival rate can change significantly over a single service time, so that the PSA view is no longer appropriate; Figure 1 of [21] dramatically shows the performance degradation of PSA with longer service times.

Even in this setting with longer service times, forecasting is important. A direct statistical data analysis of arrival data, analyzing all days, is likely to be highly misleading if it ignores systematic effects like the day of the week. In good practice, the uncertainty is typically addressed by becoming familiar with special features of the system and applying forecasting methods. With proper understanding of the system and forecasting, the model introduced here or even an NHPP may be found to be appropriate.

We emphasize that it is far from automatic that arrival processes in practice will be NHPP except possibly for uncertainty about the rate. For example, it is well known that network structure can directly cause non-Poisson properties in arrival processes. When an arrival process arises as an overflow process or departure process from another system, that structure often induces non-Poisson variability. The early literature on overflow traffic can be traced from [26], which was aimed at creating a relatively simple approximation; see [27, 28] for recent work on loss models.

1.4 An Example: Many-Server Queues in Series

Suppose that arrivals to a high-demand service system must go through two or more stages of service, where each stage is staffed by a large number of servers working in parallel. Suppose that the arrival process to the system is an NHPP, but with significant time variability, and that the service times in each stage can be regarded as i.i.d. random variables. However, as is usually the case, suppose that the service times in the first stage of service are *not* exponentially distributed. A natural model is the $M_t/GI/s_{1,t} \rightarrow \cdot/GI/s_{2,t}$ system, possibly with abandonment of some waiting customers. This model has an NHPP arrival process, time-varying staffing levels that need to be determined (the $s_{j,t}$) and general service-time distributions (the two GI). The first stage can be analyzed with established methods, but analysis of the second stage is complicated by the non-exponential service times at the first stage.

A delayed-infinite-server modified-offered-load (DIS-MOL) staffing algorithm for this system was developed in [32]. That staffing algorithm was found to be remarkably effective, except when the second service stage requires a high quality of service, while the first stage is staffed to meet a low quality of service; see §8.1 for a performance summary of the DIS-MOL algorithm and see §4 for a study of the arrival process to the second stage, i.e., the departure process from the first stage.

There turns out to be a relatively simple explanation for the performance degradation of DIS-MOL in this one case: When the first stage operates with a high quality of service, the departure process from the first stage tends to be approximately an NHPP, but when the first stage operates with a low quality of service, the departure process from the first stage tends not to be approximately an NHPP. Instead, the departure process from the first stage tends to behave like the G_t arrival process model introduced here, with complex stochastic variability generated from the non-exponential service times in the first stage plus the time-varying arrival rate. Given that the service times come from an i.i.d. sequence with a fixed distribution, it seems reasonable to expect that the level of stochastic variability in the departure process should be approximately constant over time.

Indeed, §8.3 of [32] suggested that a fruitful next research step would be what we do in this paper. We here find that a variant of the approach proposed there can indeed be carried out and that it performs well. Thus, we are solving the open problem there. At the same time, we are providing means to solve a larger class of problems.

1.5 Our Contributions

For the $M_t/GI/s_t$ queueing model, which has arrivals according to an NHPP (M_t) with time-varying arrival rate function $\lambda \equiv \lambda(t)$ and independent and identically distributed (i.i.d.) service times with a general (non-exponential) service-time cumulative distribution function (cdf) G, successful approaches to the staffing problem were developed in [21]; see reviews in [10, 18]. Since then, further advances have been made in [9, 12, 28, 30, 32, 46, 55]. For the more general $G_t/GI/s_t$ queueing model, an offered-load (OL) normal approximation was proposed in §§5 and 6 of [21], but that has never been tested. Here are our contributions, and the place they appear in the paper:

(i) In §2 we develop **a general non-Poisson G_t arrival process model that encompasses the nonstationary Cox process**, based on methods of composition. In particular, we represent the arrival counting process as the composition of a stationary counting process and a deterministic cumulative arrival rate function, separately treating the stochastic variability and the deterministic variability of the arrival rate over time.

We propose a parsimonious partial characterization of the component stationary stochastic counting process in terms of the asymptotic variability parameter c_A^2 arising in its central limit theorem. In §2.2 we elaborate on that general model by giving the stationary stochastic counting process the structure of a stationary Cox process and showing how to compute its asymptotic variability parameter. In §5 we indicate how the key asymptotic variability parameter c_A^2 can be computed in more specific stochastic models and estimated from system data without constructing any model by estimating the IDC.

(ii) In §3 we develop **a new staffing algorithm for this $G_t/GI/s_t$ model**, which extends the modified-offered-load (MOL) algorithm for the $M_t/M/s_t$ model developed in §4 of [21] by exploiting the many-server heavy-traffic (MSHT) approximations for the stationary $G/GI/s$ model in [51]. We represent this new MOL algorithm as a square-root-staffing formula. In doing so, we exploit the peakedness (the ratio of the variance to the mean of an associated infinite-server model), as in [27] and references therein. The use of peakedness was also suggested in §6 of [21] as part of a more elementary offered-load (OL) approach to staffing, but that was never tested. Because the MOL algorithm has proven to be superior to the OL algorithm for M_t arrivals, it is evident that the MOL approach here should be preferred.

In §4 we combine the contributions above to provide simple formulas to quantify the performance impact of the non-Poisson arrivals upon the staffing decisions (here the number of servers), in order to achieve the same service level. We estimate how many more (or possibly fewer) servers are needed because the arrival process is G_t instead of M_t with the same arrival rate function; that difference can be significant.

(iii) Next, in §6 we develop **an extension of our staffing algorithm to the $G_t/GI/s_t + GI$ model having customer abandonment** according to a general (non-exponential) patience-time cdf F (the $+GI$), drawing upon [15, 56]. As emphasized in [15], including abandonment in the model is often important in service systems, because it often occurs and significantly affects performance. Moreover, the patience distribution is often non-exponential [6].

(iv) Finally, we have conducted **extensive simulation experiments verifying that the new algorithm is effective and robust**; our numerical experiments cover cases with various performance targets, large and small system sizes, and various arrival processes, service-time and patience-time distributions. In §§7 and §8, we report our simulation results for the $G_t/GI/s_t$ and $G_t/GI/s_t + GI$ models.

We draw conclusions in §9. We present additional simulation results in an appendix.

2 The Non-Poisson Nonstationary Arrival Process Model

Our arrival process model has two key features: (i) a time-varying deterministic arrival-rate function $\lambda \equiv \{\lambda(t) : t \geq 0\}$ and (ii) non-Poisson stochastic variability characterized parsimoniously by the single parameter c_A^2. As usual, the arrival-rate function λ characterizes the predictable deterministic variability over time, whereas the parameter c_A^2 characterizes the additional stochastic variability. The reference cases are $c_A^2 = 0$ for a deterministic process, without any stochastic variability at all, and $c_A^2 = 1$ for a Poisson process. Thus a nonhomogeneous Poisson process (NHPP) will be covered as a special case of the general model with $c_A^2 = 1$.

2.1 A General Model Based on Composition

We construct the various stochastic processes considered here exploiting composition, as in §7 of [37] and [16, 29, 34, 38, 53]. Let $A(t)$ count the number of arrivals in the interval $[0, t]$ for $t \geq 0$. We represent our general nonstationary arrival counting process A as the composition of a stochastic counting process N and a deterministic cumulative arrival rate function Λ, using the composition function $\circ$, with $(x \circ y)(t) \equiv x(y(t))$, $t \geq 0$. In particular, we represent our arrival process as

$$A \equiv N \circ \Lambda \quad \text{or, equivalently,} \quad A(t) \equiv N(\Lambda(t)), \quad t \geq 0, \tag{2.1}$$

where N is a stochastic counting process with nondecreasing nonnegative integer-valued sample paths, while the deterministic function Λ is the cumulative arrival rate function satisfying

$$\Lambda(t) = \int_0^t \lambda(s)\, ds, \quad t \geq 0, \tag{2.2}$$

with $0 < \lambda_{LB} \leq \lambda(t) \leq \lambda_{UB} < \infty$ for positive numbers λ_{LB} and λ_{UB}. As a consequence, Λ is continuous and strictly increasing, so that it has a well defined continuous and strictly increasing inverse Λ^{-1}. Given Λ and a general nonstationary arrival process A, the counting process N could be recovered by letting $N = A \circ \Lambda^{-1}$.

Since 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 make the asymptotic assumption that $N(t)/t \rightarrow 1$ with probability 1 (w.p.1). Thus, N could be a renewal process with mean interarrival time 1 as well as its stationary (or equilibrium) version, as in §V.3 of [3]. Our key stochastic assumption is that N obeys a central limit theorem (CLT):

$$t^{-1/2}[N(t) - t] \Rightarrow N(0, c_N^2) \quad \text{as} \quad t \rightarrow \infty, \tag{2.3}$$

where $\Rightarrow$ denotes convergence in distribution and $N(m, \sigma^2)$ denotes a random variable with the normal (Gaussian) distribution having mean m and variance σ^2. As an immediate consequence, we obtain an associated CLT for A,

$$\Lambda(t)^{-1/2}[A(t) - \Lambda(t)] \Rightarrow N(0, c_N^2) \quad \text{as} \quad t \rightarrow \infty, \tag{2.4}$$

so that $c_A^2 \equiv c_N^2$ is the asymptotic variability parameter of A, based on its CLT. (We remark that, at some places in this section, the technical development is facilitated by applying functional limit theorems as in [50], but for simplicity and brevity we omit that.)

For general stationary point processes, the asymptotic variability parameter c_A^2 can be characterized and estimated from data via its representation as the limit of the the index of dispersion for counts $I \equiv \{I(t) : t > 0\}$, i.e.,

$$c_A^2 = \lim_{t\rightarrow\infty} I_A(t) = \lim_{t\rightarrow\infty} I_N(t) = c_N^2, \quad \text{where} \quad I_A(t) \equiv \frac{Var(A(t))}{E[A(t)]} = I_N(\Lambda(t)). \tag{2.5}$$

see [8, 13, 32, 45]. We are assuming that $I(t)$ is well defined and finite, and that a finite limit in (2.5) exists. For an NHPP, $I(t) = 1$ for all t.

Remark 2.1 (Many-Server Heavy-Traffic Limits for Queues) Strong theoretical support for characterizing the arrival process by its CLT behavior is provided by many-server heavy-traffic (MSHT) limit theorems, because established MSHT limits depend on the arrival process only through its CLT.

In particular, established MSHT limits for the stationary Markovian $M/M/\infty$, $M/M/s$, $M/M/s/r$ and $M/M/s + M$ models extend to the associated $G/M/\infty$, $G/M/s$, $G/M/s/r$ and $G/M/s + M$ models, where the G arrival process can be N in §2.1, as reviewed in §7.3 of [40], which affects the limit only through the parameter c_A^2. Moreover, the same is true for nonstationary arrival process in MSHT limits established for the $G_t/G/\infty$ infinite-server (IS) model in [41, 43] and the $G_t/M/s_t + GI$ model in [31].

Remark 2.2 (The Composition Construction Is Restrictive) Even though the composition construction in (2.1) is useful and quite general, including several natural models as special cases, (2.1) is restrictive. It is a special construction, treating only a subclass of all non-Poisson nonstationary arrival processes. To understand the restriction on the G_t arrival process more generally, it is helpful to consider the special case in which the process N_a is a rate-1 Markov modulated Poisson process (MMPP) with a finite-state continuous-time Markov environment process, yielding an arrival rate of γ_k in state k [14]. The composition construction in (2.1) implies that the arrival rate of A at time t when the environment process is in state k is simply the product $\lambda(t)\gamma_k$. More generally, a nonstationary MMPP with a finite-state Markov environment process could have arrival rate $\gamma_k(t)$, which is a general function of the two variables k and t. Clearly, the construction here yields only a subset of all possible cases, but nevertheless we believe that it usefully goes beyond the M_t model. It allows some characterization of the stochastic variability of the arrival and service processes instead of none at all. It remains to determine how useful is the "one-dimensional" characterization of non-Poisson stochastic variability in the non-M_t G_t arrival process. Since non-M_t properties often arise through structural features such as having arrivals be departures or overflows from another queue, as illustrated by §1.4, there is good reason to expect that the present approach will prove useful. Moreover, the heavy-traffic limit identifies parsimonious characterizations of the stochastic variability in the arrival and service processes, as discussed in Remark 2.1.

2.2 A More Detailed Model Based on Composition

We now explain how our general model encompasses the Cox process (or doubly-stochastic Poisson process) mentioned in §1. For that purpose, we introduce a more detailed model. We now represent the stochastic counting process N as the composition of two other stochastic processes, writing

$$N \equiv M \circ C \quad \text{or, equivalently,} \quad N(t) = M(C(t)), \quad t \geq 0, \tag{2.6}$$

where M is a stochastic counting process with nondecreasing nonnegative integer-valued sample paths and C is a stochastic cumulative process, expressed as

$$C(t) \equiv \int_0^t Z(s)\, ds, \quad t \geq 0, \tag{2.7}$$

with $\{Z(t) : t \geq 0\}$ being a stochastic "rate" process (SRP) with nonnegative sample paths. We assume that the component stochastic processes M and C are mutually independent. Combining representations (2.1) and (2.6) gives a three-fold composition representation for the overall arrival process A: $A = M \circ C \circ \Lambda$.

This representation of N reduces to a stationary Cox process if we assume that M is a Poisson process. The most familiar stationary Cox process is a Markov-modulated Poisson process (MMPP), which arises when the SRP Z is a function of a continuous-time Markov chain (CTMC); see [14]. A further special case of an MMPP is an interrupted Poisson process (IPP), which is an MMPP with a two-state environment process, where the rate of the Poisson process is 0 in one of the two environment states. An IPP is equivalent to a renewal process with hyperexponetial (H_2) intervals between renewals; see [26] and §2.3.1 of [14].

Our key stochastic assumption in this new framework is the validity of CLT's for the two stochastic processes M and C. Given that we want N to asymptotically have rate 1 and C to specify the cumulative rate, We assume that $M(t)/t \Rightarrow 1$ and $C(t)/t \Rightarrow 1$ w.p.1 as $t \rightarrow \infty$. Our key stochastic assumption in this new framework is the validity of CLT's for the two independent stochastic processes M and C.

$$t^{-1/2}[M(t) - t] \Rightarrow N(0, c_M^2) \quad \text{and} \quad t^{-1/2}[C(t) - t] \Rightarrow N(0, c_C^2) \tag{2.8}$$

These together imply a CLT for N and A as in (2.3) and (2.4) with

$$c_A^2 = c_N^2 = c_M^2 + c_C^2, \tag{2.9}$$

as in Example 9.6.2 of [50]. For additional details on the derivation of (2.9), see Theorem 11.4.4 and §13.3 of [50].

3 The New Staffing Algorithm

We consider the general $G_t/GI/s_t$ model, which has unlimited waiting space and i.i.d. service times that are independent of the arrival process specified in §2. We let the service times be distributed as a random variable S with mean $E[S] = \mu^{-1}$ and general cdf G.

Our proposed staffing algorithm for the general $G_t/GI/s_t$ model is designed to stabilize the (virtual) delay probability, i.e., the probability that a potential arrival at time t must wait before starting service, $P(W(t) > 0) = P(Q(t) \geq s(t))$, where $Q(t)$ denotes the number of customers in the system at time t. The algorithm is an extension of the OL approach developed in [21] and reviewed in [18], which leads to the classical *square-root staffing* (SRS) formula.

3.1 The Square Root Staffing Formula

Our SRS formula stipulates that the staffing (number of servers) at time t be

$$s(t) = m(t) + \beta_\alpha \sqrt{m(t)}, \quad \text{with} \quad \beta_\alpha \equiv \beta_\alpha(z) \equiv \beta\sqrt{z} \quad \text{and} \quad \beta \equiv \beta_\alpha(1), \tag{3.1}$$

when the targeted delay probability is α, where $m(t)$ is the OL, i.e., the mean number of busy servers in the associated $G_t/GI/\infty$ infinite-server (IS) model with the same arrival and service processes, $\beta_\alpha(1)$ is the previous quality-of-service (QoS) parameter for the modified-offered-load (MOL) approximation for the $M_t/M/s_t$ (Erlang-C) Markovian model based on the MSHT limit in [19], and z is a one-parameter characterization of all non-Markov variability in the associated stationary $G/GI/\infty$ IS model, with arrival process N, i.e., acting as if $\Lambda(t) = t$. Since the number of servers is necessarily an integer, we round to the next largest integer in all staffing formulas.

3.2 Explicit Formulas

We now specify the key parameters $m(t)$, $\beta_\alpha(1)$ and z explicitly. First,

$$m(t) = \int_{-\infty}^{t} \lambda(s)\bar{G}(t-s)\,ds, \quad t \geq 0, \tag{3.2}$$

where $\lambda(t)$ is the deterministic arrival rate at time t (assumed to start in the indefinite past, but we could have $\lambda(s) = 0$ for $s \leq t_0$) and $\bar{G}(s) \equiv 1 - G(s)$. Second,

$$\beta_\alpha(1) = H^{-1}(\alpha), \tag{3.3}$$

where H^{-1} is the inverse of the strictly increasing continuous function

$$H(\beta) = [1 + \beta\Phi(\beta)/\phi(\beta)]^{-1}, \quad 0 < \beta < \infty, \tag{3.4}$$

and Φ (ϕ) is the cdf (pdf) of a standard (mean 0, variance 1) normal random variable. Third,

$$z \equiv z(c_A^2, G) = 1 + (c_A^2 - 1)\mu \int_0^\infty \bar{G}(x)^2\,dx, \tag{3.5}$$

where μ^{-1} is the mean service time and c_A^2 is an arrival process variability parameter specified in §2.

3.3 Additional Justification

These choices can be further justified. First, as observed in §5 of [21], formula (3.2) is (exactly) the same as for the much more elementary $M_t/GI/\infty$ model, which has a Poisson number of busy servers at each time t. Second, the particular way $\beta_\alpha(1)$ and z are combined in (3.1) draws on the MSHT approximation for the stationary $G/GI/s$ model developed in [51]. The refined MOL staffing formula proposed for the Markovian $M_t/M/s_t$ model in §4 of [21] is (3.1) with $z = 1$. The MSHT limit assumes that $\lambda \to \infty$ and $s \to \infty$ with the SRS in (3.1) holding asymptotically. When there is customer abandonment (discussed in §6), we use the related MSHT limits from [15] and [56].

The parameter z in (3.5) is the heavy-traffic limit (letting the arrival rate grow) of the *peakedness* in the associated stationary $G/GI/\infty$ IS model, with a stationary version of the G_t arrival process (the process N in §2), where the peakedness is the ratio of the variance to the mean of the steady-state number of busy servers. Formula (3.5) is discussed further in [27, 42] and references therein. Consistent with established theory for the $M/GI/\infty$ model, $z = 1$ for all service-time cdf's if $c_A^2 = 1$, which occurs if the arrival process is Poisson. We have $z \geq (\leq) 1$ if and only if $c_A^2 \geq (\leq) 1$.

In §7 we will show that the new staffing algorithm in §3 is effective for the $G_t/GI/s_t$ model and that it provides a significant improvement over the corresponding staffing algorithm from [21], which is obtained by using $z = 1$ in (3.1); e.g., see §7.4.

4 Predicting the Impact of the Non-Markov Features

The SRS formula in (3.1) and the peakedness formula z in (3.5) allow us to predict the staffing implications of the non-Markovian stochastic features in the model (having G_t instead of M_t), assuming that we want to maintain the same QoS: The (approximately constant) difference in the staffing level is simply

$$s_G(t) - s_M(t) = \beta_\alpha(1)(\sqrt{z} - 1)\sqrt{m(t)} \approx \beta_\alpha(1)(\sqrt{z} - 1)\sqrt{s(t)}. \tag{4.1}$$

(As $m(t)$ grows, formula (3.1) implies that $s(t)/m(t) \to 1$.) The QoS parameter $\beta_\alpha(1)$ in (3.1) should usually satisfy $0.5 \leq \beta_\alpha(1) \leq 2.0$; see the Halfin-Whitt (HW) curve in Figure 2 of [18]. If we take $\beta_\alpha(1) = 1$ as a typical reference case, then we see that the non-Markovian structure should lead to changing the number of servers by $(\sqrt{z} - 1)\sqrt{s(t)}$. If $s(t) = 100$, then the change would by $10(\sqrt{z} - 1)$ servers. This usually means additional servers, but it could mean fewer servers, because we could have $0 \leq z < 1$ as well as $z \geq 1$.

An important practical reference case is exponential M service, yielding $z = (c_A^2 + 1)/2$. For this case, we see right away that the approximate performance impact when $\beta_\alpha(1) = 1$ is

$$s_G - s_M = (\sqrt{z} - 1)\sqrt{s_M} = \left(\sqrt{(c_A^2 + 1)/2} - 1\right)\sqrt{s_M} \quad \text{servers.} \tag{4.2}$$

Hence, when $\beta_\alpha(1) = 1$, $c_A^2 = 4$ and $s_M = 100$, we need $10(\sqrt{2.5} - 1)/2 = 5.8$ additional servers compared to the Markovian case. Very roughly, this is about 6% more servers.

Another important reference case for the peakedness z is a deterministic service cdf, yielding $z = c_A^2$. Surprisingly, perhaps, if the service-time cdf were changed from M to D in the numerical example above with $\beta_\alpha(1) = 1$, $c_A^2 = 4$ and $s(t) = 100$, the number of extra servers required to achieve the same QoS would increase from 5.8 servers to 10 servers. Clearly, the impact becomes much greater if c_A^2 is larger. These formulas allow quick back-of-the-envelope calculations.

Given the common case in which $c_A^2 > 1$, z is *decreasing* as the variability of G increases. (As the variability increases for fixed mean, $\mu \int_0^\infty \bar{G}(x)^2 \, dx \to 0$. Think of a two-point distribution with mean 1 having a very small probability p of a very large $1/p$ and otherwise being 0. Understanding this phenomenon is facilitated by the integral representation in (11) of [42]. See [54] for an early discussion of this phenomenon.) For $c_A^2 > 1$, the largest possible value of z occurs with deterministic service times, yielding $z = c_A^2$. Overall, the possible values of z as a function of c_A^2 are

$$z \equiv z(c_A^2, G) \quad \text{in} \quad [c_A^2 \wedge 1, c_A^2 \vee 1], \tag{4.3}$$

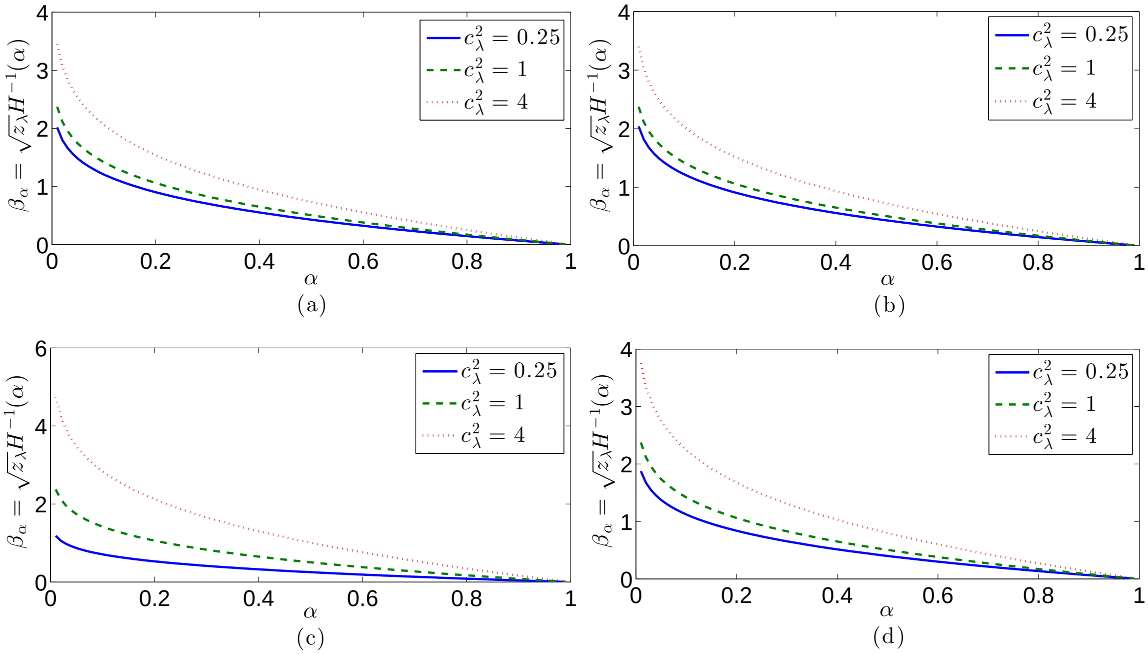
where $a \wedge b \equiv \min\{a, b\}$, $a \vee b \equiv \max\{a, b\}$. Moreover, all possible values of z can be attained (possibly asymptotically). The range of possible z values as a function of c_A^2 increases as $|c_A^2 - 1|$ increases for either $c_A^2 \geq 1$ or $c_A^2 \leq 1$.

Table 1 shows peakedness values $z \equiv z(c_A^2, G)$ as a function of the arrival variability parameter c_A^2 and common service-time cdf's G: lognormal ($LN(\mu^{-1}, c_s^2)$), deterministic, Erlang (of order 2, E_2), hyperexponential ($H_2(\mu^{-1}, c_s^2)$) and exponential. The mean service times μ^{-1} are chosen to be 1, but z is independent of the mean. The second service-time parameter c_s^2 is the squared coefficient of variation (scv, variance divided by the square of the mean). The third parameter of the H_2 distribution is fixed by using balanced means, as on p. 137 of [47]. Only modest levels of variability, as measured by c_A^2 and z, are considered in Table 1.

c_A^2	D	E_2	M	$LN(1, 0.25)$	$LN(1, 1)$	$LN(1, 4)$	$H_2(1, 1.5)$	$H_2(1, 2)$	$H_2(1, 4)$
0.25	0.25	0.53	0.63	0.45	0.58	0.72	0.66	0.69	0.74
0.50	0.50	0.69	0.75	0.63	0.72	0.82	0.78	0.79	0.83
1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2.00	2.00	1.63	1.50	1.74	1.56	1.37	1.45	1.42	1.35
3.00	3.00	2.25	2.00	2.48	2.11	1.74	1.90	1.83	1.70
4.00	4.00	2.88	2.50	3.22	2.67	2.11	2.35	2.25	2.05

Table 1: Values of the peakedness $z \equiv z(c_A^2, G)$ for six different arrival process variability parameters c_A^2 and nine different service distributions.

Analysis of service-time data by [6] and others has shown that service system service-time cdf's often fit the $LN(1, 1)$ lognormal cdf quite well, but simulation experiments show that the performance impact of that distribution is not very different from the commonly assumed exponential distribution. Table 1 is consistent with that, showing that $z(c_A^2, LN(1, 1)) \approx z(c_A^2, M)$. This



(a) (b) (c) (d)

Figure 1: The QoS parameter $\beta_\alpha \equiv \beta_\alpha(z)$ as a function of α, for three different arrival variability parameters $c_\lambda^2 = 0.25, 1, 4$ and four different service-time distributions: (a) $LN(1,4)$, (b) $H_2(1,4)$, (c) deterministic (D) (d) exponential (M).

suggests that assuming exponential service times is unlikely to seriously invalidate performance predictions. However, the non-Poisson arrival process is an important feature. Note that the peakedness z for $LN(1,1)$ is relatively large in Table 1 for $c_A^2 > 1$, e.g., for $c_A^2 = 4$. In particular, note that the peakedness $z \equiv z(c_A^2, LN(1,\sigma^2))$ for $c_A^2 > 1$ and $LN(1,\sigma^2)$ service is *decreasing* in σ^2, so that the relatively small variance seen in estimated lognormal service times does not help when the arrival process is more bursty than Poisson.

To summarize, for service-time cdf's something like exponential (as measured by z), we roughly need $(\sqrt{(c_A^2-1)/2}-1)\sqrt{s(t)}$ additional servers at time t compared to the same model with Poisson arrivals.

To illustrate the consequence of the non-Markov variability on the approximation, we display the QoS parameter $\beta_\alpha \equiv \beta_\alpha(z)$ as a function of α for three arrival process variability parameters c_A^2 $(0.25, 1.00, 4.00)$ and four service-time distributions: (a) $LN(1,4)$, (b) $H_2(1,4)$, (c) deterministic (D), and (d) exponential (M) in Figure 1.

5 Parameter Specification

The processes N and M introduced in §2.1 and §2.2 above are understood to be conventional rate-1 stationary counting processes, so interest centers on the variability parameters, which we already have discussed in general terms. We now elaborate for the more structured model in §2.2.

5.1 Calculating Variability Parameters for Stochastic Models

As indicated in (2.7), the process C in (2.6) is an integral of the SRP Z. In most applications, the SRP is a regenerative process, which makes C a cumulative process as in [17] or §VI.3 of [3]. That commonly occurring structure provides general sufficient conditions for the FCLT for $\mathbf{C}_n$ to hold, but the parameters are expressed in terms of relatively complicated variables associated with the underlying regenerative cycles. However, these usually can be numerically calculated or estimated in simulations. In general, the rate λ_C is just the steady-state mean $E[Z(\infty)]$, assuming that $Z(t) \Rightarrow Z(\infty)$ as $t \to \infty$ and $E[Z(\infty)] < \infty$.

A relatively convenient model for N is an MMPP, because it is not difficult to simulate and analyze. One natural construction is to let M be a rate-1 Poisson process and let $Z(t) = f(\Gamma(t))$, $t \geq 0$, where $\Gamma \equiv \{\Gamma(t) : t \geq 0\}$ is a CTMC taking values in the finite set $\{1, 2, \dots, m\}$. Then $f(i) = \lambda_i$, where λ_i is the deterministic arrival rate that prevails whenever $\Gamma(t) = i$.

With this convention, not only is M a rate-1 Poisson process, but C is a special cumulative process, with successive visits of the underlying CTMC Γ to any fixed state constituting regenerative cycles. In this setting,

$$E[Z(\infty)] = \lim_{t\to\infty} t^{-1} E[C(t)] = 1, \tag{5.1}$$

Cumulative processes associated with functions of DTMC's and CTMC's are discussed, respectively, in §I.7 of [3] and [48] (and many references therein). Formulas and algorithms to compute c_C^2 are given in (12) and Corollary 3 of [48]. More elementary formulas and algorithms for birth-and-death processes are given in (6) (Proposition 1) and Remarks 1-3 of [48].

The key parameters of an MMPP can also be obtained from [14], but it is important to recognize that it is a different representation. They directly represent N and do not separately exploit the rate-1 Poisson process M. Nevertheless, expressions for a general rate λ_N and c_N^2 can be obtained from [14]. We can obtain λ_N from expressions for the mean $E[N(t)]$. In particular, in [14] we see that $\lambda_N = \pi\lambda = \sum_{j=1}^{m} \pi_j \lambda_j$ from the first term on the right of (23). Here π_j is the steady-state probability that the CTMC is in state j and λ_j is the rate of the MMPP when the CTMC is in state

j. Similarly, we can obtain the variability parameter c_N^2 from the related expressions for $E[N(t)^2]$ in (25) and (26) of [14]. We close this section by noting that the MMPP is a special case of the batch Markovian arrival process (also known as the versatile Markovian process or Neuts process), for which asymptotic variability parameters can be found in §5.4 of [39] and [7].

5.2 Estimating the Arrival Process Variability Parameter Directly from Data

Since the arrival process beyond its determinitic rate $\lambda(t)$ affects the staffing algorithm in (3.1) only through the asymptotic variability parameter $c_A^2 = c_N^2$ in the peakedness z in (3.5), in many applications it may be convenient to directly estimate c_A^2 from arrival process data. That can be done using the IDC characterization in (2.5). Since the limits of $I_A(t)$ and $I_N(t)$ as $t \to \infty$ are identical, we can directly work with the nonstationary arrival process A and estimate $I_A(t)$, estimating c_A^2 by the estimated limit of $I_A(t)$ as $t \to \infty$.

Unfortunately, this estimation is not entirely straightforward, tending to require large samples. Large samples present relatively little problem with simulation, but they may not be possible with arrival data. See [13] and §4 of [32] for examples involving single-server and many-server queues, respectively.

Remark 5.1 (*detecting model violations*) Model violations from excessive variability sometimes can be identified from divergence of $I(t)$ as $t \to \infty$. For example, if $N(t) = \Pi(Xt)$, where Π is a unit-rate Poisson process and X is a nonnegative random variable with $E[X] = 1$ and $0 < Var(X) < \infty$, then $E[N(t)] = t$ for all t, but

$$Var(N(t)) = Var(E[N(t)|X]) + E[Var(N(t)|X)] = Var(X)t^2 + t,$$

so that $I(t) = 1 + Var(X)t \to \infty \to \infty$ as $t \to \infty$. Our model requires that both the variance and the mean of $N(t)$ grow linearly in t.

6 Extension to Models with Customer Abandonment

We next extend our MSHT MOL SRS algorithm based on (3.1) to the corresponding $G_t/GI/s_t{+}GI$ model with customer abandonment. We first consider cases in which the service and patience distributions are exponential (M). We next extend to the framework non-exponential service and patience times.

6.1 Extension of the Algorithm for the Model with Exponential Patience Times

In particular, it is natural to use (3.1) with the same peakedness z in (3.5), but with the MSHT QoS parameter $\beta_\alpha(1) \equiv H^{-1}(\alpha)$ in (3.3) replaced by $G^{-1}(\alpha)$, where G is the Garnett MSHT PoD function from pp. 217-218 of [15], which is based on the QED MSHT limit for the stationary $M/M/s+M$ model. As noted previously, the MSHT limit for the $M/M/s+M$ model extends to the associated $G/M/s+M$ model by §7.3 of [40]. Just as for the $G_t/GI/s_t$ that we have considered, the extensions to GI service and GI abandonment are heuristic.

From (3.9) of [18], the Garnett PoD function can be written as

$$G(\beta) \equiv G(\beta, \theta_{rat}) \equiv \left[1 + \frac{\sqrt{\theta_{rat}}h(\beta/\sqrt{\theta_{rat}})}{h(-\beta)}\right]^{-1}, \quad -\infty < \beta < \infty, \tag{6.1}$$

where $\theta_{rat} \equiv \theta/\mu$ and $h(x) \equiv \phi(x)/\bar{\Phi}(x)) = \phi(x)/(1-\Phi(x))$ is the hazard rate of the standard normal distribution.

Unfortunately, there are typographical errors in other representation of the Garnett function. First, an alternative expression is given for the Garnett function G in (11) on p. 331 of [12], but there is a typo in the definition of $\hat{\beta}$ below (11). It should be $\hat{\beta} \equiv \beta\sqrt{\mu/\theta}$ or $\hat{\beta} \equiv \beta/\sqrt{\theta/\mu}$ instead of $\hat{\beta} \equiv \beta\sqrt{\theta/\mu}$. Yet another alternative expression of G is given in (4) on p. 1553 of [30], but it too has a problem. The intended formula for h there is $h(x) \equiv \phi(x)/\bar{\Phi}(x)$, where $\bar{\Phi}(x) \equiv 1 - \Phi(x)$, consistent with established notation in that paper, but the bar cannot be seen.

6.2 Extension for Non-Exponential Service and Patience Times

We also conducted simulation experiments for several $G_t/GI/s_t + GI$ models with non-exponential service times and patience times. We have found that our approach for the $G_t/M/s_t + M$ model in §8 of the main paper continues to work well in many cases, but needs refinement in some cases. We see stable performance in all cases, but not always at the desired target. The major difficulty encountered was for non-exponential patience times. A basis for extension to the $G_t/M/s_t + GI$ model is provided by results in Theorem 4.1 of [56] (see also p. 1196 of [36]), which suggests a refined Garnett function

$$G^*(\beta) \equiv G^*(\beta, \theta^*_{rat}) \equiv \left[1 + \frac{\sqrt{\theta^*_{rat}}h(\beta/\sqrt{\theta^*_{rat}})}{h(-\beta)}\right]^{-1}, \quad -\infty < \beta < \infty, \tag{6.2}$$

where $\theta^*_{rat} \equiv f(0)/\mu$, $f(0)$ is the patience-time pdf at $x = 0$ and $h(x) \equiv \phi(x)/\bar{\Phi}(x) = \phi(x)/(1-\Phi(x))$ is the hazard rate of the standard normal distribution. This generalization from M abandonment to GI abandonment is quite intuitive: because the system is in the QED regime where waiting times are asymptotically negligible, the patience-time distribution plays an role only through the patience hazard rate at 0, that is $h_F(0) = f(0)/\bar{F}(0) = f(0)$. Even though there are not yet any supporting MSHT limits for the more general stationary $G/GI/s+GI$ model with non-M service, we propose the same approximation based on (6.2) for the $G_t/GI/s_t+GI$ model too. In particular, to capture non-M abandonment we use (6.2) instead of (6.1); to cope with non-M service, we again rely on the peakedness z in (3.5), which depends on the non-M service.

7 Simulation Experiments

We now report results of simulation experiments to evaluate the new MOL staffing algorithm for the $G_t/GI/s_t$ model given in §2 and §3.

7.1 The Simulation Models

For all the examples, the system starts empty, the service time has mean 1 and the G_t arrival process has deterministic sinusoidal arrival rate

$$\lambda(t) = \bar{\lambda}(1 + \psi_\lambda \sin(\gamma_\lambda t + \phi_\lambda)), \quad t \in [0, 96], \tag{7.1}$$

with average arrival rate $\bar{\lambda}$, relative amplitude ψ_λ, $0 \le \psi_\lambda \le 1$, period (cycle length) $2\pi/\gamma_\lambda$ and phase shift ϕ_λ. Our base case has $\bar{\lambda} = 100$, $\psi_\lambda = 0.2$, $\gamma_\lambda = 1$ and $\phi_\lambda = 0$. Explicit formulas for the associated offered load $m(t)$ for this sinusoidal arrival rate are given in [11] and (19) of [30].

We construct the arrival process as indicated in §2. In each case, we let the stochastic counting process N be a rate-1 stationary counting process. Our base case is an $H_2(1,4)$ renewal process, which is also an IPP, the special MMPP with two states in the underlying CTMC with the rate in one state being 0. The H_2 distribution was characterized for Table 1. For $H_2(1,4)$, the probabilities on the two exponential components are $p_1 \equiv p = (5 + \sqrt{15})/10 = 0.8873$ and $1 - p \equiv 0.1127$, while the rates (reciprocals of the two means) are $\mu_1 = 2p = 1.7745$ and $\mu_2 = 2(1-p) = 0.2254$. From §2.3.11 of [14], the associated IPP parameters are: rate in the on state $\lambda_{on} = 4p^2 + 4(1-p)^2 = 1.60$, the mean time in the on state is $1/\mu_{on} = 1/0.15 = 6.667$ and the mean time in the off state is $1/\mu_{off} = 1/0.40 = 2.500$. Our overall base case is the $H_2^t(1,4)/LN(1,4)/s_t$ model.

We also consider variations on our base case. For the arrival process N, we consider other rate-1 renewal processes with non-H_2 inter-renewal times and other non-renewal MMPP's. For the service-time cdf, we also consider the other service-time cdf's in Table 1. We make the renewal arrival process stationary by letting the first interval have the equilibrium stationary-excess cdf, as in §V.3 of [3].

7.2 Simulation implementation

The simulation experiments were performed with MATLAB. Since we are interested in the virtual waiting time, i.e. the delay of a potential arrival at each time t, we generate virtual customers at each fixed time $\triangle t, 2\triangle t, 3\triangle t, \ldots$, with $\triangle t = 0.05$. Those virtual customers are different from real customers, because once they enter the service, they leave the service immediately, so that they do not occupy any service resource. They are not counted in queue length. If the number of servers needs to decrease while all servers are busy, we wait until the next customer to finish service then remove that server.

System performance measures are measured at the fixed time points $\triangle t, 2\triangle t, \ldots$. We record the queue length $\hat{Q}(t)$ then take the average over all replications. We also calculate potential waiting time $\hat{W}(t)$ which is defined as the waiting time of a virtual costumer that arrives at time t, then take the average of all replications. The estimated probability of delay $\hat{P}_D(t)$ is calculated as the average of the indicator variable $\mathbf{1}_{\{\hat{W}(t)>0\}}$ over all replications.

We ran 1000 independent replications to obtain the estimates of all the performance measures. To understand why this yields adequate statistical precision, note that for a delay probability of about 0.1 at a single time t, our approach corresponds to looking at the average of 1000 i.i.d Bernoulli random variables with approximate mean 0.1 and variance $0.09 \approx 0.1$, making the sample mean have mean 0.1 and sample variance of about $\bar{s}_n^2 \approx 10^{-4}$ with associated sample standard deviation of about $\bar{s}_n \approx 10^{-2}$. Thus the halfwidth of a 95% confidence interval would be approximately 0.00067, which is about 0.7% of the mean 0.10. As in [21], the larger oscillations we see in simulation estimates are primarily due to the significant impact of changing a single agent. (Recall that the staffing is in integer values.)

7.3 Performance Estimates in the Base Case

We now report results for the $H_2^t(1,4)/LN(1,4)/s_t$ base case, with the distributions of the i.i.d. interarrival times of N and the service times as specified in §4. First, Figure 2 shows the estimated time-varying probability of delay (PoD) for the $H_2^t(1,4)/LN(1,4)/s_t$ base case with $z = 2.11$, for five PoD targets α using the MOL SRS formula (3.1) (left) and using one server less (right). All plots here show an initial transient associated with starting empty, but stabile performance is seen after

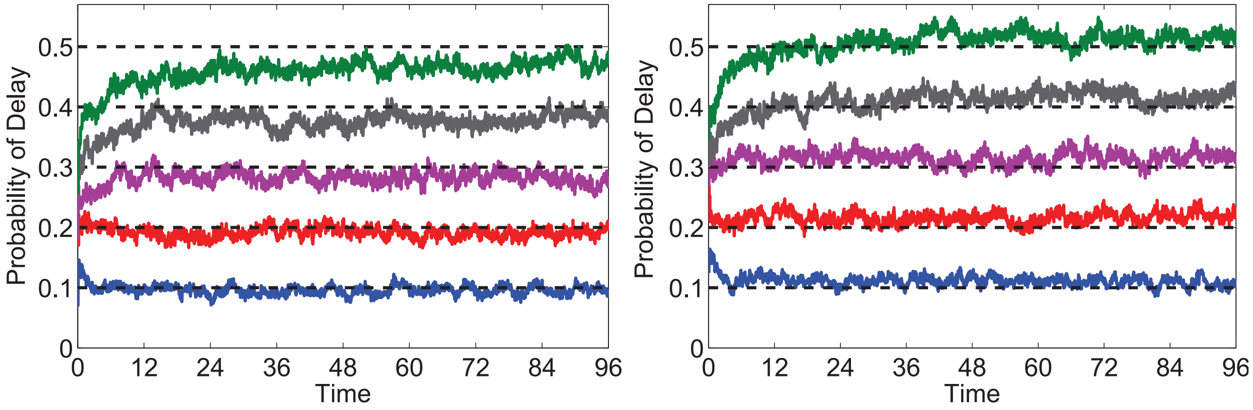


Figure 2: Estimated time-varying probability of delay for the $H_2^t(1,4)/LN(1,4)/s_t$ model ($z = 2.11$) with the MOL SRS staffing (3.1) $s(t)$ (left) and one less server $s(t) - 1$ (right), for five delay probability targets α.

a short time. (The mean service time is 1.) We only show targets $\alpha \leq 0.5$, because higher targets tend to be inconsistent with practical staffing levels without customer abandonment (which will be discussed in §6). Higher targets α tends to move the system out of the quality-and-efficiency-driven (QED) regime into the more heavily loaded efficiency-driven (ED) regime. To provide evidence, we show the average traffic intensity for each of the five cases of Figure 2 in Table 2.

To show that our extension of the MSHT MOL SRS algorithm in (3.1) performs just as well for the non-Markov $H_2^t(1,4)$ arrival process as the previous MSHT MOL SRS algorithm with $z = 1$ in [21] performs for the $M_t/M/s_t$ model, Figure 1 of the EC shows the estimated time-varying probability of delay (PoD) for the $M_t/M/s_t$ model with $z = 1$ on the left and for the $H_2^t(1,4)/LN(1,4)/s_t$ base case with $z = 2.11$ on the right.

To drill down deeper into the results in Figure 2, we display the average, maximum and minimum of the PoD for $t \in [36, 96]$ as a function of the target for the base model with the specified staffing ($s(t)$) and for one less server ($s(t) - 1$) in Table 3, also see Figure 2 for plot comparison. For all five targets, the average PoD falls below the target, while the average PoD with one less server lies above the target. The fact that the maximum estimated PoD for all time points is above the target, while the minimum with one less server is below the target, indicates that (i) the performance is indeed stabilized over time, after an initial transient, and (ii) the performance and statistical precision are

α	0.1	0.2	0.3	0.4	0.5
$\bar{\rho}$	0.828	0.865	0.891	0.912	0.930

Table 2: Time average $\bar{\rho}$ of the instantaneous traffic intensity $\rho(t)$ for the $H_2^t(1,4)/LN(1,4)/s_t$ model using the MOL SRS staffing

	Average(±HW) (diff. to target)		Max (diff. to target)		Min (diff. to target)	
Target	$s(t)$	$s(t)-1$	$s(t)$	$s(t)-1$	$s(t)$	$s(t)-1$
0.5	0.468(±0.0219) (-0.032)	0.516(±0.0219) (+0.016)	0.503 (+0.003)	0.550 (+0.050)	0.437 (-0.063)	0.478 (-0.022)
0.4	0.377(±0.0212) (-0.023)	0.418(±0.0216) (+0.018)	0.416 (+0.016)	0.449 (+0.049)	0.344 (-0.056)	0.387 (-0.013)
0.3	0.282(±0.0197) (-0.018)	0.315(±0.0203) (+0.015)	0.316 (+0.016)	0.352 (+0.052)	0.251 (-0.049)	0.282 (-0.018)
0.2	0.192(±0.0172) (-0.008)	0.217(±0.0181) (+0.017)	0.219 (+0.019)	0.247 (+0.047)	0.166 (-0.034)	0.188 (-0.012)
0.1	0.0956(±0.0129) (-0.0044)	0.111(±0.0137) (+0.011)	0.123 (+0.023)	0.134 (+0.034)	0.0755 (-0.0245)	0.0855 (-0.0145)

Table 3: Average, maximum and minimum of the probability of delay for $t \in [36, 96]$ as a function of the target for the base model $H_2^t(1,4)/LN(1,4)/s_t$ with the specified staffing ($s(t)$) and for one less server ($s(t)-1$). The halfwidths (HW) of 95% confidence intervals are shown.

within the difference caused by the change of a single server. In addition, the change of one server plays a bigger role for higher α (smaller $s(t)$) and a smaller role for lower α (bigger $s(t)$). We later demonstrate in §8.1 the effect of changing one server for a smaller system with $\bar{\lambda} = 10$.

We emphasize that this staffing algorithm is not simply choosing the staffing to make the time-varying instantaneous traffic intensity $\rho(t) \equiv \lambda(t)/\mu s(t)$ constant. Figure 3 shows the instantaneous traffic intensity resulting from the MOL algorithm applied to the base case for three PoD targets: $\alpha = 0.1, 0.3, 0.5$. See Figures 1-3 of [21] to see that other staffing alternatives such as PSA and constant staffing at the average load perform very badly.

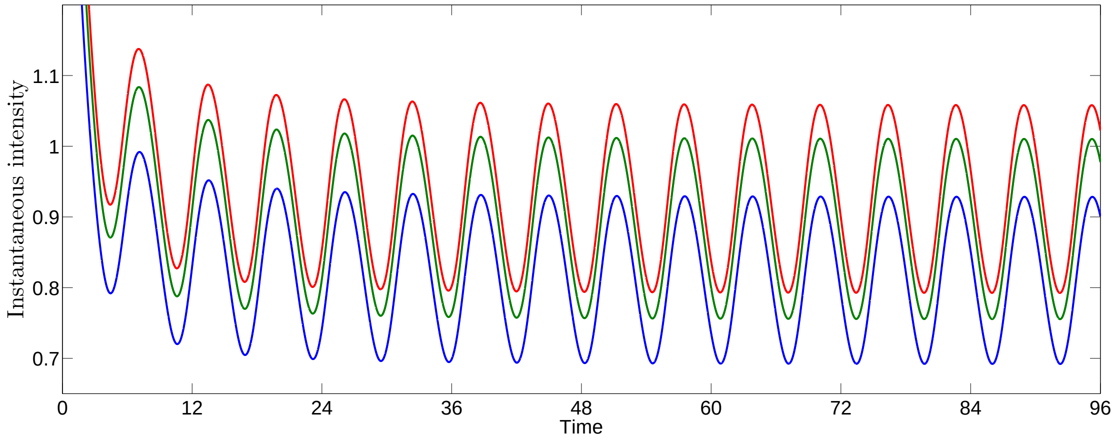


Figure 3: The instantaneous traffic intensities $\rho(t) \equiv \lambda(t)/\mu s(t)$ for the $H_2^t(1,4)/LN(1,4)/s_t$ model with $\mu = 1$ and MOL SRS staffing for $\alpha = 0.1, 0.3, 0.5$ from bottom to top.

We now investigate the extent to which other performance measures are stabilized by the MOL SRS staffing algorithm. Figure 4 shows the estimated time-varying mean queue length $E[Q(t)]$ (left) and mean waiting time $E[W(t)]$ (right) for the base model. As in all previous studies, we find that the mean waiting times are not always stabilized, but all performance measures tend to be stabilized with low PoD targets, where we aim to provide high QoS; e.g., see Table 4 of [21] and §3 of the e-companion to [12]. While [12] primarily focuses on an interative simulation algorithm (ISA) for staffing, it also provides strong support for the OL approach using the SRS formula by showing that the implied empirical quality of service $\beta^{ISA}(t) \equiv (s^{ISA}(t) - m(t))/\sqrt{m(t)}$ in (10) of [12] is stabilized by ISA; see Figures 3 and 12 of the e-companion to [12]. Significant fluctuations were observed in both the expected waiting times in the $M_t/M/s_t$ model and in the abandonment probabilities in the $M_t/M/s_t + M$ model; see Figures 6 and 13 of the e-companion to [12]. These observations are confirmed by Figure 4.

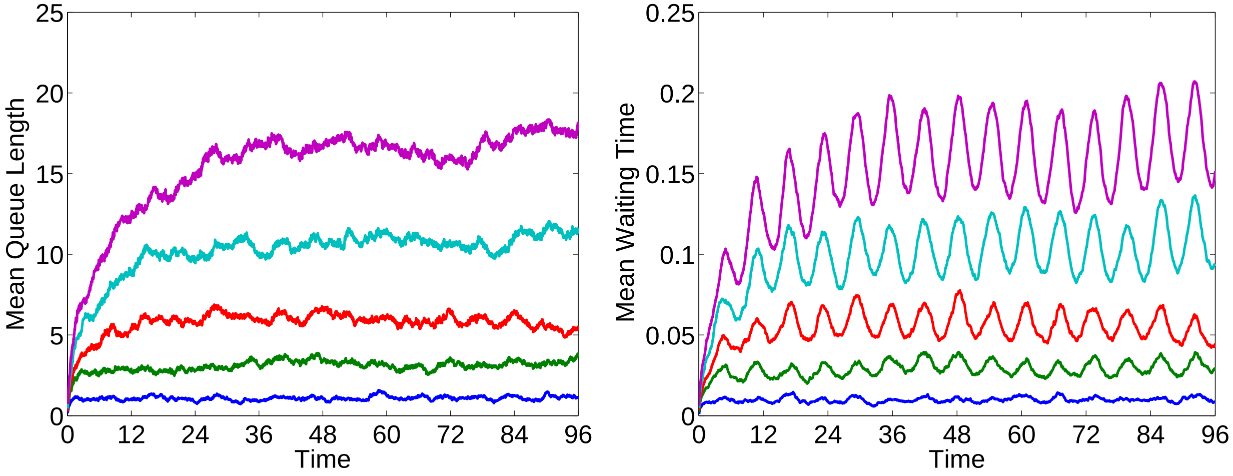


Figure 4: Estimated time-varying mean queue length $E[Q(t)]$ (left) and mean waiting time $E[W(t)]$ (right) for the $H_2^t(1,4)/LN(1,4)/s_t$ model with $z = 2.14$ using the MOL SRS formula (3.1) for the five delay probability targets $\alpha = 0.1, 0.2, 0.3, 0.4, 0.5$.

7.4 The Consequence of Using the Old MOL SRS Algorithm

We now show the consequence of using the old MOL SRS staffing, i.e., (3.1) with $z = 1$. Figure 5 shows the performance of the MOL SRS staffing with $z = 1$ applied to the $H_2^t(1,4)/LN(1,4)/s_t$ model, with $z = 2.11$, for targets $\alpha = 0.1, 0.3, 0.5$. Figure 5 shows that the staffing algorithm with $z = 1$ still stabilizes performance; the refinements are needed only to hit the PoD target α. Figure 5 also shows the significantly higher staffing levels required with the higher value of z.

8 Variations of the Base Model

In this section we report results of the MSHT MOL SRS algorithm for variations of the base model. We first consider higher QoS (lower α targets) and smaller scale. Then we consider alternative ar-

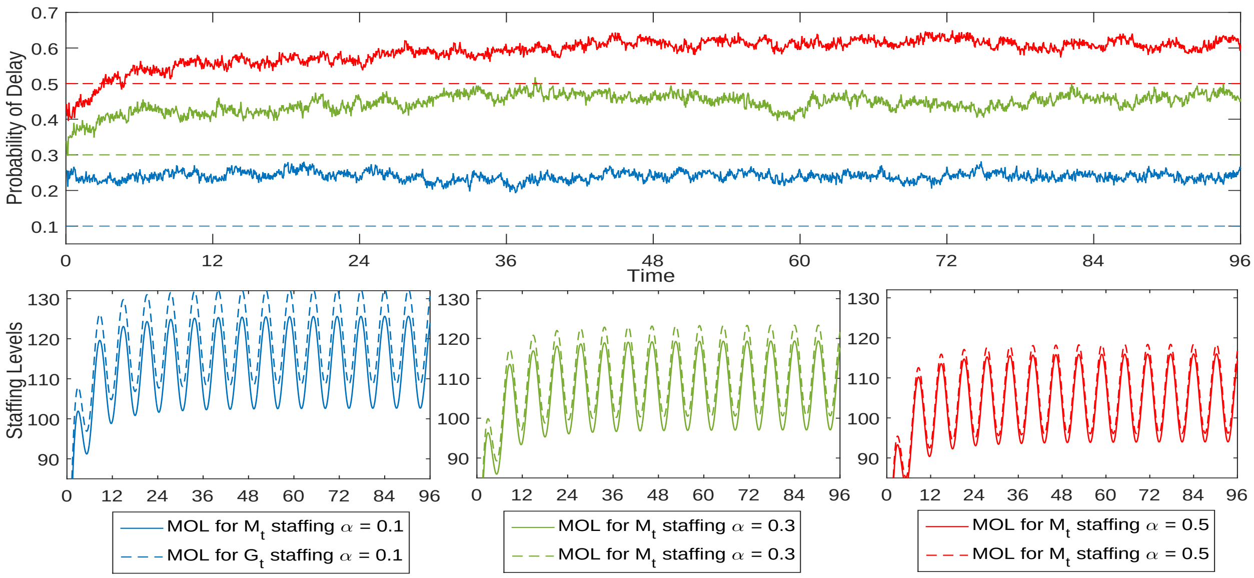


Figure 5: Estimated probability of delay for the $H_2^t(1,4)/LN(1,4)/s_t$ model with $z = 2.11$ using the MOL staffing algorithm in (3.1) with $z = 1$ as would be done for the same model with an M_t arrival process, for targets $\alpha = 0.1, 0.3, 0.5$ (above) and comparison of the associated staffing levels using the MOL staffing for $H_2^t(1,4)$ and M_t arrivals (below).

rival processes and service-time distributions. Finally, we report results evaluating the performance of our MOL algorithm for models with customer abandonment.

8.1 Lower Targets and Lower Arrival Rates

In this section we consider the performance of the SRS MOL staffing algorithm with lower targets α (higher QoS) and for lower average arrival rate, and thus smaller scale (fewer servers).

First, Figure 6 shows on the top the estimated PoD for the base $H_2^t(1,4)/LN(1,4)/s_t$ model with four low targets α less than 0.1, ranging from 0.02 to 0.08. On the bottom of Figure 6 is shown the associated higher time-varying staffing levels required for the target $\alpha = 0.02$.

Next, Figure 7 are displayed the estimated PoD's for the base $H_2^t(1,4)/LN(1,4)/s_t$ model with the average arrival rate $\bar{\lambda}$ reduced from 100 to 10, i.e., for the arrival rate function $\lambda(t) = 10+2\sin(t)$. The reduced offered load leads to reduced staffing accordingly; the old OL $m(t)$ in (3.2) is now simply divided by 10, while the peakedness z is unchanged. Hence, unlike the case on the left, each single server matters much more. Figure 7 shows that the MSHT MOL SRS algorithm in (3.1) still stabilizes the delay probability in these new cases. However, the performance falls further below the target at the higher PoD targets (left-hand plot in Figure 7). But note that a single server makes a much greater difference now (right-hand plot in Figure 7). Despite the rather unconvincing left plot, from both plots, we can see that the stabilization at the target α has been achieved as well

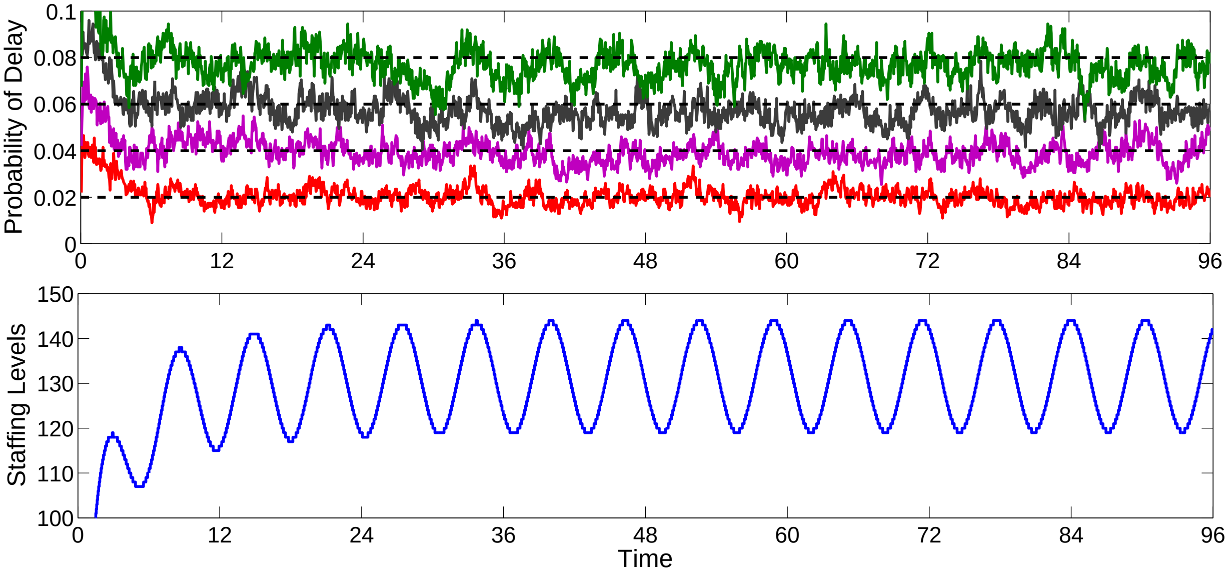


Figure 6: Estimated time-varying probability of delay for the $H_2^t(1,4)/LN(1,4)/s_t$ model with four low targets α less than 0.1 (top) and associated staffing for the case $\alpha = 0.02$ (bottom).

as possible, because there is a substantial gap for $s(t)$, but understaffing with $s(t)-1$. With fewer servers, each server matters more; there is a limit to what is possible. Figure 14 in the e-companion shows similar performance for the more challenging case $\bar{\lambda} = 4$.

8.2 Alternative Arrival Processes

We now consider the MOL SRS staffing algorithm to the base $H_2^t(1,4)/LN(1,4)/s_t$ model except that we change the arrival process. First, we considered the performance for a deterministic D^t arrival process and an E_2^t Erlang renewal arrival process, which have the same deterministic arrival rate function, but has N a stationary D and E_2 renewal process. These processes are less variable than a Poisson process, having asymptotic variability parameters (equal to the interarrival times scv) of $c_A^2 = 0$ and $c_A^2 = 0.5$, respectively. Such low-variability arrival processes commonly occur in service systems with arrivals by appointment. Figure 15 in the e-companion shows that the same excellent performance holds in these low-variability examples.

As noted in §2.2, our base $H_2^t(1,4)$ arrival process is constructed from an $H_2(1,4)$ renewal process, which also is an IPP (a special MMPP). We next consider non-renewal MMPP's as the arrival process. In particular, we consider an MMPP with an underlying CTMC $\{\Gamma(t), t \geq 0\}$ that is a birth-and-death process having three states 0, 1 and 2. Let $Z(t) = f(\Gamma(t))$ with state-dependent

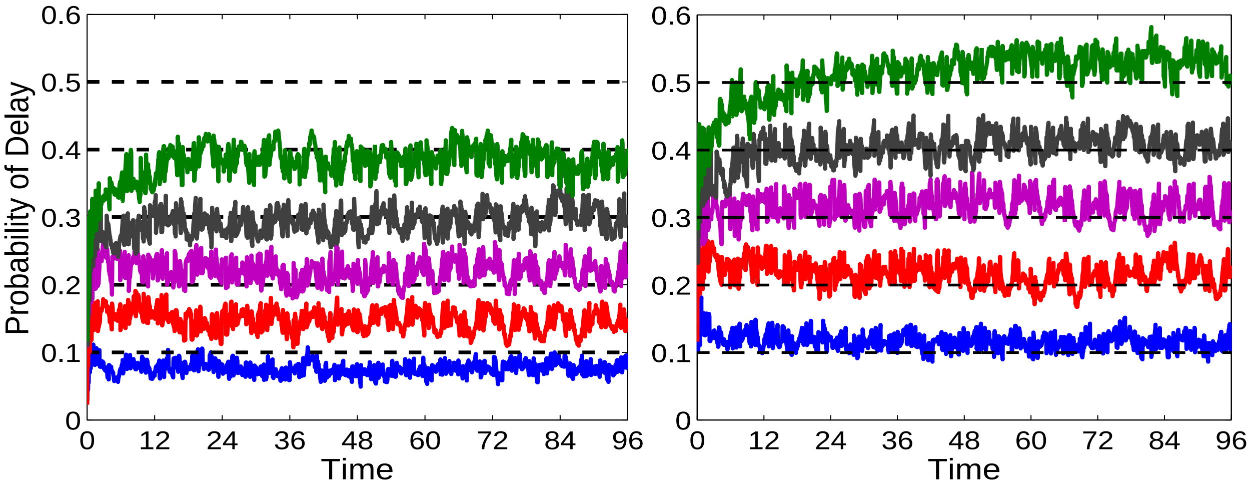


Figure 7: Estimated time-varying probability of delay in the base $H_2^t(1,4)/LN(1,4)/s_t$ model with the same targets as before, but with the average arrival rate $\bar{\lambda}$ reduced from 100 to 10, using the MOL SRS formula (3.1) $s(t)$ (left) and $s(t)-1$ (right).

rate $f(i) = \lambda_i$, where $(\lambda_0, \lambda_1, \lambda_2) = (3, 1, 1/3)$. The long-run rate of the MMPP is

$$\lambda_C = \lim_{t\to\infty} t^{-1}C(t) = \lim_{t\to\infty} t^{-1}\int_0^t Z(s)ds = \lim_{t\to\infty} t^{-1}\int_0^t f(\gamma(s))ds = \sum_{j=0}^{2} \pi_j\lambda_j \equiv \lambda^*,$$

where $\pi \equiv (\pi_0, \pi_1, \pi_2)$ is the steady state distribution for the CTMC. We consider two sets of birth and deaths rates (i) $\hat{\lambda}_0 = 2$, $\hat{\lambda}_1 = 1.5$, $\hat{\mu}_1 = \hat{\mu}_2 = 1$ and (ii) $\hat{\lambda}_0 = 20/27$, $\hat{\lambda}_1 = 5/9$, $\hat{\mu}_1 = \hat{\mu}_2 = 10/27$, which yield the same steady state $\pi = (1/6, 1/3, 1/2)$ and asymptotic rate of MMPP $\lambda_C = \lambda^* = 1$, but different variability parameter of C: (i) $c_C^2 = 10/9$ and (ii) $c_C^2 = 3$, where c_C^2 is given by

$$c_C^2 = \frac{\bar{\sigma}_C^2}{\lambda_C} = \bar{\sigma}_C^2 = 2\sum_{j=0}^{1} \frac{1}{\hat{\lambda}_j\pi_j}\left[\sum_{i=0}^{j}(\lambda_i - \lambda^*)\pi_i\right]^2.$$

See Proposition 1 of [48] for details, also see [14]. Because M is a rate-1 Poisson process, the stochastic variability parameters for the G_t arrival are (i) $c_A^2 = c_M^2 + c_C^2 = 1 + 10/9 = 19/9$ and (ii) $c_A^2 = 1 + 3 = 4$. Figure 8 shows the time-varying delay probability for different targets α with the $MMPP^t/LN(1,4)/s_t$ model having MMPP arrivals with $c_A^2 = 19/9$ (left) and $c_A^2 = 4$ (right). Clearly the performance is again excellent.

8.3 Alternative Service-Time Distributions

We also conducted experiments for the base $H_2^t(1,4)/LN(1,4)/s_t$ model with different service distributions. Figure 9 shows that the same stable plots of the delay probability hold for exponential (M) and lognormal $LN(1,0.25)$ service times.

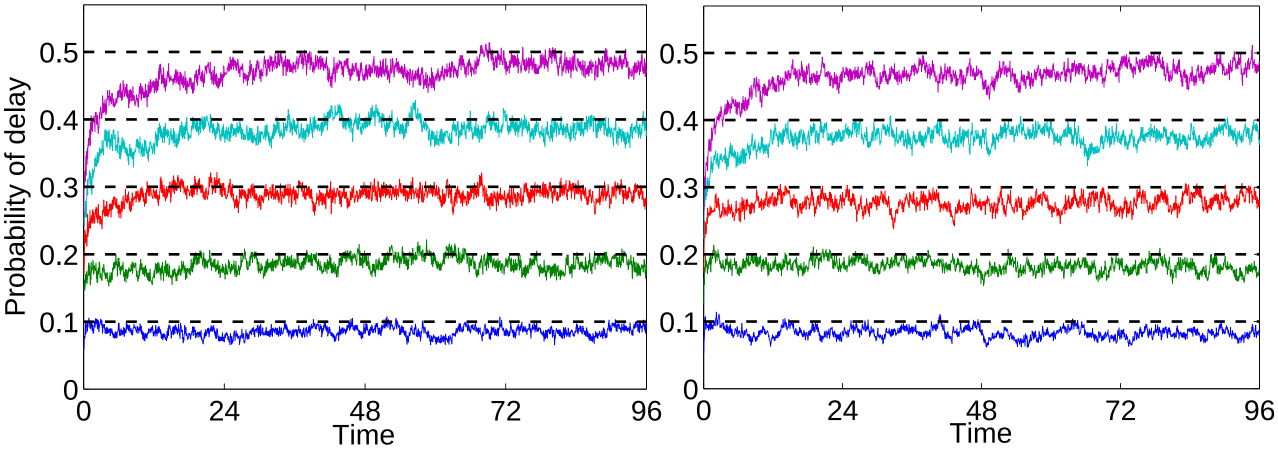


Figure 8: Estimated time-varying probability of delay for the $MMPP^t(1, c_A^2)/LN(1, 4)/s_t$ model with $c_A^2 = 19/9$ (left) and $c_A^2 = 4$ (right), using the MOL SRS formula (3.1) for five delay probability targets α.

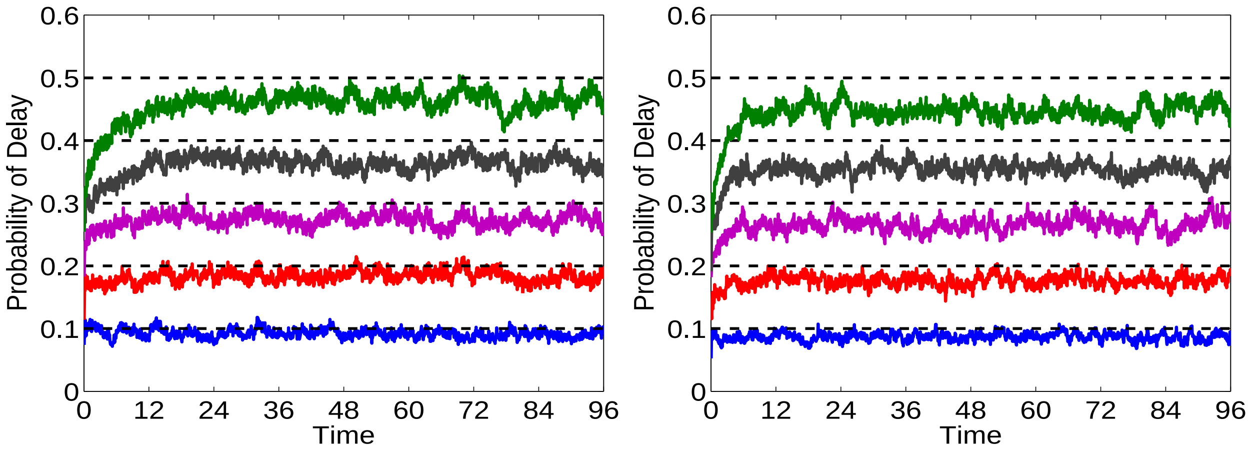


Figure 9: Estimated time-varying probability of delay for the $H_2^t(1, 4)/M/s_t$ model with exponential service times yielding $z = 2.5$ (left) and for the $H_2^t(1, 4)/LN(1, 0.25)/s_t$ model with the low-variabiity lognormal $LN(1, 0.25)$ service times yielding $z = 3.25$ (right) using the MOL SRS formula (1) of the main paper for five delay probability targets α.

8.4 Models with Customer Abandonment

Finally, we conducted simulation experiments evaluating the performance of our new MOL SRS algorithm for the general $G_t/GI/s_t + GI$ model. using the refined Garnett functions in (6.1) and (6.2).

Exponential service and patience times. Figure 10 reports simulation results of this staffing algorithm applied to the $H_2^t(1,4)/M/s_t + M$ model, having our base arrival process and exponential service times with mean $1/\mu = 1$, but now also with customer abandonment for a range of abandonment rates θ from 1/16 to 16. Figure 10 shows that the staffing algorithm is effective for all θ and all delay probability targets $0.1 \leq \alpha \leq 0.9$.

Non-exponential service and patience times. Figure 11 shows the results for the $H_2^t(1,4)/H_2(1,4)/s_t+$ $H_2(1,4)$ model and $H_2^t(1,4)/E_2(1)/s_t + H_2(1,4)$ model. In the e-companion we show corresponding results for models with low-variability, service times and arrival processes, in particular, for the $E_2^t/LN(1,4)/s_t + H_2(1,4)$ and $D^t/LN(1,4)/s_t + H_2(1,4)$ models, having the process N be a renewal process with E_2 and D times between renewals. We find that the performance is stabilized at all targets in all these cases.

Smaller Arrival Rates. Figure 12 shows the results for $\bar{\lambda} = 10$ and $\bar{\lambda} = 4$ for our main $H_2^t(1,4)/LN(1,4)/s_t + H_2(1,4)$ example. We see that (i) our staffing method continue to stabilize the performance for a wide range of targets; and (ii) a single agent matters more with a smaller OL.

9 Conclusions

We have developed (i) a new non-Poisson nonstationary arrival process model in §2 that includes the nonstationary Cox (doubly stochastic Poisson) process as a special case and (ii) a new many-server heavy-traffic (MSHT) modified-offered-load (MOL) square-root-staffing (SRS) algorithm in §3 and §6 for the general $G_t/GI/s_t$ and $G_t/GI/s_t + GI$ models with that arrival process. We have shown that the algorithm is effective for stabilizing the probability of delay with this model by conducting simulation experiments in §§7-8.

In §2 we have shown how to construct and usefully characterize general arrival processes that combine non-standard stochastic variability with significant time variability. First, in §2.1 we

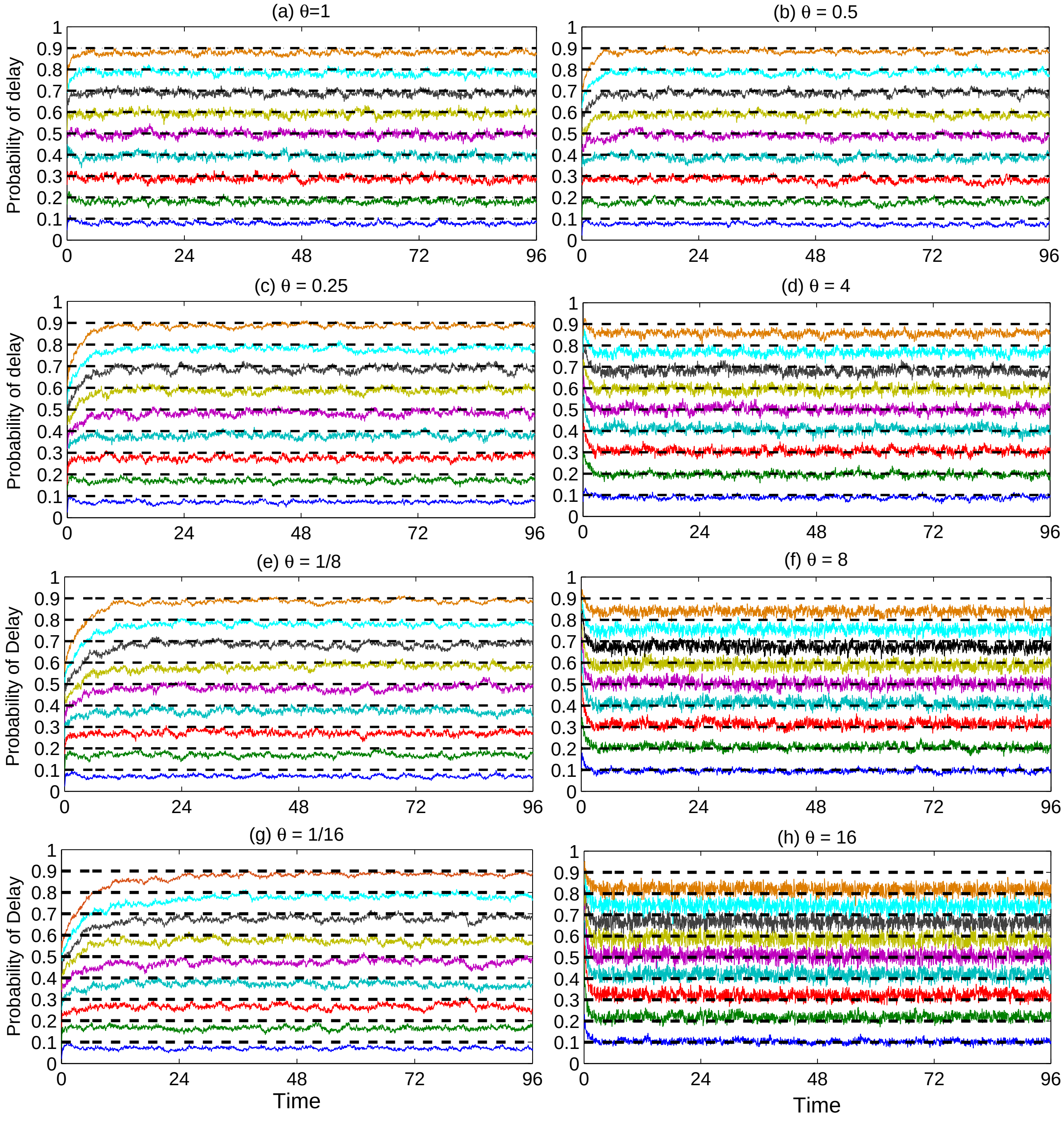


Figure 10: Estimated time-varying probability of delay with nine targets $\alpha = 0.1, \ldots, 0.9$, for the $H_2^t(1,4)/M/s_t + M$ model with $\mu = 1$ and different θ, ranging from 1/16 to 16.

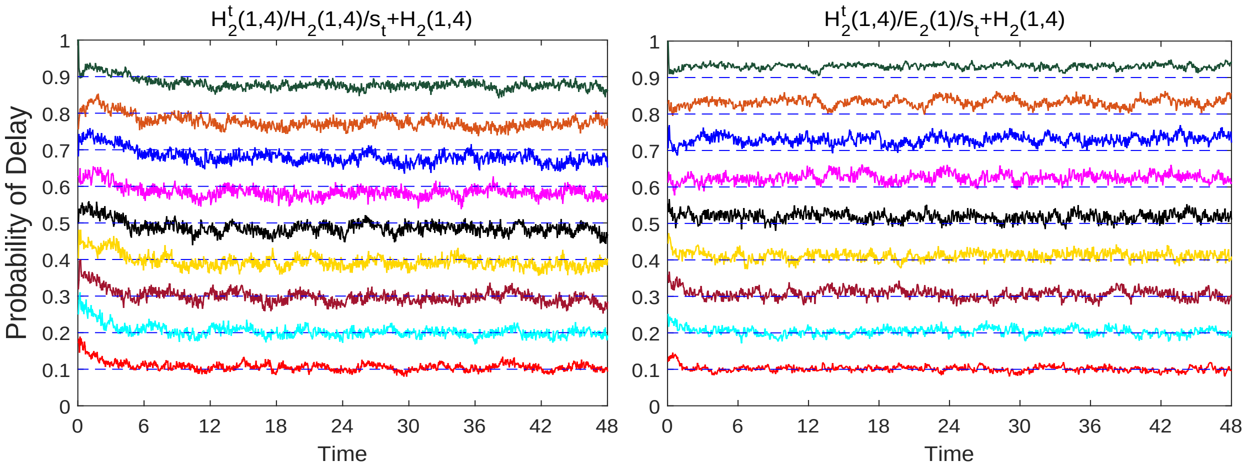


Figure 11: Estimated time-varying probability of delay for the $H_2^t(1,4)/GI/s_t + H_2(1,4)$ model with (a) $H_2(1,4)$ service times yielding $z = 2.05$ (left) and (b) $E_2(1)$ service times yielding $z = 2.88$ (right), and i.i.d. $H_2(1,4)$ patience times, yielding $\theta = \mu = 1$, using the MOL SRS formula (1) and the Zeltyn-Mandelbaum (2005) refinement to the Garnett function in (41) of the main paper for nine delay probability targets α, ranging from 0.1 to 0.9.

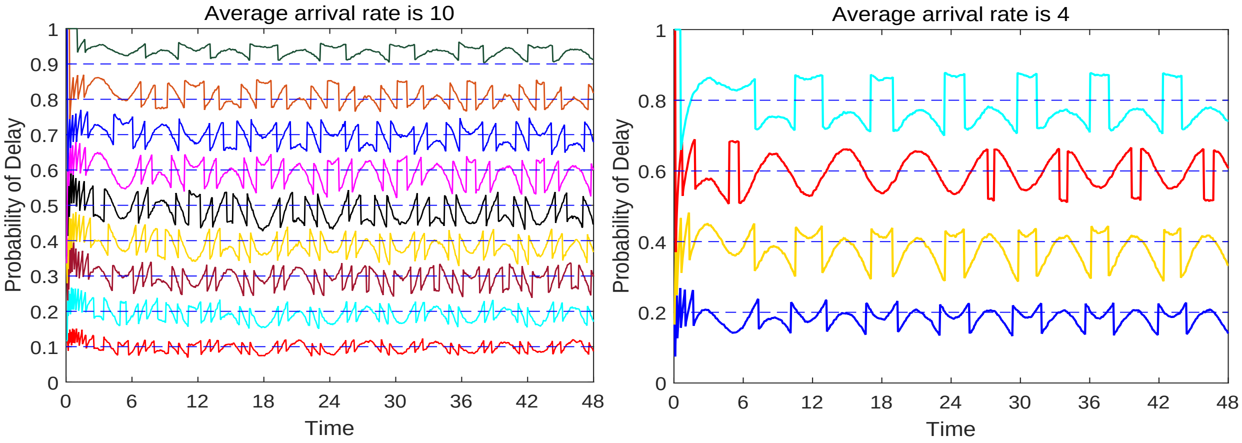


Figure 12: Estimated time-varying probability of delay for the $H_2^t(1,4)/LN(1,4)/s_t + H_2(1,4)$ model with a wide range of targets, but with the average arrival rate $\bar{\lambda}$ reduced from 100 to 10 (left) and to 4 (right).

constructed a general model exploiting composition. In §2.2 we exhibited a special case, which includes the nonstationary Cox process, i.e., a nonhomogeneous Poisson process with a rate function that is itself a stochastic process. In §5 we showed how to compute the asymptotic variability parameter of the arrival process, c_A^2, from stochastic models and estimate it from data without constructing a specific stochastic model, by estimating the index of dispersion $I(t)$ for large t.

The new MSHT MOL SRS algorithm in §3 exploits the approximation for the steady-state delay probability in the stationary $G/GI/s$ model in [51], which is based on the many-server heavy-traffic (MSHT) limit for the $GI/M/s$ model in [19], extended to the $G/M/s$ model by §7.3 of [40]. The new algorithm extends the MSHT MOL approach to staffing introduced for the $M_t/M/s$ model in [21]. The extension exploits the MSHT limit of the peakedness z, i.e., the ratio of the variance to the mean of the steady-state number of busy servers in the associated infinite-server (IS) model, which is supported by the MSHT limits in [31, 41, 43]. The MSHT limit of the peakedness in (3.5) succinctly captures the important nontrivial combined impact of the service-time distribution and the variability in the arrival process on system performance.

Broadly, this paper is useful for showing one way to model and staff for more complex non-Poisson nonstationary arrival processes. Moreover, the analysis in this paper yields useful insights about the impact of stochastic variability upon the performance of many-server queues. First, our analysis supports the conclusion that the variability in the arrival process primarily affects performance and staffing through the asymptotic variability parameter c_A^2 arising in the CLT. Second, there is a complicated interaction between the service-time distribution and the arrival process in their impact upon performance, which tends to be captured by the MSHT limit of the peakedness, as in MSHT limits for the $G/G/\infty$ infinite-server queue in [41, 43]. As discussed in §4, the peakedness representation shows the impact of the service-time variance σ^2 on performance and staffing with a lognormal $LN(1, \sigma^2)$ service distribution. Counter to conventional wisdom, for an arrival process that is more variable than Poisson, the congestion tends to be decreasing in σ^2, so that the commonly found $\sigma^2 \approx 1$ is not helpful compared to a higher variance such as $\sigma^2 \approx 4$ or more.

It is significant that the new staffing algorithm in (3.1) and §3 is relatively simple, being a variant of the widely used square-root-staffing formula. Our results show that even the basic algorithm with $z = 1$ stabilizes performance for our general models. The refinement is important for hitting the delay probability target α. The robustness suggests that variants of our proposed algorithm might be useful in other complex settings.

Nevertheless, it remains to investigate how this new staffing algorithm works in applications with non-Poisson nonstationary arrival processes. Moreover, it remains to develop alternative approaches and compare them. For example, it may prove useful to consider other variants of the SRS algorithm in (3.1), such as the alternative staffing formula $s(t) = m(t) + \beta m(t)^c$ for $c \neq 2$ investigated by [35].

Acknowledgment We thank the National Science Foundation for support: NSF grants CMMI 1362310 (first and second authors) and CMMI 1265070 (third author).

References

[1] Aksin, O. Z., Armony, M. and Mehrotra, V. (2007). The modern call center: a multi-disciplinary perspective on operations management research. *Production and Operations Management* 16:665–688.

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

[3] Asmussen, S. (2003). *Applied Probability and Queues.* New York: Sprnger, 2nd edition.

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

[5] Bassamboo, A. and Zeevi, A. (2009). On a data-driven method for staffing large call centers. *Operations Research* 57(3):714–726.

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

[7] Choudhury, G. L. and Whitt, W. (1994). Heavy-traffic asymptotic expansions for the asymptotic decay rates in the $BMAP/G/1$ queue. *Stochastic Models* 10(2):453–498.

[8] Cox, D. R. and Lewis, P. A. W. (1966). *The Statistical Analysis of Series of Events.* London: Methuen.

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

[10] Defraeye, M. and van Nieuwenhuyse, I. (2015). Staffing and scheduling under nonstationary demand for service: A literature review. *Omega* 58:4–25.

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

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

[13] Fendick, K. W. and Whitt, W. (1989). Measurements and approximations to describe the offered traffic and predict the average workload in a single-server queue. *Proceedings of the IEEE* 71(1):171–194.

[14] Fischer, W. and Meier-Hellstern, K. (1992). The Markov-modulated Poisson process (MMPP) cookbook. *Performance Evaluation* 18:149–171.

[15] Garnett, O., Mandelbaum, A. and Reiman, M. I. (2002). Designing a call center with impatient customers. *Manufacturing and Service Operations Management* 4(3):208–227.

[16] Gerhardt, I. and Nelson, B. L. (2009). Transforming renewal processes for simulation of nonstationary arrival processes. *INFORMS Journal on Computing* 21:630–640.

[17] Glynn, P. W. and Whitt, W. (1993). Limit theorems for cumulative processes. *Stochastic Processes and their Applications* 47:299–314.

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

[19] Halfin, S. and Whitt, W. (1981). Heavy-traffic limits for queues with many exponential servers. *Operations Research* 29(3):567–588.

[20] Ibrahim, R., L'Ecuyer, P., Regnard, N. and Shen, H. (2012). On the modeling and forecasting of call center arrivals. *Proceedings of the 2012 Winter Simulation Conference* 2012:256–267.

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

[22] Jongbloed, G. and Koole, G. (2001). Managing uncertainty in call centers using Poisson mixtures. *Applied Stochastic Models in Business and Industry* 17:307–318.

[23] Kim, S., Vel, P., Whitt, W. and Cha, W. C. (2015). Poisson and non-Poisson properties of appointment-generated arrival processes: The case of an endocrinology clinic. *Operations Research Letters* 43:247–253.

[24] Kim, S.-H. and Whitt, W. (2014). Are call center and hospital arrivals well modeled by nonhomogeneous Poisson processes? *Manufacturing and Service Operations Management* 16(3):464–480.

[25] Kim, S.-H. and Whitt, W. (2014). Choosing arrival process models for service systems: Tests of a nonhomogeneous Poisson process. *Naval Research Logistics,* 61(1):66–90.

[26] Kuczura, A. (1973). The interrupted Poisson process as an overflow process. *Bell System Tech J* 52(3):437–448.

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

[28] Li, A., Whitt, W. and Zhao, J. (2016). Staffing to stabilize blocking in loss models with time-varying arrival rates. *Probability in the Engineering and Informational Sciences* x:nnn.

[29] Liu, R., Liu, Y. and Wilson, J. R. (2014). Modeling and simulation of nonstationary non-Poisson processes. Working paper, North Carolina State University, Raleigh, NC (2014).

[30] Liu, Y. and Whitt, W. (2012). Stabilizing customer abandonment in many-server queues with time-varying arrivals. *Oper Res* 60(6):1551–1564.

[31] Liu, Y. and Whitt, W. (2014). Many-server heavy-traffic limits for queues with time-varying parameters. *Annals of Applied Probability* 24(1):378–421.

[32] Liu, Y. and Whitt, W. (2014). Stabilizing performance in networks of queues with time-varying arrival rates. *Probability in the Engineering and Informational Sciences,* 28(4):419–449.

[33] Liu, Y. and Whitt, W. (2016). Stabilizing performance in a service system with time-varying arrivals and customer feedback. Columbia University, Available at: http://www.columbia.edu/∼ww2040/allpapers.html.

[34] Ma, N. and Whitt, W. (2015). Efficient simulation of non-Poisson non-stationary point processes to study queueing approximations. *Statistics and Probability Letters* 102:202–207.

[35] Maman, S. (2009). Uncertainty in demand for service: the case of call centers and emergency departments. PhD. dissertation, The Technion, Israel, http://iew3.technion.ac.il/serveng.

[36] Mandelbaum, A. and Zeltyn, S. (2009). Staffing many-server queues with impatient customers: constraint satisfaction in call centers. *Operations Research* 57(9):1189–1205.

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

[38] Nelson, B. L. and Gerhardt, I. (2011). Modeling and simulating renewal nonstationary arrival processes to facilitate analysis. *Journal of Simulation* 5:3–8.

[39] Neuts, M. F. (1989). *Structured Stochastic Matrices of $M/G/1$ Type and Their Applications.* New York: Marcel Dekker.

[40] Pang, G., Talreja, R. and Whitt, W. (2007). Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Probabiity Surveys* 4:193–267.

[41] Pang, G. and Whitt, W. (2010). Two-parameter heavy-traffic limits for infinite-server queues. *Queueing Systems* 65:325–364.

[42] Pang, G. and Whitt, W. (2012). The impact of dependent service times on large-scale service systems. *Manufacturing and Service Operations Management* 14(2):262–278.

[43] Pang, G. and Whitt, W. (2013). Two-parameter heavy-traffic limits for infinite-server queues with dependent service times. *Queueing Systems* 73(2):119–146.

[44] Shen, H. and Huang, J. Z. (2008). Interday forecasting ind intraday updating of call center arrivals. *Manufacturing and Service Operations Management* 10(3):391–410.

[45] Sriram, K. and Whitt, W. (1986). Characterizing superposition arrival processes in packet multiplexers for voice and data. *IEEE Journal on Selected Areas in Communications* SAC-4(6):833–846.

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

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

[48] Whitt, W. (1992). Asymptotic formulas for Markov processes with applications to simulation. *Operations Research* 40(2):279–291.

[49] Whitt, W. (1992). Understanding the efficiency of multi-server service systems. *Management Science* 38(5):708–723.

[50] Whitt, W. (2002). *Stochastic-Process Limits.* New York: Springer.

[51] Whitt, W. (2004). A diffusion approximation for the $G/GI/n/m$ queue. *Operations Research* 52(6):922–941.

[52] Whitt, W. (2006). Staffing a call center with uncertain arrival rate and absenteeism. *Production and Operations Management* 15(1):88–102.

[53] Whitt, W. (2015). Stabilizing performance in a single-server queue with time-varying arrival rate. *Queueing Systems* 81:341–378.

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

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

[56] Zeltyn, S. and Mandelbaum, A. (2005). Call centers with impatient customers: many-server asymptotics of the $M/M/n+G$ queue. *Queueing Systems* 51(5):361–402.

[57] Zhang, X., Hong, L. J. and Glynn, P. W. (2014). Timescales in modeling call center arrivals. Working paper, Department of Industrial Engineering and Logistics Management, The Hong Kong University of Science and Technology.

APPENDIX

This is an appendix to the main paper. We display additional results from simulation experiments that examine the performance of the proposed staffing algorithm.

After giving a brief review in §A, we consider the base model with very low arrival rate in §B, in particular, the average arrival rate $\bar{\lambda}$ reduced from 100 to 4. In §C we consider the performance of the base model modified to have different arrival processes, in particular, with variability less variable than Poisson, instead of more variable than Poisson. In §D we consider the performance for additional models with customer abandonment.

A Brief Review

Recall that we applied the MOL SRS staffing algorithm to the $H_2^t(1,4)/LN(1,4)/s_t$ base model with the sinusoidal arrival rate function

$$\lambda(t) = \bar{\lambda}(1 + \psi_\lambda \sin(\gamma_\lambda t + \phi_\lambda)), \quad t \in [0, 96], \tag{A.1}$$

with average arrival rate $\bar{\lambda}$, relative amplitude ψ_λ, $0 \leq \psi_\lambda \leq 1$, period (cycle length) $2\pi/\gamma_\lambda$ and phase shift ϕ_λ. Our base case has $\bar{\lambda} = 100$, $\psi_\lambda = 0.2$, $\gamma_\lambda = 1$ and $\phi_\lambda = 0$. The arrival process is constructed from an H_2 renewal process (having hyperexponential inter-renewal times), which is a special MMPP. The service-time distribution is lognormal. These $H_2(1,4)$ and $LN(1,4)$ distributions are specified in §4 and §6.1 of the main paper.

First, Figure 13 shows the estimated time-varying probability of delay (PoD) for the $M_t/M/s_t$ model with $z = 1$ on the left and for the $H_2^t(1,4)/LN(1,4)/s_t$ base case with $z = 2.11$ on the right, using the MOL SRS formula (10) in the main paper for five PoD targets α. Of course, the plots on the left in Figure 13 just confirm the results of [21]. The plots on the right in Figure 13 show that our extension of the MSHT MOL approximation performs just as well for the non-Markov $H_2^t(1,4)$ arrival process.

B Low Arrival Rates

We first consider the $H_2^t(1,4)/LN(1,4)/s_t$ base model having the sinusoidal arrival rate in (A.1) with $\bar{\lambda} = 100$ reduced to $\bar{\lambda} = 4$. Figure 14 shows the performance with the SRS staffing $s(t)$ (left) and $s(t) - 1$ (right). We observe that the change of a single server now makes an even greater difference to the performance than the case $\bar{\lambda} = 10$, shown in Figure 8 of the main paper. Because the overall staffing levels are low, the change of one server (as time evolves) account for the relatively large fluctuations of the PoD.

C Alternative Arrival Processes

Figure 15 shows the performance for the $H_2^t(1,4)/LN(1,4)/s_t$ base model with $\bar{\lambda} = 100$ modified to have deterministic D^t arrival process (left) and an E_2^t Erlang renewal arrival process, which has the same deterministic arrival rate function but has N a stationary D and E_2 renewal process. These processes are less variable than a Poisson process, having asymptotic variability parameters (equal to the interarrival times scv) of $c_A^2 = 0$ and $c_A^2 = 0.5$, respectively.

The associated peakedness in these two cases is $z = 0.64$ and $z = 0.82$, both less than 1. Just as for the MOL algorithm from [21] for the Markovian $M_t/LN(1,4)/s_t$ on the left and for our new MOL algorithm for the base case $H_2^t(1,4)/LN(1,4)/s_t$ model on the right in Figure 2 of the main paper, we see that the performance target is met perfectly at the lower PoD targets, but there is some gap at the higher PoD targets, but we have seen that this is due to the impact of a single server.

D Customer Abandonment

In this final section we show the results of additional simulation experiments for $G_t/GI/s_t + GI$ models with non-exponential service times and patience times.

D.1 Different Service Variability and Abandonment Rates

Figure 16 shows the performance for the $H_2^t(1,4)/LN(1,v)/s_t + M(m)$ for all combinations of three variances $v = 0.25$, 1.0, 4 of the mean-1 lognormal service-time distribution and three mean values for the exponential patience distributions: $m = 0.25$ and 4.0. The performance is consistently good, except in the case of a low-variability $LN(1,0.25)$ service distribution and a high-abandonment-rate $M(0.25)$ patience distribution, with mean 0.25 and abandonment rate of 4, appearing in the top-left plot of Figure 16. The performance remains good for the low targets, but we see under-staffing at the high targets.

D.2 Nonexponential Patience Distributions

The left-hand plots in Figures 17 and 18 show two cases with $H_2(1,4)$ patience distribution, the first for $M(1)$ service and the second for $LN(1,4)$ service. Again we see stable performance, but both give evidence of over-staffing, because the delay probabilities fall below the targets for the higher targets. These cases do not yet benefit from the refined steady-state delay-probability approximation in [56].

From [56], we know that the steady-state distribution for the non-M abandonment requires a modification of the Garnett function. The right-hand plots in Figures 17 and 18 show the performance in the left-hand plots after the refinement has been made. These new figures show significant improvement, notably at the higher targets.

D.3 Low Arrival Variability

We consider models with low arrival variabilities in Figure 19 for the $E_2^t/LN(1,4)/s_t + H_2(1,4)$ and $D^t/LN(1,4)/s_t + H_2(1,4)$ models, where the base process N is a renewal process with E_2 and D times between renewals, and show that the performance is stabilized at all targets in all these cases.

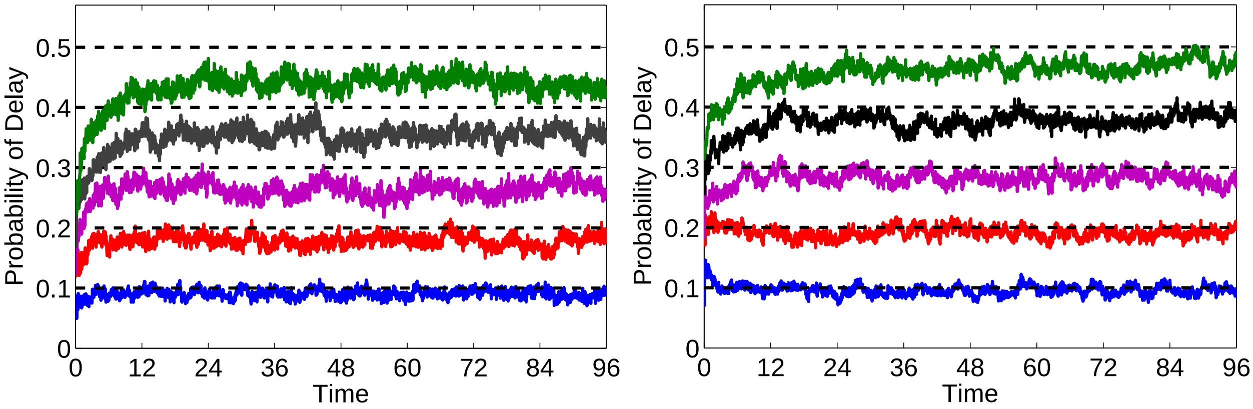


Figure 13: Estimated time-varying probability of delay for the $M_t/M/s_t$ model ($z = 1$, left) and the $H_2^t(1,4)/LN(1,4)/s_t$ model ($z = 2.11$, right) using the MOL SRS formula (3.1) for five delay probability targets α.

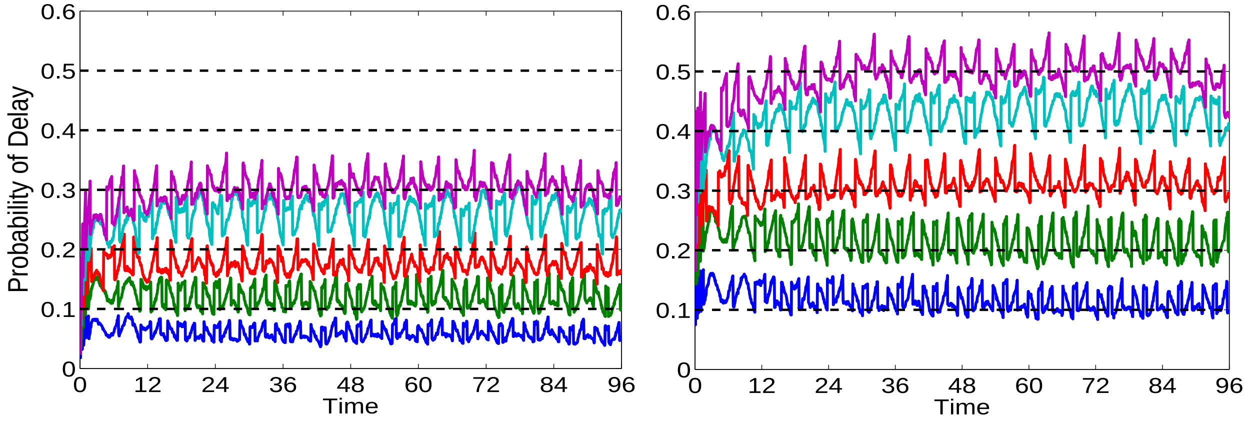


Figure 14: Estimated time-varying probability of delay in the base $H_2^t(1,4)/LN(1,4)/s_t$ model with the same targets as before, but with the average arrival rate $\bar{\lambda}$ reduced from 100 to 4, using the MOL SRS formula $s(t)$ (left) and $s(t) - 1$ (right).

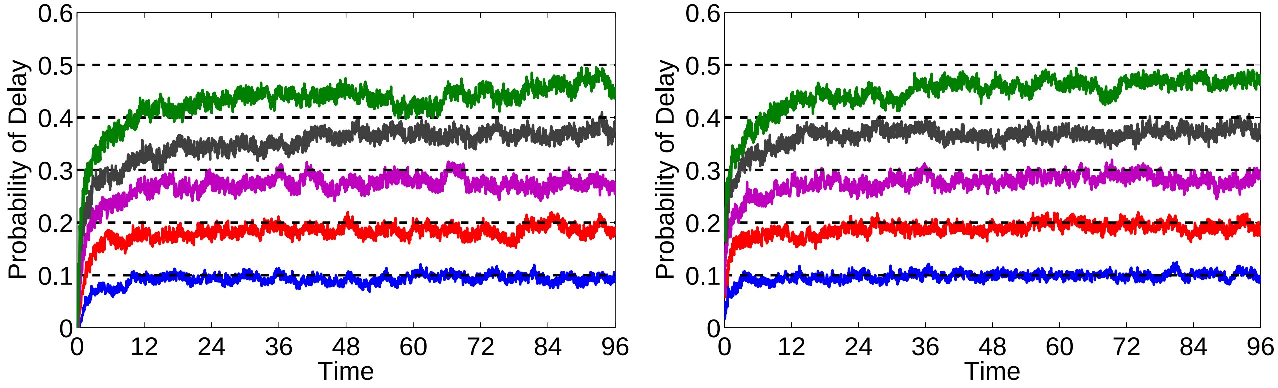


Figure 15: Estimated time-varying probability of delay for the $D^t/LN(1,4)/s_t$ model with $z = 0.64$ (left) and the $E_2^t/LN(1,4)/s_t$ model with $z = 0.82$ (right) using the MOL SRS formula in (1) of the main paper for five delay probability targets α.

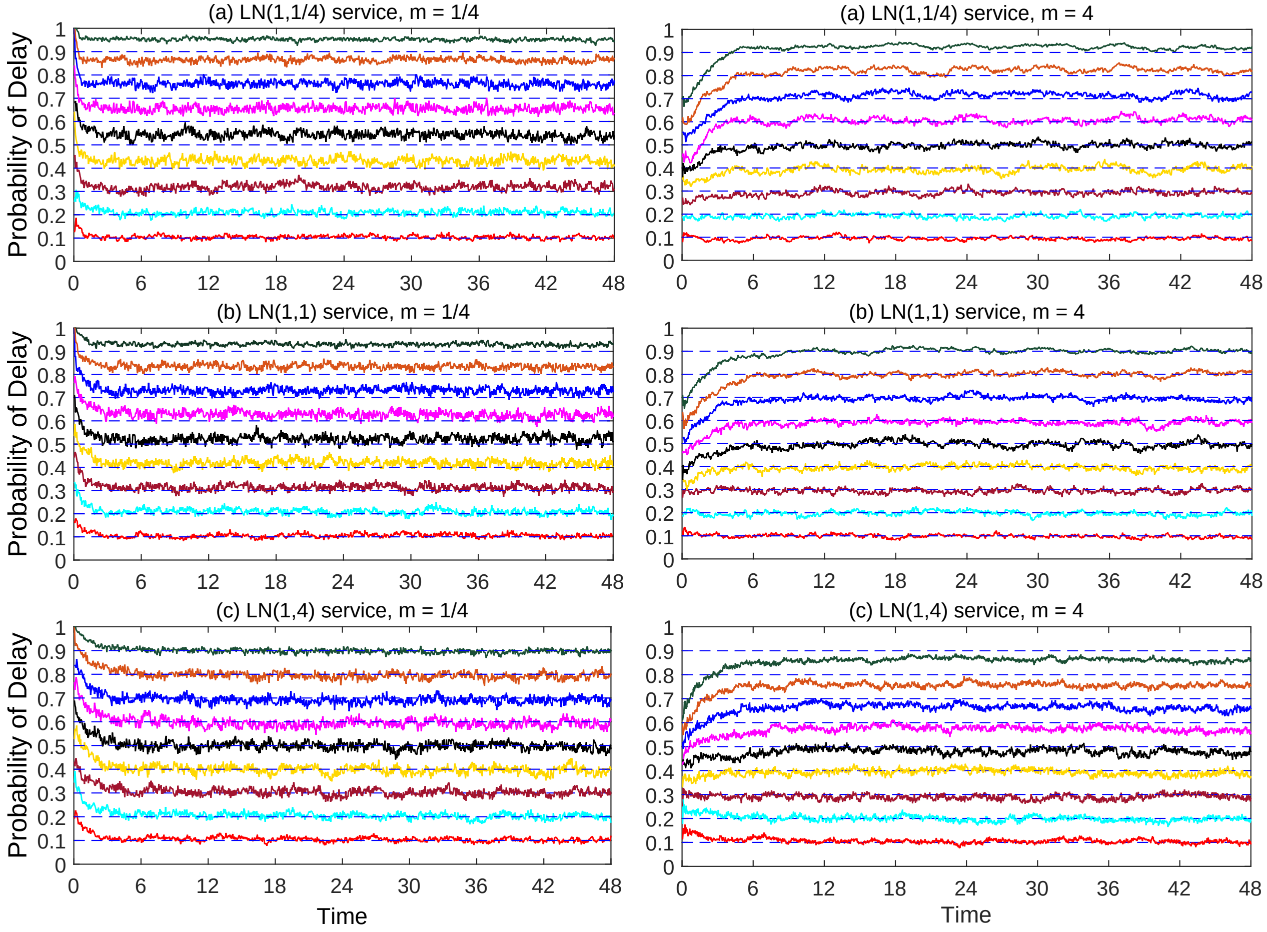


Figure 16: Estimated time-varying probability of delay for the $H_2^t(1,4)/LN(1,s)/s_t + M(m)$ model with $LN(1,v)$ lognormal service times for $v = 0.25$, 1.0 and 4.0 and exponential abandonment with mean $m = 1/4$ (left) and $m = 4$ (right) using the MOL SRS formula (1) and the Garnett function in (41) of the main paper for nine delay probability targets α, ranging from 0.1 to 0.9.

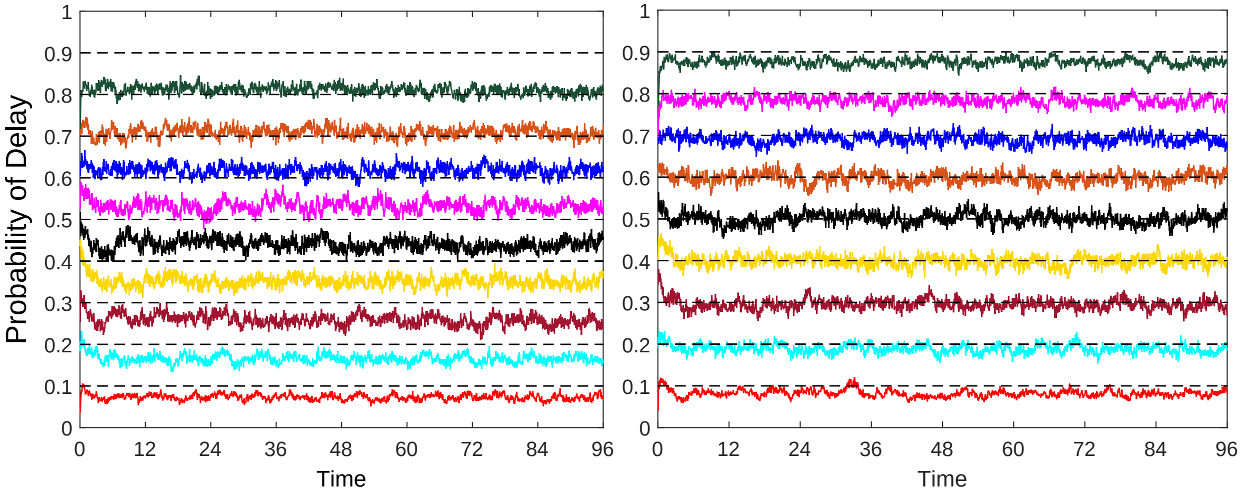


Figure 17: Estimated time-varying probability of delay for the $H_2^t(1,4)/M(1)/s_t + H_2(1,4)$ model with $M(1)$ exponential service times and i.i.d. $H_2(1,4)$ patience times, yielding $\theta = \mu = 1$, using the MOL SRS formula (1), and (a) the Garnett function (left) and (b) the Zeltyn-Mandelbaum (2005) refinement to the Garnett function (right), in (41) of the main paper for nine delay probability targets α (left) and (ii) , ranging from 0.1 to 0.9.

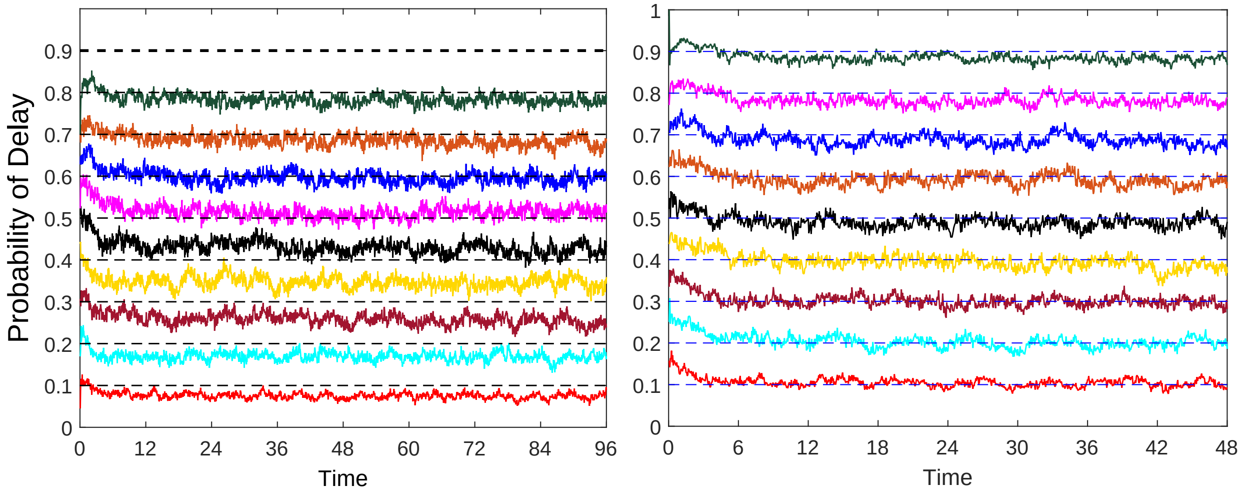


Figure 18: Estimated time-varying probability of delay for the $H_2^t(1,4)/LN(1,4)/s_t + H_2(1,4)$ model with $LN(1,4)$ lognormal service times as in the base case, yielding $z = 2.11$, and i.i.d. $H_2(1,4)$ patience times, yielding $\theta = \mu = 1$, using the MOL SRS formula (1), (a) the Garnett function (left) and (b) the Zeltyn-Mandelbaum (2005) refinement to the Garnett function (right), in (41) of the main paper for nine delay probability targets α, ranging from 0.1 to 0.9.

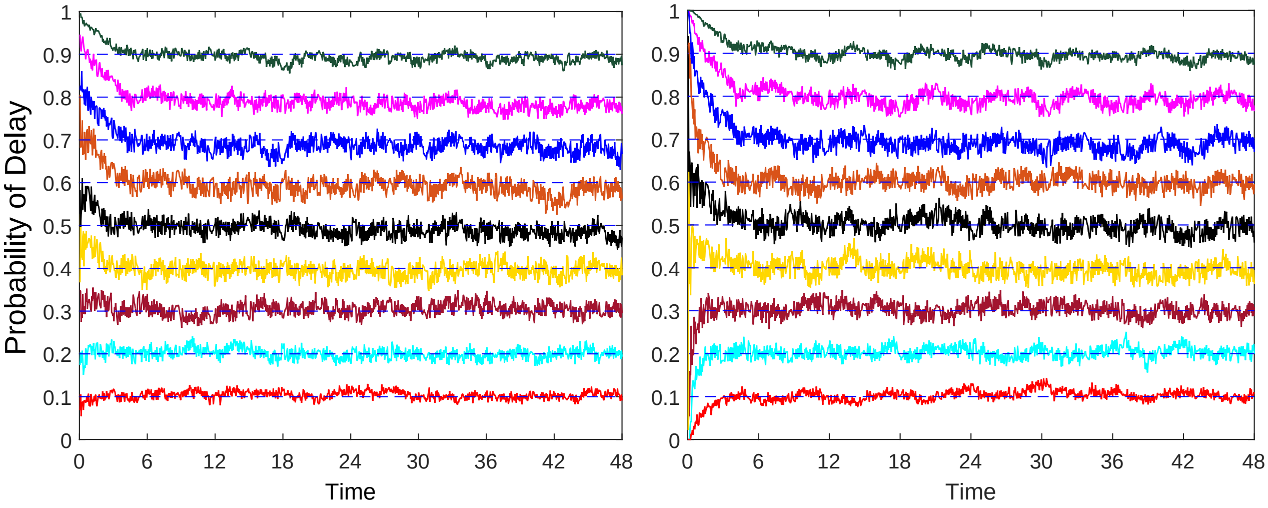


Figure 19: Estimated time-varying probability of delay for the $G_t/M(1)/s_t + H_2(1,4)$ model with low arrival variabilities: (i) E_2^t arrivals (left) and (ii) D^t arrivals (right), $M(1)$ exponential service times and i.i.d. $H_2(1,4)$ patience times, yielding $\theta = \mu = 1$, using the MOL SRS formula (1) and the Zeltyn-Mandelbaum (2005) refinement to the Garnett function in (41) of the main paper for nine delay probability targets α, ranging from 0.1 to 0.9.