http://pubsonline.informs.org/journal/opre/

OPERATIONS RESEARCH
Articles in Advance, pp. 1–21
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Staffing to Stabilize the Tail Probability of Delay in Service Systems with Time-Varying Demand

Yunan Liu[a]

[a] Department of Industrial and Systems Engineering, North Carolina State University, Raleigh, North Carolina 27695
Contact: yliu48@ncsu.edu, http://orcid.org/0000-0001-9961-2610 (YL)

Received: December 20, 2016
Revised: April 14, 2017
Accepted: July 27, 2017
Published Online in Articles in Advance: February 8, 2018

Subject Classifications: probability: stochastic model applications; queues: applications
Area of Review: Stochastic Models.

https://doi.org/10.1287/opre.2017.1678

Abstract. Analytic formulas are developed to set the time-dependent number of servers to stabilize the tail probability of customer waiting times for the $G_t/GI/s_t + GI$ queueing model, which has a nonstationary non-Poisson arrival process (the G_t), nonexponential service times (the first GI), and allows customer abandonment according to a nonexponential patience distribution (the $+GI$). Specifically, for any delay target $w > 0$ and probability target $\alpha \in (0, 1)$, we determine appropriate staffing levels (the s_t) so that the time-varying probability that the waiting time exceeds a maximum acceptable value w is stabilized at α at all times. In addition, effective approximating formulas are provided for other important performance functions such as the probabilities of delay and abandonment, and the means of delay and queue length. Many-server heavy-traffic limit theorems in the efficiency-driven regime are developed to show that (i) the proposed staffing function achieves the goal asymptotically as the scale increases, and (ii) the proposed approximating formulas for other performance measures are asymptotically accurate as the scale increases. Extensive simulations show that both the staffing functions and the performance approximations are effective, even for smaller systems having an average of three servers.

Funding: This research was supported by National Science Foundation [Grant CMMI 1362310].
Supplemental Material: The e-companion is available at https://doi.org/10.1287/opre.2017.1678.

Keywords: staffing algorithms • service systems • capacity planning • many-server queues • efficiency-driven • time-varying arrivals • queues with abandonment • nonstationary queues • nonexponential distributions

1. Introduction

Queueing systems have been widely adopted to model and analyze service systems, such as healthcare systems and customer contact centers (Armony et al. 2014, Brown et al. 2005). Because the demands in service systems typically vary significantly over time, an important issue is how to efficiently allocate critical resources such as staffing (represented as the number of servers) to achieve required targets for the desired performance. In a hospital setting, this can be interpreted as staffing the number of doctors and nurses to ensure that the patients' waiting times before treatments are below certain targets (e.g., six hours). Because inefficient staffing policies can cause excessive suffering, low care quality, degradation of treatment outcomes, and significant mortality increase (Cram et al. 2004, Donahue 2013, Stanton 2004), there is a growing need for developing theories and methods to help improve staffing efficiency.

We develop new formulas to generate appropriate staffing recommendations that would stabilize performance in multiserver queueing systems with time-varying demands and realistic model features. First, abandonment by waiting customers, which corresponds to patients leaving without being seen by a care provider, or to callers hanging up in a call center, can significantly alter system performance (Zeltyn and Madelbaum 2005, Yom-Tov and Mandelbaum 2014). Second, we include the challenging case of the relatively long service times, which commonly occurs in healthcare systems but has been proven difficult to treat (Green et al. 2007), because the influence of the time variability of demands is extended significantly after their arrival times by the long service durations. Next, empirical studies show that neither service time nor abandonment time is exponentially distributed (Armony et al. 2014, Brown et al. 2005, Shi et al. 2016), which motivates us to build general models beyond the conventional Erlang models that have convenient Markovian probability structure.

In particular, we consider a $G_t/GI/s_t + GI$ model with a nonstationary arrival process (the G_t), *independent and identically distributed* (i.i.d.) nonexponential service times (the GI), time-varying staffing levels (the s_t, which is to be determined), and customer abandonment with i.i.d. nonexponential abandonment times (the $+GI$). *We will develop an analytic staffing formula* $s(t)$ *to stabilize the tail probability of delay* (*TPoD*), *the*

probability that the waiting time exceeds a delay target (i.e., a maximum acceptable value), at a desired constant probability target. More precisely, for given quality-of-service (QoS) targets: (i) delay $w > 0$ and (ii) probability $\alpha \in (0,1)$, our staffing function aims to achieve a time-stable TPoD:

$$p(t,w) \equiv P(V(t) > w) \approx \alpha, \quad \text{for } 0 \leqslant t \leqslant T, \qquad (1)$$

where T is a finite time (e.g., a day or a week), and $V(t)$ is the time-dependent *offered waiting time*, that is the delay of an arrival at time t assuming this arrival is infinitely patient.

Applied Relevance of the Tail Probability of Delay. To emphasize that TPoD is an important performance indicator in practice, we next give evidence arising from call centers and healthcare.

(1) *The 80-20 rule in call centers.* A common *telephone service factor* (TSF) in call centers is the well-known 80-20 rule, meaning 80% of calls should be answered in 20 seconds, which is equivalent to 20% of calls (i.e., $\alpha = 0.2$) having to wait for more than 20 seconds (i.e., $w = 1/3$ minute). See Aksin et al. (2007), Brown et al. (2005), Gans et al. (2003) for more background.

(2) *The six-hour service level in Singapore hospitals.* TPoD is also an important service metric in healthcare. Shi et al. (2016) showed that it is an important problem to keep delays below $w = 6$ hours in the impatient departments of Singapore hospitals; their empirical studies show that under inefficient staffing, the probability the delay is below six hours, dubbed "six-hour service level," varies significantly over time during a day from $\alpha = 4\%$ to 37% in a Singapore hospital; see Shi et al. (2016, figure 1(b)).

(3) *Canadian triage and acuity scale.* Patients in Canadian emergency departments are classified to 5 levels, with 1 representing the sickest and 5 the least sick. According to the Canadian triage and acuity scale (CTAS) guideline (Bullard et al. 2014, p. 1), "CTAS level i patients need to be seen by a physician within w_i minutes $100\alpha_i$% of the time," with

$$(w_1, w_2, w_3, w_4, w_5) = (0, 15, 30, 60, 120) \text{ mins},$$
$$(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5) = (0.98, 0.95, 0.9, 0.85, 0.8).$$

The CTAS standard can be directly translated to our TPoD performance with delay target w_i and probability target α_i for level-i patients, $i = 1, \ldots, 5$.

Although important, TPoD has not yet been seriously treated in service systems with time-varying arrival rates. We next review related literature on performance stabilization via optimal staffing levels.

Related literature. Pointwise stationary approximation (PSA) has been proven useful in staffing systems with shorter service times, which follows the basic idea of approximating a nonstationary model at each time by a stationary model; see Green et al. (2007) for a review. The *modified-offered-load* (MOL) approximation has been developed to design staffing functions to control performance functions including the *probability of delay* (PoD), mean waiting time, and *probability of abandonment* (PoA). A key step of MOL is to study the corresponding infinite-server queue and to compute its offered-load function (that is the total service resource needed if there were no constraint on the capacity), see He et al. (2016), Jennings et al. (1996), Yom-Tov and Mandelbaum (2014), Li et al. (2016), Liu and Whitt (2012a, 2014b, 2017), Whitt and Zhao (2017). Feldman et al. (2008), Defraeye and van Nieuwenhuyse (2013) developed a simulation-based *iterative staffing algorithm* (ISA). The basic idea of ISA includes three steps: Step 1, ISA starts with a candidate staffing function; Step 2, under the candidate staffing, ISA estimates the performance (e.g., PoD) via Monte Carlo simulations; Step 3, based on the simulation results, ISA makes adjustments to the candidate staffing and return to Step 2, until the performance target is met. Because many independent sample paths (such as 5,000) have to be generated in each iteration to produce satisfying statistical estimates, a major disadvantage of ISA is its computational expensiveness. Furthermore, ISA does not provide insightful explicit staffing formulas because it is generated by simulations. Except for the simulation-based algorithm in Defraeye and van Nieuwenhuyse (2013), none of the above methods treats TPoD. This motivates us to develop an analytic formula-based staffing method. (We will show that our new analytic staffing recommendation can treat the challenging realistic example in Defraeye and van Nieuwenhuyse 2013 without needing simulations; see Section 4.)

Challenges of treating TPoD. Controlling PoA is equivalent to controlling the mean delay $E[V(t)]$; see Liu and Whitt (2012a). Comparing to PoA and mean delay, TPoD is a much more flexible QoS metric because the manager will be able to choose two desired QoS parameters, w and α, where w indicates the desired delay target and α measures the tolerance level of excessive delay. For instance, to carefully classify the 5 patient acuity levels, CTAS gives 5 distinct values of (w_i, α_i), for $i = 1, \ldots, 5$.

Two main factors contribute to the complications of treating TPoD (versus other measures such as the mean waiting time and PoD). First, controlling TPoD requires the understanding of the full distribution of waiting time $V(t)$, which is far more difficult than stabilizing the mean waiting time (because it involves only the first moment information $E[V(t)]$). Second, when PoD $P(V(t) > 0)$ is controlled at a probability $0 < \alpha < 1$, the system is nearly critically loaded and has negligible delay ($V(t) \approx 0$). In this case the system performance is close to that of its corresponding infinite-server model,

which is the basis of previous staffing literature. However, controlling TPoD with a delay target $w > 0$ that is not close to 0 should operate the system in the overloaded regime so it no longer performs closely to the infinite-server model. These two factors motivate us to develop new methodologies that characterize the full waiting time distribution; our new approach is no longer limited to the frameworks of infinite-server model. Based on heavy-traffic Gaussian approximations, we develop a new staffing prescription that is the sum of two staffing terms, dubbed *two-term Gaussian-approximation-based* (TTGA) staffing method.

Our contributions. We summarize our contributions in four directions.

First, we introduce an analytic staffing method TTGA to stabilize the TPoD at desired QoS targets. Unlike simulation-based ISA and the numerical-search-based MOL, TTGA gives simple analytic formulas that are extremely easy to compute and implement. Unlike simulation-based algorithms, TTGA's explicit staffing structure gives useful engineering insights; see Remarks 1, 3, and 5, and Corollaries 1–3. To the best of our knowledge, there exists no analytic formula-based staffing method to stabilize TPoD prior to our work.

Second, we substantiate the remarkable performance of TTGA by establishing a heavy-traffic limit theorem, assuming an exponential service distribution. This theorem guarantees that as the scale increases, TTGA is asymptotically the correct staffing level for stabilizing the TPoD at any desired probability $\alpha \in (0, 1)$, with any delay target $w > 0$.

Third, although the TPoD can be stabilized, the time-varying effect of the arrival process cannot be completely eliminated, because all other performance metrics (e.g., PoD, PoA, means of delay and queue length, and utilization) are still highly time varying. Therefore, there is a need for seeking effective time-varying performance estimators (Feldman et al. 2008, Liu and Whitt 2012a). We provide analytic formulas to approximate these other important performance measures and establish a second heavy-traffic limit theorem showing the asymptotic accuracy of these estimators as the scale increases.

Finally, we conduct extensive simulation experiments to verify the effectiveness of robustness of TTGA. Our examples include various service and patience distributions (e.g., exponential, hyperexponential, and lognormal), small-scale and large-scale systems (e.g., from 3 servers to 1,000 servers), a wide range of probability targets α (e.g., from 0.02 to 0.9), various delay targets (e.g., from $w \approx 0$ to $w = 6$), various arrival rate function (e.g., sinusoidal, quadratic, piecewise linear) and realistic arrival rates estimated by real hospital and call center data. Our numerical experiments show that TTGA is effective in both (i) overloaded systems (i.e., *efficiency-driven* (ED) regime) and (ii) nearly critically loaded systems (i.e., *quality-and-efficiency driven* (QED) regime). In addition, we show that TTGA continues to work effectively in systems with fewer servers (e.g., hospitals); we can treat the challenging case in Defraeye and van Nieuwenhuyse (2013) without using simulations.

An example. Before describing the details of TTGA, we first give an example using the arrival rates estimated from a call center of a U.S. bank, obtained from SEEStat center (SEE Center 2014). To confirm that TTGA can achieve the 80-20 rule, we let TTGA parameters be $w = 20$ seconds and $\alpha = 1\% - 80\% = 0.2$, meaning 80% of calls are answered in 20 seconds. In this example, we consider an $M_t/M/s_t + M$ model with mean service time 2 minutes and mean patience time 4 minutes. Figure 1 plots (i) the call center arrival rate in a day (top panel); (ii) TTGA staffing level (middle panel); and (iii) estimated TPoD by simulations (bottom panel). Detailed experiment description and simulation procedure are given in Section 4. Figure 1 shows that the TPoD is indeed controlled below $\alpha = 0.2$ in throughout a day.

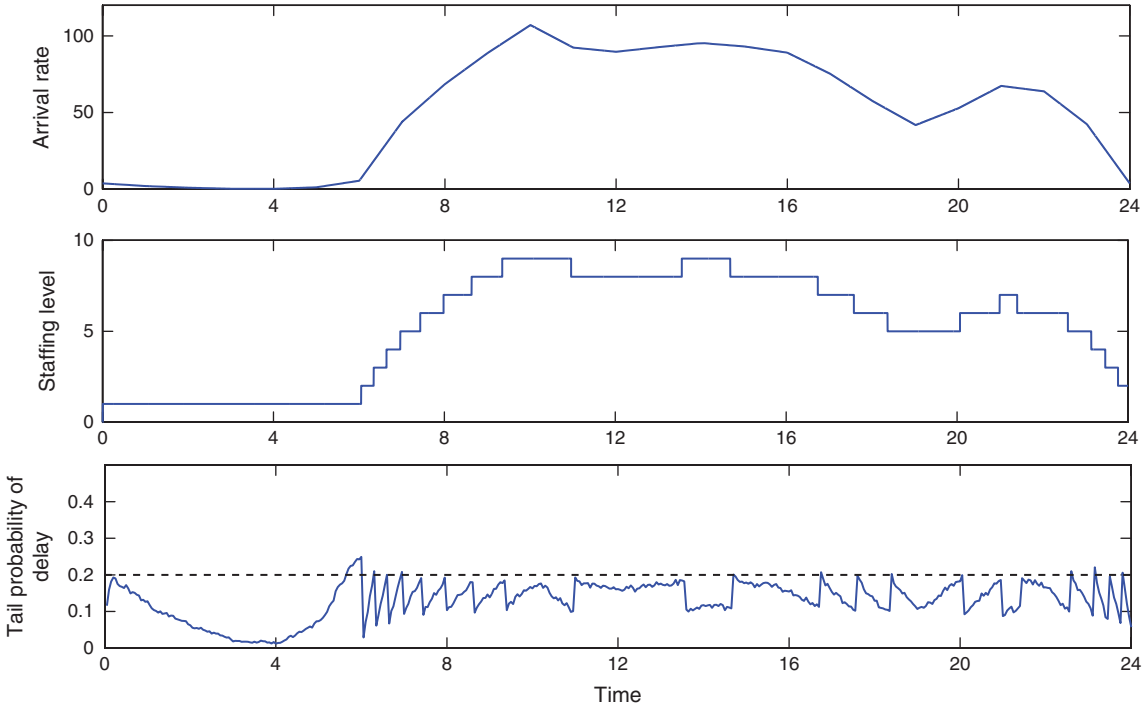
We remark that this is a small-scale model having an average of five servers. The variability of TPoD (bottom plot of Figure 1) is caused by adding and removing a server. More comprehensive examples are given in Section 4. We will show that the TPoD performance becomes more time stable as the system scale increases.

Organization of the Paper. In Section 2 we develop the analytic TTGA staffing formulas as a function of both the model input parameters and the QoS parameters; we give an asymptotic stability theorem to support the effectiveness of TTGA; we also provide various simplified staffing formulas in structural special cases. In Section 3 we provide useful analytic approximating formulas for other important performance measures such as the PoD, PoA, mean waiting time, and mean queue length. In Section 4 we conduct comprehensive numerical experiments to substantiate the effectiveness of TTGA. In Section 5 we prove the asymptotic stability of TTGA and the asymptotic accuracy of the approximating formulas of other performance functions as the scale increases. We draw conclusions in Section 6. Additional supporting materials appear in the e-companion and a longer online appendix (Liu 2017).

2. The TTGA Staffing and Asymptotic Stability

In Section 2.1 we describe the $G_t/GI/s_t + GI$ queueing system. In Section 2.2 we develop the TTGA staffing functions and establish a heavy-traffic limit theorem on the asymptotic stability of TTGA assuming an exponential service distribution. We give staffing formulas in structural special cases in Section 2.3.

Figure 1. (Color online) The QED Performance with the 80-20 Rule: Arrival Rate, TPoDs, and TTGA Staffing Functions for the Model with Real Call-Center Arrival Rate



Note. The QoS targets are $w = 20$ seconds and $\alpha = 20\% = 0.2$.

2.1. The $G_t/GI/s_t + GI$ Model

We consider the $G_t/GI/s_t + GI$ model, with arrival process following a *nonstationary non-Poisson process* (NNPP) with rate $\lambda(t)$, i.i.d. service times following a general *cumulative distribution function* (cdf) $G(x)$, and i.i.d. patience times following a general cdf $F(x)$. Let $\bar{F}(x)$, $f(x)$ and $h_F(x) \equiv f(x)/\bar{F}(x)$ be the *complementary cdf* (ccdf), *probability density function* (pdf), and hazard rate of the patience times. We assume the system is initially empty. Let the average arrival rate in the interval $[0, T]$ be

$$\bar{\lambda} \equiv \frac{1}{T}\int_0^T \lambda(u)\,du. \tag{2}$$

The G_t arrival process. Let $N(t)$ be the total number of arrivals by time $t \geqslant 0$. The Poisson or non-Poisson property of $N(t)$ can be effectively studied with the *index of dispersion for counts* (IDC) $I(t) \equiv \mathrm{Var}(N(t))/\mathrm{E}[N(t)]$. For a *nonhomogeneous Poisson process* (NHPP), $I(t) = 1$ for all t. However, recent studies (Kim and Whitt 2014, Nelson and Gerhardt 2011) have shown that the IDC of arrival processes in service systems differ significantly from 1, which implies that the Poisson assumption can be unrealistic. This motivates us to consider the case of general G_t arrivals.

Following Nelson and Gerhardt (2011), Liu et al. (2018), we consider a G_t arrival model that has two key features: (i) deterministic variability in time characterized by a time-varying arrival-rate function $\lambda(t)$ and (ii) non-Poisson stochastic variability characterized parsimoniously by the single parameter $c_\lambda^2 > 0$. Our NNPP satisfies

$$E[N(t)] = \Lambda(t) \equiv \int_0^t \lambda(s)\,ds \quad \text{and} \quad \lim_{t\to\infty} I(t) = c_\lambda^2 > 0. \tag{3}$$

The reference cases are $c_\lambda^2 = 0$ for a deterministic process and $c_\lambda^2 = 1$ for a Poisson process. Hence, *NHPP will be covered as a special case.* Our NNPP can model a process that is more (less) volatile than NHPP if we let $c_\lambda^2 > 1$ ($c_\lambda^2 < 1$).

A simple way to obtain such an NNPP is to use the composition method. Let $N_0(t)$ be a rate-1 stationary counting process satisfying a *central limit theorem* (CLT): $(N(t) - t)/\sqrt{t} \Rightarrow \mathcal{N}(0, c_\lambda^2)$ as $t \to \infty$, where $\Rightarrow$ denotes convergence distribution and $\mathcal{N}(m, \sigma^2)$ denotes a Gaussian distribution with mean m and variance σ^2. For example, $N_0(t)$ can be a renewal process with interrenewal times having mean 1 and variance c_λ^2. We can construct an NNPP $N(t) \equiv N_0(\Lambda(t))$ that satisfies (3). See Nelson and Gerhardt (2011), Liu et al. (2018) for details.

2.2. The TTGA Staffing Formula

We now present our TTGA staffing formula. The basis of TTGA is the *many-server heavy-traffic* (MSHT) *function weak law of large numbers* (FWLLN) and *functional central limit theorem* (FCLT) for the $G_t/GI/s_t + GI$ queue

(Liu and Whitt 2012b, 2014a). These results show that the properly scaled performance functions (e.g., queue length and waiting times) converge in distribution to convenient Brownian motion based limiting processes as the scale increases. See Liu and Whitt (2014a). Taking the offered waiting time $V(t)$ for an example: When the system is in large scale, the FCLT supports the following Gaussian approximation:

$$V(t) \approx \mathcal{N}(w_s(t), \sigma_s^2(t)), \quad \mathrm{E}[V(t)] \approx w_s(t) \quad \text{and} \quad \mathrm{Var}(V(t)) \approx \sigma_s^2(t), \tag{4}$$

where the approximating mean and variance $w_s(t)$ and $\sigma_s^2(t)$ are both functions of the staffing function $s(t)$ (denoted by the subscript "s") while assuming other model parameters $\lambda(t)$, F and G are fixed. Detailed formulas of $w_s(t)$ and $\sigma_s^2(t)$ are given in Liu and Whitt (2014a).

The first staffing term of TTGA. To stabilize TPoD for given QoS targets w and α, our first step is to choose nominal staffing levels to center $V(t)$ around the delay target w (i.e., placing $\mathrm{E}[V(t)]$ close to w). Therefore, we obtain a staffing function $s_w^{(1)}(t)$ by analytically solving the equation $w_s(t) = w$, $t \geqslant 0$. Because the probability target α is not yet included, this first staffing function is a function only of the QoS target w and model parameters $\lambda(t)$, F, and G:

$$s_w^{(1)}(t) \equiv \bar{F}(w) \int_0^{t-w} \lambda(u)(1 - G(t - w - u))\, du. \tag{5}$$

Detailed derivation of (5) is given in Section 5.

If we round $s^{(1)}(t)$ to the closest integer values (e.g., $\lceil s^{(1)}(t) \rceil$) and set the staffing level accordingly, the mean delay $\mathrm{E}[V(t)]$ will be asymptotically stabilized at the delay target w as expected. We next formally state the asymptotic stability result. Its proof is given in the e-companion.

Theorem 1 (Asymptotic Stability of Mean Delays). *Consider a $G_t/GI/s_t + GI$ queueing model having arrival rate $\lambda(t)$ in the interval $[0, T]$ with an average $\bar{\lambda}$, service cdf G and patience cdf F. If the staffing function is $s(t) = \lceil s_w^{(1)}(t) \rceil$ in (5) for $w > 0$, then the mean delay is stabilized at w in the interval $[0, T]$ asymptotically for a large $\bar{\lambda}$, namely,*

$$\sup_{0<t\leqslant T} |\mathrm{E}[V(t)] - w| \to 0, \quad \text{as } \bar{\lambda} \to \infty,\ w > 0,\ T > 0.$$

We remark that our first staffing term $s^{(1)}(t)$ in (5) coincides with the offered-load function of the *delayed infinite-server* (DIS) approximation (Liu and Whitt 2012a).

Unfortunately, $s^{(1)}(t)$ is not effective for stabilizing TPoD, because $s^{(1)}(t)$ does not include the probability target α. Indeed, under $s^{(1)}(t)$, the mean waiting time $\mathrm{E}[V(t)] \approx w$ so that $V(t) \approx \mathcal{N}(w, \sigma_s^2(t))$, which implies that the TPoD $P(V(t) > w) \approx P(\mathcal{N}(w, \sigma_s^2(t)) > w) = P(\mathcal{N}(0, \sigma_s^2(t)) > 0) = 0.5$. So the TPoD should be close to 0.5 if we directly staff according to $s_w^{(1)}(t)$; $s^{(1)}(t)$ is too crude for a desired α that is not close to 0.5 (e.g., $\alpha = 0.2$ for the 80-20 rule).

Our strategy is to refine (5) by adding a second-order staffing term that is a function of α. We envision a staffing function consisting of two parts, in particular,

$$s_{w,\alpha}(t) = s_w^{(1)}(t) + s_{w,\alpha}^{(2)}(t), \tag{6}$$

where $s_w^{(1)}(t)$ is given in (5) (a function of w), and $s_{w,\alpha}^{(2)}$ is some secondary staffing term (a function of both w and α) that gives finer adjustments.

The secondary staffing term $s_{w,\alpha}^{(2)}(t)$. The delay target $w > 0$ can be understood as the "first-order" QoS target because it roughly determines the average (principle) behavior of the random delay $V(t)$, while the probability $0 < \alpha < 1$ can be understood as the "second-order" QoS target because it measures the variability of $V(t)$ around its mean value (which is of a smaller order). Since the first staffing term in (6) controls the w and the second term in (6) controls α, we expect $s_{w,\alpha}^{(2)}(t)$ to be a smaller order term compared to $s_w^{(1)}(t)$. The second term $s_{w,\alpha}^{(2)}(t)$ should slightly tilt $\mathrm{E}[V(t)]$ from w to some $w_{s^{(2)}}$ (a function of $s_{w,\alpha}^{(2)}(t)$), making it possible to shift the TPoD to any $\alpha \in (0, 1)$. Under TTGA (6), we have the following approximation for TPoD:

$$\begin{aligned} &P(V(t) > w) \\ &\quad \approx P(\mathcal{N}(w_{s^{(2)}}, \sigma^2_{s^{(1)}+s^{(2)}}(t)) > w) \\ &\quad = P\left(\mathcal{N}(0,1) > \frac{w - w_{s^{(2)}}}{\sigma_{s^{(1)}+s^{(2)}}(t)}\right) = 1 - \Phi\left(\frac{w - w_{s^{(2)}}}{\sigma_{s^{(1)}+s^{(2)}}(t)}\right), \end{aligned}$$

where $\Phi(\cdot)$ is the cdf of an $\mathcal{N}(0,1)$ Gaussian distribution. Matching the right-hand side of TPoD to the desired probability target $\alpha \in (0, 1)$, we can solve for the desired second term $s_{w,\alpha}^{(2)}(t)$ in (6). The detailed derivation is given in the proof of Theorem 2.

As a function of $(s_{w,\alpha}^{(1)}, c_\lambda^2, F, G, w, \alpha)$, we now provide the analytic form of $s_{w,\alpha}^{(2)}(t)$:

$$s_{w,\alpha}^{(2)}(t) \equiv z_\alpha e^{-\mu t} \left(Z(t) - (\mu - h_F(w)) \int_w^{t \vee w} Z(u)\, du \right), \tag{7}$$

where

$$Z(t) \equiv e^{(\mu - h_F(w))t} \cdot \sqrt{\int_w^{t \vee w} e^{2h_F(w)x} (\hat{C}^2 (\mu s_w^{(1)}(x) + \dot{s}_w^{(1)}(x)) - \dot{s}_w^{(1)}(x))\, dx}, \tag{8}$$

$\hat{C} \equiv \sqrt{(c_\lambda^2 - 1)F^c(w) + 1 + c_s^2}$, $c_s^2 \equiv \mathrm{Var}(S)/\mathrm{E}[S]^2$ is the squared coefficient of variation (SCV) of a generic service time S, z_α is the α-percentile of the standard

Gaussian random variable $\mathcal{N}(0,1)$, namely, $z_\alpha \equiv \Phi^{-1} \cdot (1-\alpha)$, $x \vee y \equiv \max(x,y)$, and $\dot{u}(x) \equiv du(x)/dx$ is the derivative of the function $u(x)$. Since both $s_w^{(1)}$ and $s_{w,\alpha}^{(2)}$ are continuous functions, we can simply round the formula (6) (e.g., $\lceil s_{w,\alpha} \rceil$) to set the actual integer-valued staffing levels in practice (this is what we do in our simulation examples).

Remark 1 (Understanding the Second Term $s_{w,\alpha}^{(2)}(t)$). We provide some intuitive explanations for the seemingly complicated term $s_{w,\alpha}^{(2)}(t)$.

(i) *Sign of* $s_{w,\alpha}^{(2)}$. The TTGA formula (6) simply reduces to $s_w^{(1)}(t)$ when $\alpha = 0.5$, because $s_{w,0.5}^{(2)} = z_{0.5} = 0$. This agrees with our intuition because directly staffing according to (5) should already stabilize the TPoD at 0.5. Since the impact of α factors out in the term z_α, the second term $s_{w,\alpha}^{(2)} > 0$ (< 0) when $\alpha < 0.5$ (> 0.5) (because $z_\alpha > 0$ (< 0)). This is also consistent with our intuition that the first staffing term $s_w^{(1)}$ (5) understaffs (overstaffs) the system when the desired target $\alpha < 0.5$ (> 0.5). So we rely on the secondary term $s_{w,\alpha}^{(2)}$ to add (remove) an appropriate "second-order" number of servers, setting the TPoD at a desired $\alpha \in (0,1)$.

(ii) *Magnitude of* $s_{w,\alpha}^{(2)}$. The analytic formulas (7)–(8) also help estimate the magnitude of $s_{w,\alpha}^{(2)}$. It is easy to see that $s_{w,\alpha}^{(2)} = O(\sqrt{s_w^{(1)}})$, where we say $f = O(g)$ if $\sup_{t \geqslant 0} |f(t)/g(t)| \leqslant C$ for some $C > 0$. This observation justifies why we refer to $s_w^{(1)}$ and $s_{w,\alpha}^{(2)}$ as the "first-order" and "second-order" staffing levels and w and α as the "first-order" and "second-order" QoS targets. See Figure 4 for a comparison of magnitudes of $s_w^{(1)}$ and $s_{w,\alpha}^{(2)}$. To provide more insights, we show in Section 2.3 that the TTGA formula in (6) simplifies to the square-root staffing formula in special cases.

(iii) *Dependence on model parameters*. The secondary staffing term $s_{w,\alpha}^{(2)}$ is an analytic function of the QoS targets w and α, the main staffing term $s_w^{(1)}$, and the model parameters F, μ, c_s^2, and c_λ^2. We note that $s_{w,\alpha}^{(2)}$ depends on the service cdf G only via its mean $1/\mu$ and variance c_s^2/μ^2. In addition, $s_{w,\alpha}^{(2)}$ is independent of $\lambda(t)$. In contrast, both the service cdf G and the time-varying arrival rate $\lambda(t)$ are captured by the first-order term $s_w^{(1)}$.

We next state a heavy-traffic limit theorem for the asymptotic stability of the TTGA formula (6) when the service distribution is exponential (i.e., $G(x) = 1 - e^{-\mu x}$ and $c_s^2 = 1$). We give the proof in Section 5.

Theorem 2 (Asymptotic Stability of the TTGA Formula (6) for Stabilizing TPoD). *Consider a $G_t/M/s_t + GI$ queueing model having arrival rate $\lambda(t)$ in the interval $[0,T]$ with an average $\bar{\lambda} \equiv \int_0^T \lambda(t)\,dt/T$, exponential service times, and general patience cdf F. If the staffing function is given in (6) with $c_s^2 = 1$, then for all $w > 0$, $0 < \alpha < 1$, the TPoD is stabilized at α in the interval $[0,T]$ asymptotically for a large $\bar{\lambda}$, namely,*

$$\sup_{0<t\leqslant T} |P(V(t) > w) - \alpha| \to 0, \quad \text{as } \bar{\lambda} \to \infty.$$

Remark 2 (TTGA for Nonexponential Service Times). A rigorous heavy-traffic limit theorem for the asymptotic stability of TTGA with GI service requires an associated FCLT for the $G_t/GI/s_t + GI$ model. But such a limit theorem currently remains an open problem. The good news is that according to Theorem 1, we know that our first-order staffing term $s_w^{(1)}$ stabilizes the mean delay when the service distribution is *nonexponential*. The second-order term $s_{w,\alpha}^{(2)}$ in (7) for $c_s^2 \neq 1$ is a heuristic refinement. The basis of this refinement is to use a renewal process to approximate the departure process when the service cdf is nonexponential. We are able to include c_s^2 as a factor in $s_{w,\alpha}^{(2)}$ because the FCLT of properly scaled renewal processes is a Brownian motion multiplied by the SCV of the interrenewal time; see Whitt (2002). We substantiate the effectiveness of the staffing formula for GI service (with $c_s^2 \neq 1$) by conducting simulation experiments with nonexponential service times (e.g., lognormal and H_2 service distributions) in Section 4. We provide more detailed explanations of this heuristic refinement in the online appendix.

2.3. Staffing Functions in Structural Special Cases

In this section, we simplify the TTGA staffing function in (6) by imposing special model assumptions. In particular, we consider three cases: NHPP arrival process, constant arrival rate, and sinusoidal arrival rate. We formulate the simplified staffing formulas through the following corollaries and give the proofs in the e-companion.

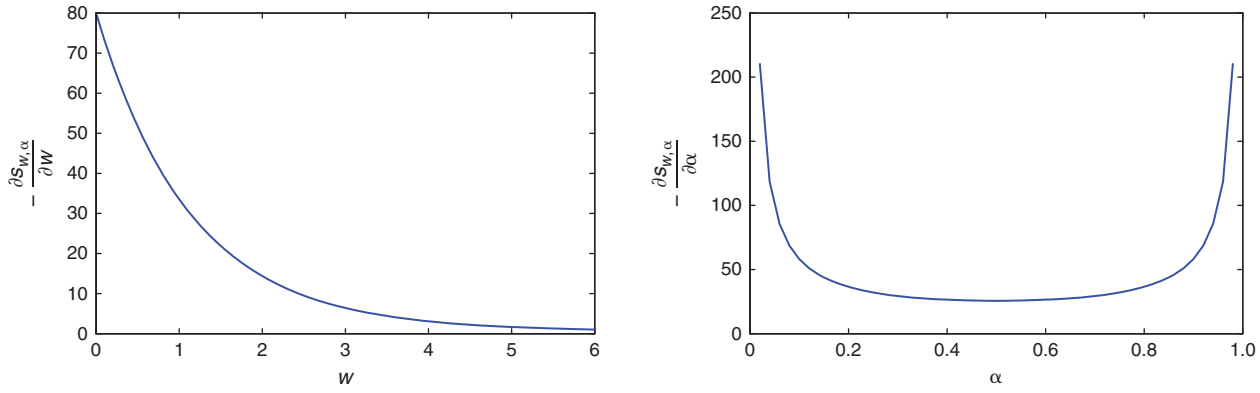
Corollary 1 (TTGA Staffing for the $G/M/s_t + GI$ Model). *If the arrival rate is a constant $\bar{\lambda}$ and service times are exponentially distributed, then staffing formula in (6) is asymptotically a classical square-root-staffing (SRS) function, in particular, as $t \to \infty$,*

$$s_{w,\alpha}(t) \sim s_w^{(1)} + \beta_{w,\alpha}\sqrt{s_w^{(1)}}, \quad \text{where } s_w^{(1)} \equiv \bar{F}(w)\frac{\lambda}{\mu}, \qquad \beta_{w,\alpha} \equiv z_\alpha \sqrt{\frac{[(c_\lambda^2 - 1)\bar{F}(w) + 2]h_F(w)}{2\mu}}, \tag{9}$$

and we say $f \sim g$ if $f(x)/g(x) \to 1$ as $x \to \infty$.

Remark 3 (Average Number of Servers and Marginal Price of Staffing). The above SRS structure can help estimate the required average number of servers. The QoS coefficient $\beta_{w,\alpha}$ is a function of both w and α; it is increasing in the arrival CSV c_λ^2 and depends on the

Figure 2. (Color online) MPS with Respect to w (with $\alpha = 0.5$, Left) and α (with $w = 0.5$, Right) When the Arrival Rate Is 100 and t Is Large



patience ccdf $\bar{F}$ and hazard rate h_F only at their values at w. In addition, these analytic staffing formulas can help estimate the *marginal price of staffing* (MPS), that is, *to improve the service to a next level* (*e.g., reducing w by* Δw*, or reducing* α *by* $\Delta\alpha$*), how many extra servers are needed*? We can answer this question by computing the partial derivatives of the staffing formula in (9) with respect to w and α. (See the e-companion for analytic formulas of the two partial derivatives.) Let $\bar{\lambda} = 100$, $c_\lambda^2 = 4$, patience cdf be a hyperexponential with mean 2 and variance 8, mean service time $1/\mu = 1$, we plot the two partial derivatives in Figure 2. We observe that the MPS is monotonically decreasing in w and the MPS is high when α is close to 0 or 1 but low when $\alpha \approx 0.5$. For instance, for $\Delta\alpha = 0.1$, we need to add to the staffing function $(-\partial s_{0.5,\alpha}/\partial\alpha|_{\alpha=0.5}) \times \Delta\alpha \approx 30 \times 0.1 = 3$ servers if we hope to reduce α from 0.5 to $0.5 - \Delta\alpha = 0.4$. See Section EC.2.4.1 of the e-companion for more discussions of MPS.

Corollary 2 (TTGA Staffing for the $M_t/M/s_t + GI$ Model). *If the arrival process is an NHPP, and* $\mu = h_F(w)$*, then*

$$s_{w,\alpha}(t) = s_w^{(1)}(t) + \beta_{w,\alpha}\sqrt{s_w^{(1)}(t)}, \quad \text{where } \beta_{w,\alpha} \equiv z_\alpha. \tag{10}$$

Remark 4 (Connection with the Garnett Function in Garnett et al. (2002)). The assumption of an NHPP arrival process implies a time-varying SRS staffing function. If, in addition, the patience distribution is exponential with rate $\theta = \mu$, then our staffing formula (10) agrees with the Garnett SRS formula in Garnett et al. (2002). In this case, our TTGA staffing function reduces to the SRS staffing function for the full Markovian $M_t/M/s_t + M$ model studied in Feldman et al. (2008), when $w = 0$. Of course, we hereby aim at stabilizing the TPoD, which generalizes the PoD by introducing a new QoS target w. Nevertheless, the TPoD reduces to the simple PoD when $w \approx 0$.

Because sinusoidal functions capture the periodic structure in realistic arrival patterns (see Feldman et al. 2008, Yom-Tov and Mandelbaum 2014, Liu and Whitt 2012a), we next consider a sinusoidal arrival rate

$$\lambda(t) = \bar{\lambda}(1 + r\sin(\gamma t + \phi)), \tag{11}$$

with average rate $\bar{\lambda}$, relative amplitude $|r| < 1$, frequency γ, and phase ϕ. We give structural results below for the corresponding TTGA staffing function.

Corollary 3 (TTGA staffing for sinusoidal arrival rates). *If the arrival-rate is sinusoidal as in* (11)*, then*

$$s_w^{(1)}(t) \sim \frac{\bar{\lambda}}{\mu}\bar{F}(w) + \frac{r\bar{\lambda}\bar{F}(w)}{\sqrt{\mu^2+\gamma^2}}\sin(\gamma(t+\phi/\gamma - w) - \varphi)$$

$$s_{w,\alpha}^{(2)}(t) \sim \frac{z_\alpha f(w)}{\mu\sqrt{\bar{F}(w)}}\big(\bar{a} + \bar{b}_1\sin(\gamma(t+\phi/\gamma - w) - \eta) + \bar{b}_2\sin(\gamma(t+\phi/\gamma - w) - \varphi)\big)^{1/2},$$

where

$$\bar{a} \equiv \frac{\bar{\lambda}C}{2h_F(w)}, \ \bar{b}_1 \equiv \frac{r\bar{\lambda}(\mu C - 2(C-1)h_F(w))}{\sqrt{(\mu^2+\gamma^2)(4h_F^2(w)+\gamma^2)}},$$

$$\bar{b}_2 \equiv \frac{r\bar{\lambda}(C-1)}{\sqrt{\mu^2+\gamma^2}} \ C \equiv (c_\lambda^2 - 1)\bar{F}(w) + 2,$$

$$\varphi \equiv \arctan(\gamma/\mu) \quad \text{and} \quad \eta \equiv \varphi + \arctan(\gamma/(2h_F(w))).$$

Remark 5 (Asymptotic Periodic Structure). Corollary 3 indicates that if the arrival rate is sinusoidal, then the TTGA staffing function exhibits a periodic pattern. The detailed formulas here can be used to estimate useful quantities of the periodic function, such as the period, oscillation magnitude, and their sensitivities to the model parameters.

3. Approximating Other Performance Functions Under TTGA

Although the TPoD is stabilized at a desired target once the queueing model is staffed according to TTGA,

other important performance measures still remain highly time varying. Indeed, Feldman et al. (2008), Liu and Whitt (2012a) show that the time-varying effect brought by the nonstationary arrivals cannot be completely eliminated by any staffing method. For the purpose of performance approximation and forecasting, we next provide analytic formulas to approximate performance functions including the PoD $p_{de}(t)$, PoA $p_{ab}(t)$, mean delay $\mathrm{E}[W(t)]$, mean queue length $\mathrm{E}[Q(t)]$, and service utilization $u(t)$ (the ratio of the mean number of busy servers and the total number of servers). These approximating formulas are simple and analytic functions of the model parameters $(\lambda, c_\lambda^2, G, F)$ and the QoS parameters (w, α).

Let $Q(t)$ and $B(t)$ be the number of customers waiting in queue and the number of busy servers at time t. Let $X(t) \equiv Q(t) + B(t)$ be the total number of customers in the system at t. We will show in Section 5 that both $V(t)$ and $X(t)$ are approximately distributed as Gaussian random variables for a large $\bar{\lambda}$. In particular, under the TTGA staffing, we have

$$V(t) \approx \mathcal{N}(w - z_\alpha \sigma_V^*(t), {\sigma_V^*}^2(t)) \quad \text{and}$$
$$X(t) \approx \mathcal{N}(X^*(t), {\sigma_X^*}^2(t)),$$

where

$$\begin{aligned}
{\sigma_V^*}^2(t) &= \frac{e^{-2\mu t} Z^2(t)}{(\lambda(t-w)\bar{F}(w))^2} \\
X^*(t) &= \int_{t-w}^{t} \lambda(u)\bar{G}(t-u)\,du \\
&\quad - z_\alpha \lambda(t-w)\bar{F}(w)\sigma_V^*(t) + s_{w,\alpha}(t) \\
{\sigma_X^*}^2(t) &= \int_{t-w}^{t} \lambda(u)\bar{F}(t-u)[(c_\lambda^2 - 1)\bar{F}(t-u) + 1]\,du \\
&\quad + e^{-2\mu t} Z^2(t),
\end{aligned}$$

and $Z(t)$ is given in (8). See Section 5 for detailed derivations for these Gaussian formulas. Based on the above Gaussian approximations, we propose the performance approximating formulas

$$\begin{aligned}
\mathrm{E}[V(t)] &\approx \tilde{V}(t) \equiv \mathrm{E}\left[\mathcal{N}(w - z_\alpha \sigma_V^*(t), {\sigma_V^*}^2(t))^+\right], \\
p_{ab}(t) &\approx \tilde{p}_{de}(t) \equiv \int_0^\infty \Phi\left(\frac{w-x}{\sigma_V^*(t)} - z_\alpha\right) f(x)\,dx \\
\mathrm{E}[Q(t)] &\approx \tilde{Q}(t) \equiv \mathrm{E}\left[(\mathcal{N}(X^*(t), {\sigma_X^*}^2(t)) - s_{w,\alpha}(t))^+\right], \\
p_{de}(t) &\approx \tilde{p}_{de}(t) \equiv \Phi\left(\frac{w}{\sigma_V^*(t)} - z_\alpha\right), \\
u_n(t) &\approx \tilde{u}(t) \equiv \mathrm{E}\left[\mathcal{N}\left(\frac{X^*(t)}{s_{w,\alpha}(t)}, \frac{{\sigma_X^*}^2(t)}{s_{w,\alpha}^2(t)}\right)^+ \wedge 1\right],
\end{aligned} \tag{12}$$

where $x \wedge y \equiv \min(x, y)$ and $x^+ \equiv \max(x, 0)$. We add the superscript "+" in the approximating formulas because all performance functions have to be nonnegative. The computation of the approximating formulas in (12) involves numerically computing $\mathrm{E}[(a + b\mathcal{Z})^+]$ and its analogs. We remark that these computations are simple and fast because the formulas can be expressed as simple functions of the Gaussian cdf Φ and pdf ϕ. We provide the final explicit forms of (12) in Section EC.5 of the e-companion.

Next we provide a heavy-traffic limit theorem to establish the asymptotic accuracy of the approximating formulas in (12) when the service times are exponential. The proof is given in Section 5.

Theorem 3 (Asymptotic Accuracy of Performance Approximation Formulas). *Consider the $G_t/M/s_t + GI$ queueing model having arrival rate $\lambda(t)$ in the interval $[0, T]$ with an average $\bar{\lambda}$, exponential service times, and general patience cdf F. If the staffing function is given in (6) with $c_s^2 = 1$, the approximating formulas in (12) are asymptotically accurate for a large $\bar{\lambda}$, specifically,*

$$\begin{aligned}
\sup_{0\leqslant t\leqslant T} |\mathrm{E}[V(t)] - \tilde{V}(t)| &\to 0, \\
\sup_{0\leqslant t\leqslant T} |\bar{\lambda}^{-1}(\mathrm{E}[Q(t)] - \tilde{Q}(t))| &\to 0, \\
\sup_{0\leqslant t\leqslant T} |u(t) - \tilde{u}(t)| &\to 0, \\
\sup_{0\leqslant t\leqslant T} |p_{de}(t) - \tilde{p}_{de}(t)| &\to 0, \\
\sup_{0\leqslant t\leqslant T} |p_{ab}(t) - \tilde{p}_{ab}(t)| &\to 0, \quad \text{as } \bar{\lambda} \to \infty.
\end{aligned} \tag{13}$$

4. Numerical Examples

We provide extensive numerical experiments to verify the effectiveness of TTGA. First, we consider an $H_2(t)/M/s_t + H_2$ base example in Section 4.1. Next we give additional simulation experiments in Section 4.2–4.4, including cases with smaller arrival rates and number of servers, various delay targets, and different arrival-rate functions such as quadratic and piecewise constant. In Section 4.2.2 we provide another realistic example with an arrival rate estimated from the emergency department of a Belgian hospital, providing direct comparison with the performance of the simulation-based staffing algorithm in Defraeye and van Nieuwenhuyse (2013). Finally, we present an example allowing nonexponential (lognormal) service times in Section 4.5.

4.1. An $H_2(t)/M/s_t + H_2$ Main Example

We consider an $H_2(t)/M/s_t + H_2$ model, having arrivals according to an NNPP with a sinusoidal rate function given in (11), exponential service cdf $G(x) = 1 - e^{-\mu x}$, and a two-phase hyperexponential (H_2) patience cdf:

$$F(x) = 1 - p_a e^{-\theta_1 x} - (1 - p_a)e^{-\theta_2 x}. \tag{14}$$

Following Nelson and Gerhardt (2011), Liu et al. (2018), we construct an NNPP arrival process $\{N(t), t \geqslant 0\}$ by composing a rate-1 renewal process with a mean-value function $\Lambda(t)$ in (3). To make sure the arrival

process is not nearly an NHPP, we consider an H_2 interrenewal cdf

$$F_\lambda(x) = 1 - p_\lambda e^{-\lambda_1 x} - (1-p_\lambda)e^{-\lambda_2 x}. \quad (15)$$

We let $\mu = \gamma = 1$ (so the mean service time $1/\mu = 1$), $\phi = 0$, $\lambda_1 = 2p_\lambda$, $\lambda_2 = 2(1-p_\lambda)$, $\theta_1 = 2p_a\theta$, $\theta_2 = 2(1-p_a)\theta$, $\theta = 0.5$ (so the mean patience time $1/\theta = 2$), $\bar{\lambda} = 100$ (so the average arrival rate is 100), $r = 0.2$ (so the arrival rate varies from 80 to 120), and $p_\lambda = p_a = (5+\sqrt{15})/10$ such that the SCV (variance divided by the square of the mean) of the interarrival times and patience times are $c_\lambda^2 = c_a^2 \approx 4$. According to Liu et al. (2018), the arrival process has a variance-to-mean ratio $\mathrm{Var}(N(t))/\mathrm{E}[N(t)] \approx 4$ (which is not nearly Poisson).

Our staffing procedure applies to arbitrary arrival-rate functions. Here the choice of a sinusoidal function is convenient because it roughly captures the spirit of real systems having cyclic demand; see Feldman et al. (2008). An important issue for applications is the rate of fluctuation in the arrival-rate function compared to the mean service time. Because a cycle of the arrival-rate function in (11) is 2π, there will be about four cycles during the interval $[0, 24]$ (i.e., during a 24-hour day).

We use simulation experiments to show that TTGA achieves the desired time-stable performances for a wide range of α. Figure 3 shows simulation estimates of key performance measures with a delay target $w = 0.5$ and probability target $\alpha = 0.1, \ldots, 0.9$. Figure 3 shows that all TPoD (solid lines) are stabilized at desired targets α (dashed lines). We show that TTGA performs well with even smaller α (e.g., $0.02 \leqslant \alpha \leqslant 0.08$) in Section 4.3.2. In addition, while TPoD is stabilized,

Figure 3. (Color online) A Simulation Comparison: Estimated Time-Dependent TPoD, PoD, Mean Queue Length for the $H_2(t)/M/s_t + H_2$ Example with a Sinusoidal Arrival Rate $\lambda(t) = 100 + 20\sin(t)$, $w = 0.5$ and $\alpha = 0.1, \ldots, 0.9$

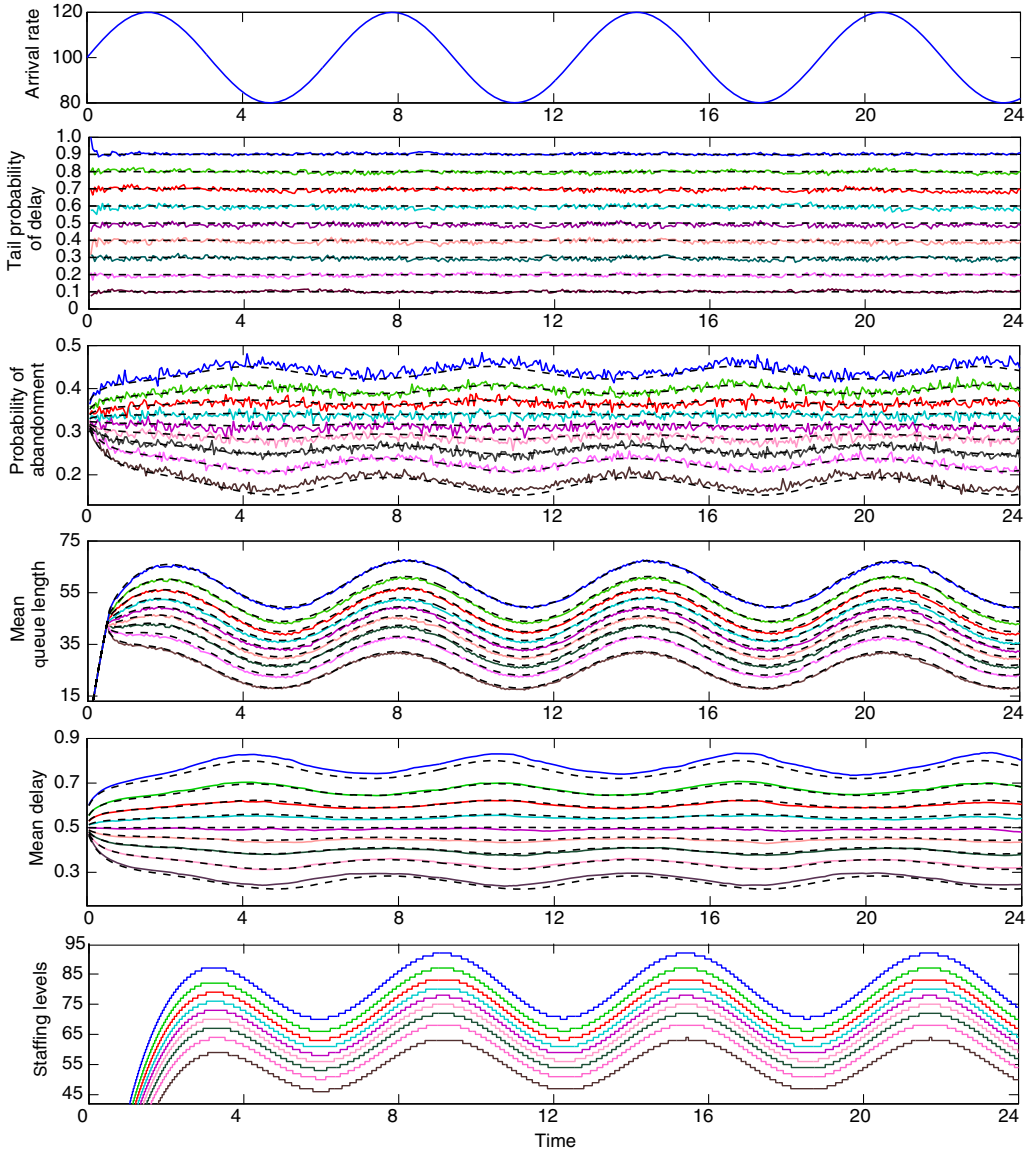


Table 1. Average, Min and Max of TPoD, and Comparison with the TPoD Targets, with $w = 0.5$ and $\lambda(t) = 100 + 20\sin(t)$

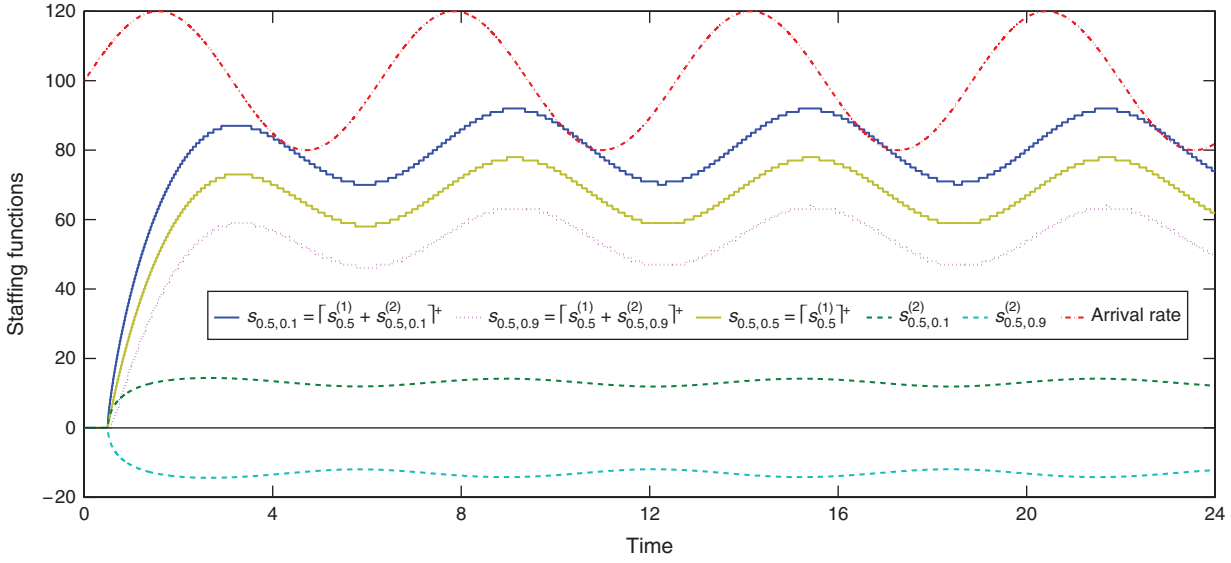
	Avg (diff. to target)		Max (diff. to target)		Min (diff. to target)	
Target	Ceiling	Flooring	Ceiling	Flooring	Ceiling	Flooring
0.9	0.9024 (+0.0024)	0.9114 (+0.0114)	0.9146 (+0.0146)	0.9222 (+0.0222)	0.8900 (−0.0100)	0.8958 (−0.0042)
0.8	0.7985 (−0.0015)	0.8120 (+0.0120)	0.8236 (+0.0236)	0.8274 (+0.0274)	0.7796 (−0.0204)	0.7916 (−0.0084)
0.7	0.6981 (−0.0019)	0.7097 (+0.0097)	0.7252 (+0.0252)	0.7330 (+0.0330)	0.6760 (−0.0240)	0.6866 (−0.0134)
0.6	0.5919 (−0.0081)	0.6103 (+0.0103)	0.6204 (+0.0204)	0.6404 (+0.0404)	0.5698 (−0.0302)	0.5842 (−0.0158)
0.5	0.4937 (−0.0063)	0.5100 (+0.0100)	0.5232 (+0.0232)	0.5422 (+0.0422)	0.4668 (−0.0332)	0.4810 (−0.0190)
0.4	0.3929 (−0.0071)	0.4071 (+0.0071)	0.4198 (+0.0198)	0.4346 (+0.0346)	0.3646 (−0.0354)	0.3788 (−0.0212)
0.3	0.2965 (−0.0035)	0.3110 (+0.0110)	0.3166 (+0.0166)	0.3330 (+0.0330)	0.2728 (−0.0272)	0.2844 (−0.0156)
0.2	0.2007 (+0.0007)	0.2098 (+0.0098)	0.2222 (+0.0222)	0.2318 (+0.0318)	0.1848 (−0.0152)	0.1880 (−0.0120)
0.1	0.1028 (+0.0028)	0.1112 (+0.0112)	0.1168 (+0.0168)	0.1266 (+0.0266)	0.0868 (−0.0132)	0.0956 (−0.0044)

other performance measures (solid lines in subplots 2–5), such as the PoA, mean delay and mean queue length (i.e., mean number of waiting customers) remain time varying; nevertheless, they agree closely with our approximating formulas (dashed lines in subplots 2–5), that are given in Section 3 for the detailed approximating formulas. These simulation estimates are obtained by averaging 5,000 independent sample paths. We give the details of the numerical implementation and computer simulations in Section EC.5 of the e-companion. The bottom panel of Figure 3 shows all nine time-varying staffing levels corresponding to the different values of α. We remark that the highest (lowest) staffing function corresponds to the lowest (highest) performance functions in other subplots.

We further quantify the good performance shown in Figure 3 by computing the minimum, maximum, and average values of the TPoD performance and compare them to the TPoD targets in Table 1. We remark that the discretization (e.g., ceiling, rounding, and flooring) does not play an important role in this example because the number of servers is relatively large (ranging from 50 to 90).

In Figure 4 we plot the TTGA staffing functions for three cases: $\alpha = 0.1, 0.5$, and 0.9 (the three functions on the top), with a fixed delay target $w = 0.5$. The

Figure 4. (Color online) The Staffing Functions with $\alpha = 0.1, \ldots, 0.9$, $w = 0.5$, and $\lambda(t) = 100 + 20\sin(t)$



detailed staffing formula of this example is given in Section EC.4 of the e-companion, as a straightforward result of the general TTGA staffing function (6)–(8) in Section 2. Indeed, we see that the secondary staffing terms $s^{(2)}_{0.5,0.1}(t) > 0$, $s^{(2)}_{0.5,0.5}(t) = 0$ and $s^{(2)}_{0.5,0.9}(t) < 0$, so that the TTGA staffing functions for these three cases are ordered as $s_{0.5,0.1} > s^{(1)}_{0.5} = s_{0.5,0.5} > s_{0.5,0.9}$. Figure 4 helps visualize the relation $s^{(2)}_{w,\alpha} = O(\sqrt{s^{(1)}_w})$ and justify the names "first-order" and "second-order" for $s^{(1)}_w$ and $s^{(2)}_{w,\alpha}$.

4.2. Small-Scale Systems

When the arrival rate is larger, e.g., $\bar{\lambda} = 1{,}000$ in (11), the TPoD is, unsurprisingly, even more stable. (This is supported by the asymptotic stability result.) We give the simulations in the online appendix. We next investigate the important case of smaller arrival rates.

4.2.1. Base Model with $\bar{\lambda} = 10$.

To make this challenging, we consider the $H_2(t)/M/s_t + H_2$ example in Section 4.1 having the same sinusoidal arrival-rate function in (11) with $\bar{\lambda} = 10$ (instead of 100 in Section 4.1). We remark that an average arrival $\bar{\lambda} = 10$ does not necessarily imply an average staffing level with around 10 servers. In fact, for a large α, the number of servers is around 3. See Figure 5 for the TTGA staffing levels for three cases: (i) $\alpha = 0.1$ (average staffing is 11), (ii) $\alpha = 0.5$ (average staffing is 7) and (iii) $\alpha = 0.9$ (average staffing is 3).

Figure 6 shows that the TTGA staffing method continues to achieve time-stable performance of TPoD for the case of smaller arrival rates with $\alpha = 0.1, \ldots, 0.9$. Other performance measures and approximations are given in the online appendix. Despite the relatively larger oscillations comparing with the example in Section 4.1 with a large $\bar{\lambda}$, the TPoD is again well

Figure 5. (Color online) Staffing Functions for $\alpha = 0.1, 0.5, 0.9$ with $w = 0.5$ and $\lambda(t) = 10 + 2\sin t$

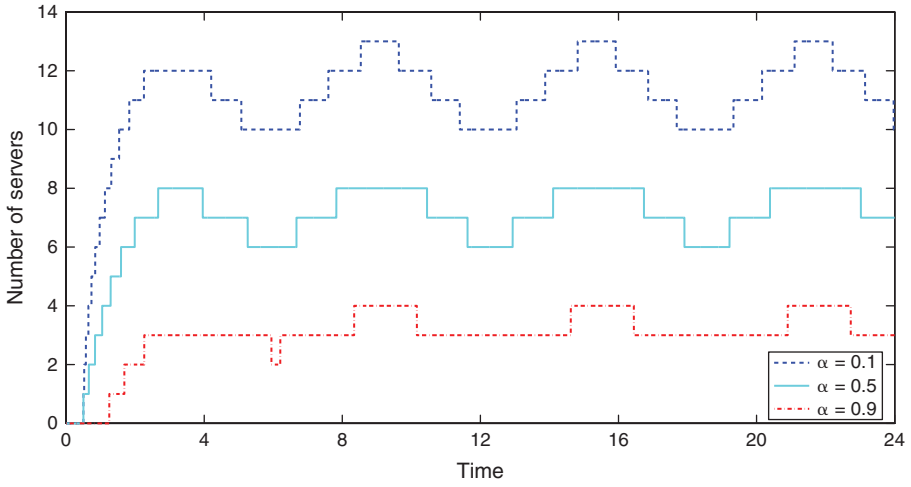
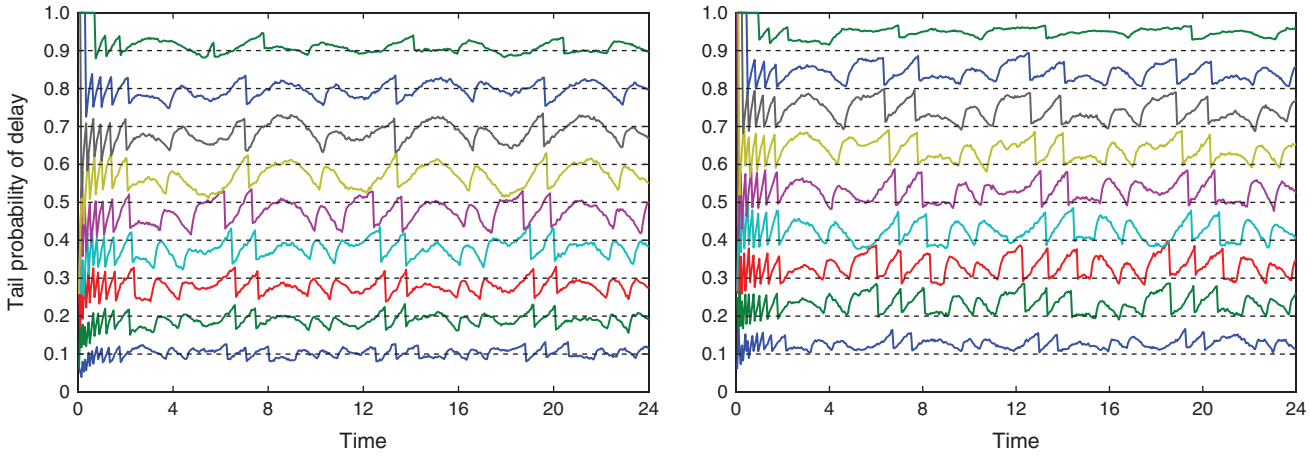


Figure 6. (Color online) Tail Probabilities in a Small System, with Staffing Functions Discretized by Ceiling (Left) and Flooring (Right), for $\alpha = 0.1, \ldots, 0.9$, $w = 0.5$ and $\lambda(t) = 10 + 2\sin t$



stabilized. We point out that the nonsmoothness of the TPoD is caused by (i) the discretization of the continuous staffing functions $s_w^{(1)}$ and $s_{w,\alpha}^{(2)}$ (see Figure 5), (ii) the choice of discretization methods (ceiling, rounding or flooring), and (iii) the Gaussian-truncation effect, which we further explain in Remarks 6 and 7.

Remark 6 (Effect of Discretization). Unlike the case $\bar{\lambda} = 100$, the performance of TPoD becomes far more sensitive to the discretization of the staffing function. When the discrete staffing levels add (remove) one server at time t, TPoD inevitably decreases (increases) quite significantly. Furthermore, TPoD tends to fluctuate quite frequently at the beginning (before time 2), because the TTGA staffing function climbs from 0 to its average value very fast, as shown in Figure 5. We also remark that the choice of discretization methods can play a much bigger role for smaller arrival rates. We compare the performance of the TPoD in Figure 6 for two staffing functions: (i) $\lceil s_w^{(1)} + s_{w,\alpha}^{(2)} \rceil$ (as in (6)) and (ii) $\lfloor s_w^{(1)} + s_{w,\alpha}^{(2)} \rfloor$. Figure 6 shows that adding a server to (removing a server from) the staffing function at all t can significantly alter the performance for smaller systems, while doing so does not make a big difference for large systems (as in Table 1).

Remark 7 (The Gaussian Truncation Effect). Because we have the lowest (highest) staffing level for $\alpha = 0.9$ ($\alpha = 0.1$) (see the bottom and top staffing functions in Figure 5), and also because the impact of the discretization increases as the staffing level decreases, we would then naturally expect the TPoD to oscillate more (less) severely for $\alpha = 0.9$ ($\alpha = 0.1$). Surprisingly, we observe from Figure 6 that the magnitude of the fluctuation of TPoD does not monotonously increase in α! In fact, the TPoD tends to fluctuate the most around $\alpha = 0.5$. We call this the *Gaussian truncation effect*, which we will explain next.

Recall that TTGA is based on a Gaussian approximation for the random delay $V(t)$. We use the first-order staffing term $s_w^{(1)}$ to center $V(t)$ near the delay target w (the first-order QoS target) and then use the second-order term $s_{w,\alpha}^{(2)}$ to make finer adjustments so that we can furthermore truncate that Gaussian random variable to obtain the desired probability α (the second-order QoS target). As discussed in Remark 6, the discretization starts to play a much bigger role for smaller systems, causing the TPoD to deviate from the desired target. We call this deviation the discretization error of staffing. When α is either small (e.g., 0.1) or big (e.g., 0.9), the truncation of the Gaussian ccdf will not deviate much from α, because the value of the Gaussian pdf is small when α is either small or large; see parts (a) and (c) in Figure 7 for an illustration. However, the worse case is $\alpha = 0.5$: even a slight discretization error will make a significant impact on the truncation, because that is where the Gaussian pdf peaks (see Figure 7(b)). This is why the TPoD is more (less) sensitive to the discretization error for a medium α (small or large α).

4.2.2. The Belgian Hospital Example. Next we consider an example with realistic arrival rate estimated from the emergency department of a Belgian hospital, studied in Defraeye and van Nieuwenhuyse (2013); we aim to provide direct comparison with the performance of the simulation-based staffing algorithm there, showing that we can achieve the same goal without needing simulations.

In Defraeye and van Nieuwenhuyse (2013), the authors used ISA to keep the TPoD below $\alpha = 0.1$ for $w = 10$ minutes in an example with realistic arrival rate estimated from the emergency department of a Belgian hospital; see Defraeye and van Nieuwenhuyse (2013, figure 4(a)) (also see the top panel of Figure 8 here). They assumed that the service and patience times are exponentially distributed with mean 30 minutes and 4 hours, respectively. Here we apply the TTGA staffing method to the same example. See Figure 8 for the daily arrival rate, TTGA staffing function, and TPoD with target $\alpha = 0.1$. This is an extremely challenging example because the arrival rate is very small (between 1 and 4),

Figure 7. (Color online) Performance Oscillation in Smaller Systems: The Gaussian Truncation Effect

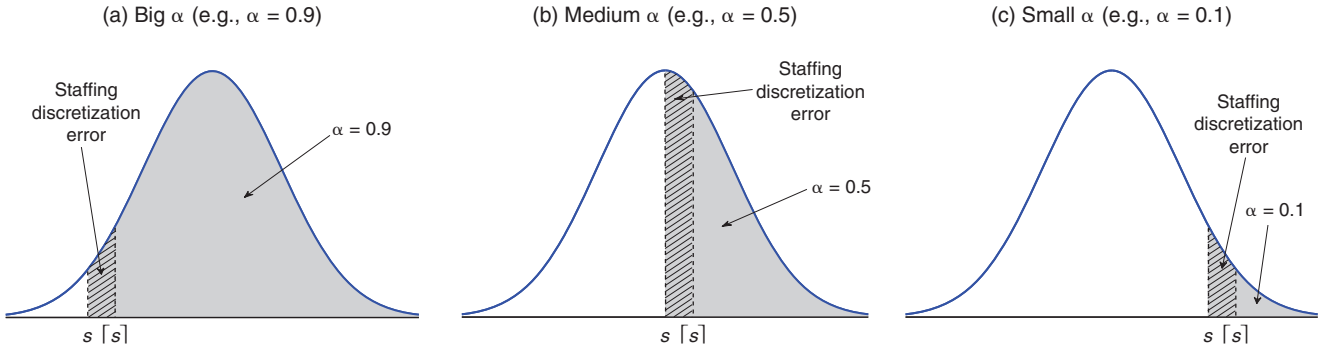
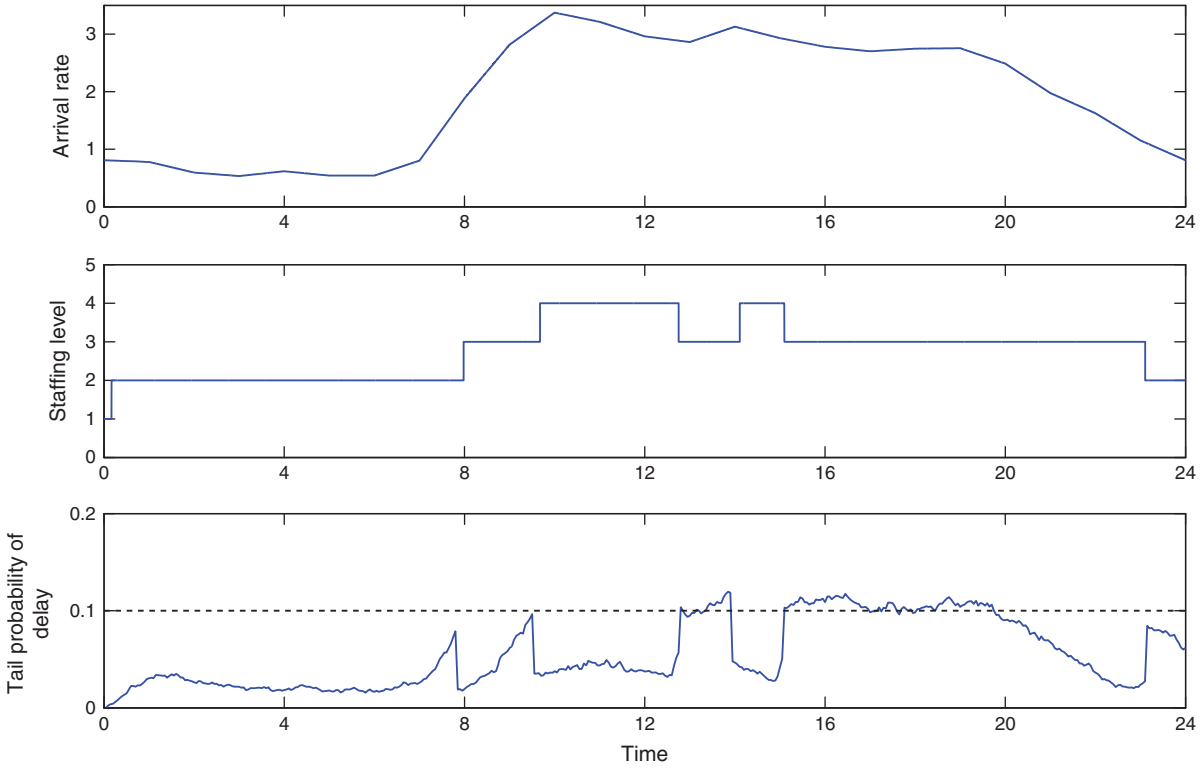


Figure 8. (Color online) Arrival Rate, TTGA Staffing Function, and TPoD for the Belgian Hospital Example, with Parameters $\mu = 2$, $\theta = 0.25$, and $w = 1/6$



and our TTGA staffing functions only change a few times (varying between 1 and 4). Comparing with figure 5 in Defraeye and van Nieuwenhuyse (2013), TTGA achieves similar good performance without using simulations.

4.3. Lightly-Loaded and Heavily-Loaded Systems

To supplement the base example in Section 4.1 with $w = 0.5$ and α ranging from 0.1 to 0.9, we now consider cases where the system is (i) lightly loaded with $w \approx 0$ (Section 4.3.1); (ii) lightly loaded with smaller α (e.g., $\alpha = 0.02$, see Section 4.3.2); and (iii) heavily loaded with bigger w (e.g., $w = 3$, Section 4.3.3).

4.3.1. High QoS with Small Delay Targets. Despite the fact that the effectiveness of TTGA is supported by a heavy-traffic limit theorem assuming $w > 0$ so the system is in the overloaded (or ED) regime, our staffing formulas still work quite well for the case of high QoS with delay target $w \approx 0$, which places the system in the QED regime where TPoD reduces to PoD. To substantiate this claim, we repeat the simulations for the example in Section 4.1 for $w = 0.05$ (i.e., 1/20 of the mean service time), with all the other model parameters unchanged. Figure 9 shows that the TPoD is stabilized and close to the desired target α ranging from 0.1 to 0.9. We also consider the extreme case of $w = 0$ in Section 6 of the online appendix.

4.3.2. High QoS with Small Probability Targets. Supplementing the good performance for a system in the

Figure 9. (Color online) Tail Probabilities of Delay for High QoS Targets: $w = 0.05$ with $\lambda(t) = 100 + 20 \sin t$

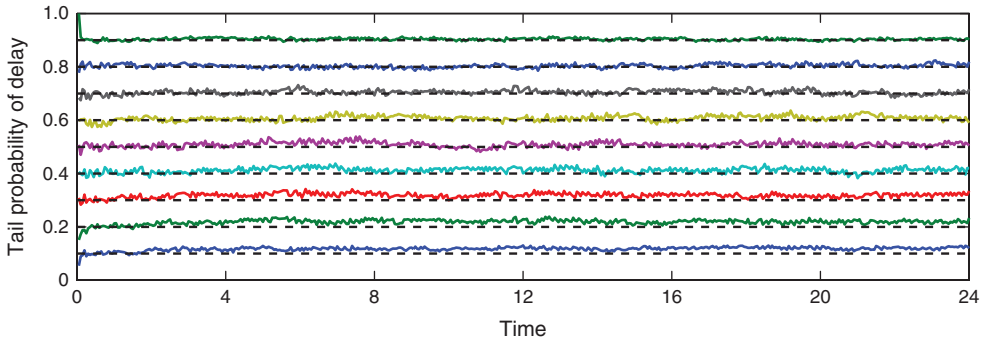
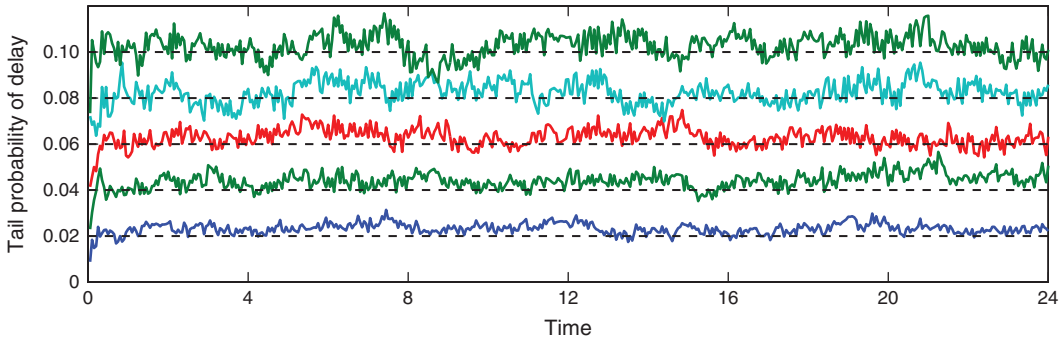


Figure 10. (Color online) Tail Probabilities of Delay for High QoS Targets: $w = 0.3$ and $\alpha = 0.02, 0.04, 0.06, 0.08$ with $\lambda(t) = 100 + 20 \sin t$



QED regime, we now consider even smaller probability target $\alpha \approx 0$. We repeat the example in Section 4.1 for $\alpha < 0.1$ and moderate delay target $w = 0.3$. Figure 10 shows that the TPoD is stabilized and close to the desired target α. We remark that the choices of smaller $0.02 \leqslant \alpha \leqslant 0.08$ have already made the system operate in the QED regime, because the corresponding PoDs range between 0.4 to 0.8; see the online appendix for more discussions.

4.3.3. Heavily Loaded Systems with Large Delay Targets. We now consider the case of relatively large delay targets by repeating the example in Section 4 with $w = 3$ and leaving all the other parameters unchanged. Examples having even larger delay targets (e.g., $w = 6$) are also considered in the online appendix.

Since all customers wait for a relatively long time, a large proportion of them eventually abandon. Therefore, we end up again investigating the performance of TTGA for a small-scale system, which supplements examples in Section 4.2. As shown in Figure 11, the average staffing levels are from 8 to 16 as α varies from 0.1 to 0.9, despite of a relatively large arrival rate $\bar{\lambda} = 100$.

In Figure 12, simulations show that TTGA continues to achieve time-stable performance for the case of large delay targets, with a wide range of α ($\alpha = 0.1, \ldots, 0.9$). Because we are again facing a smaller system (having a small number of servers), we point out that the relatively larger fluctuations of the TPoD are again caused by (i) the discretization of the continuous staffing functions $s_w^{(1)}$ and $s_{w,\alpha}^{(2)}$ and (ii) the Gaussian-truncation effect. Simulations of other performance measures and their approximations are given in the online appendix.

4.3.4. Staffing Level Comparison for Different Delay Targets. We now compare the TTGA staffing levels as a function of the delay targets. Considering the same example in Section 4, we plot the staffing functions in Figure 13 for $w = 0, 0.1, 0.3, 0.5, 1, 2, 3, 6$ with $\alpha = 0.1$ and 0.9. We observe that the TTGA staffing level decreases significantly as the delay target w increases, because larger delay leads to more abandonment and smaller system size (thus fewer servers).

We also remark that although the TTGA staffing function remains 0 at the beginning (before time w),

Figure 11. (Color online) Staffing Functions for $\alpha = 0.1, 0.5$ and 0.9, with $w = 3$ and $\lambda(t) = 100 + 20 \sin t$

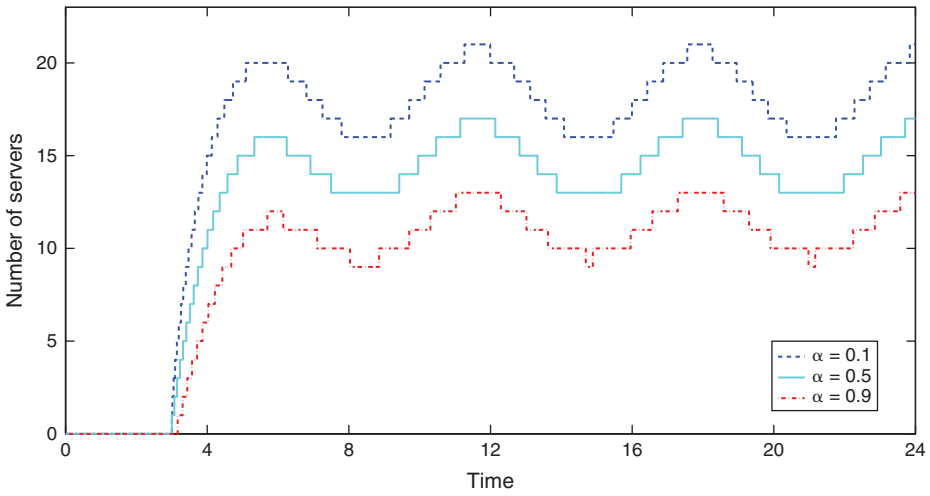


Figure 12. (Color online) Tail Probabilities of Delay with Staffing Functions Discretized by Ceiling (Left) and Flooring (Right) for $\alpha = 0.1, \ldots, 0.9$, Large Delay Target $w = 3$ and $\lambda(t) = 100 + 20 \sin t$

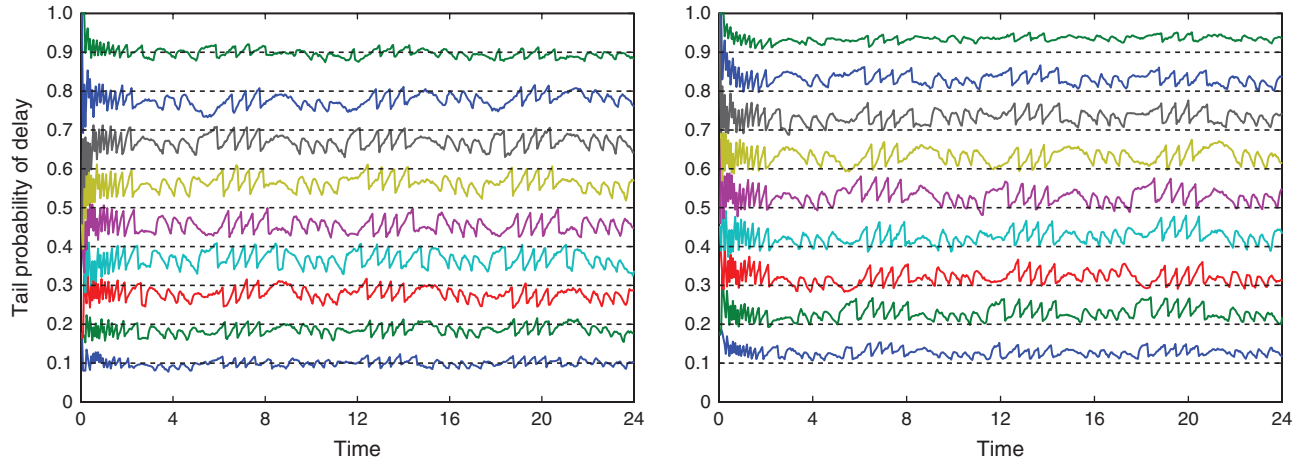
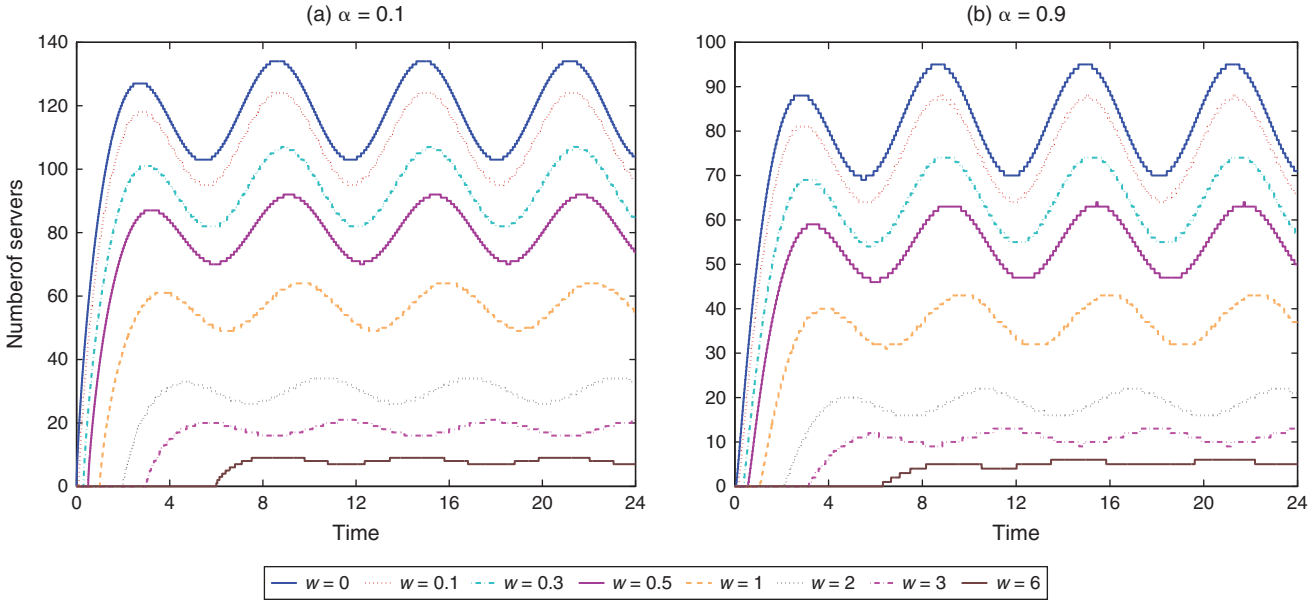


Figure 13. (Color online) TTGA Staffing Levels for Different Delay Target $0 \leqslant w \leqslant 6$, with (i) $\alpha = 0.1$ (Left) and (ii) $\alpha = 0.9$ (Right)



the TPoD is well stabilized for all time t, including the interval $[0, w]$. This is so because the delay $V(t)$ of an arriving customer at t (i.e., the potential waiting time at t) is realized when the customer enters service at a future time $t + V(t)$. Since everybody waits around w (with minor adjustment according to the second-order QoS target α), no one should enter service before w, which yields a zero staffing level at the beginning.

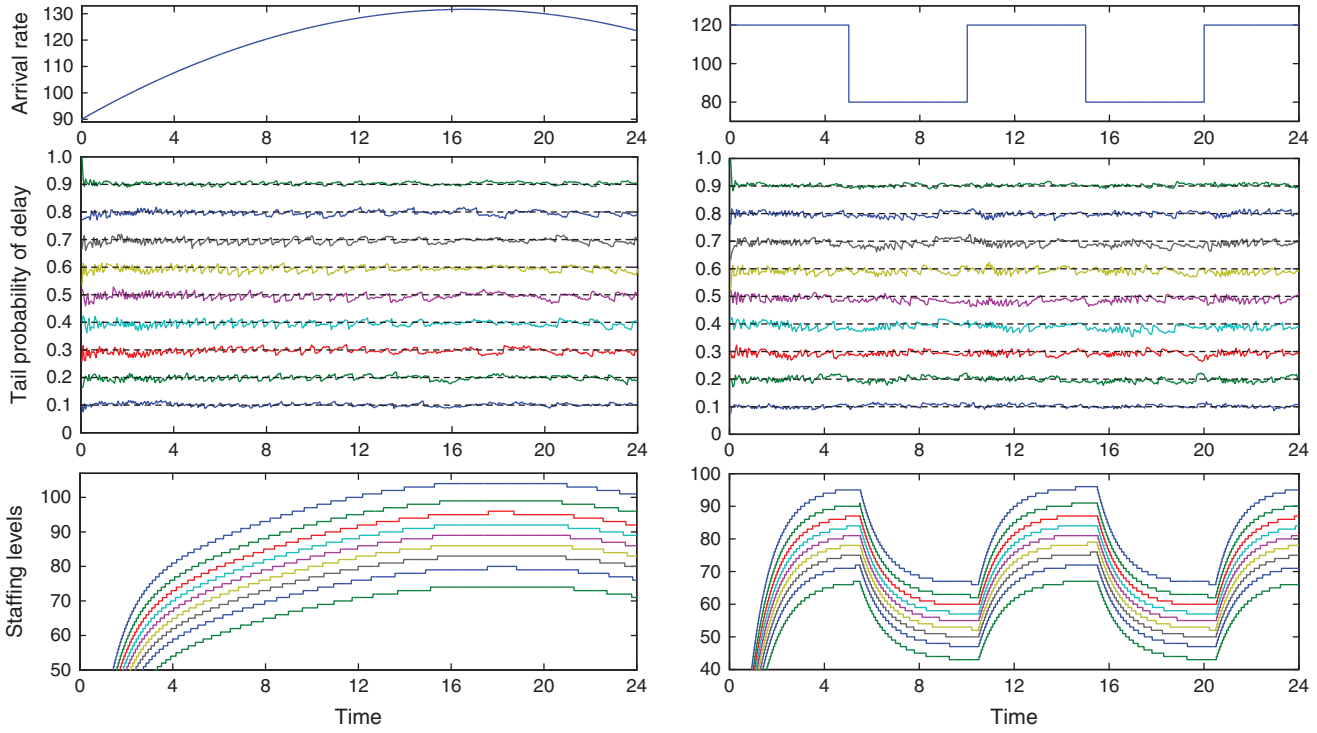
4.4. Other Arrival Rate Functions

We now consider arrival-rate functions having different structures. Specifically, we consider (i) quadratic $\lambda(t) = 90 + 5t - 0.15t^2$ and (ii) piecewise constant $\lambda(t) = 120 \cdot \mathbf{1}_A(t) + 80 \cdot \mathbf{1}_{A^c}(t)$, with the set $A \equiv [0, 5) \cup [5, 10) \cup [15, 20)$. Simulation experiments again verify the effectiveness of the TTGA staffing methods; see Figure 14, which is an analog of Figure 3. We include cases of other arrival rates, such as constant and piecewise linear functions, in the online appendix.

4.5. Nonexponential Service Distribution

Although the supporting asymptotic stability theorem of TTGA for GI service remains an open problem, we provide simulations to verify that TTGA performs well for GI service. In particular, we consider an $H_2(t)/LN/s_t + H_2$ model, having a *lognormal* (LN) service distribution and all other parameters the same as the main example in Section 4.1. We first review the basics of the LN distribution. Let S be a generic service time, we write $S = e^Z$ where $Z = \mathcal{N}(\bar{\mu}, \bar{\sigma}^2)$. Then the pdf

Figure 14. (Color online) Tail Probabilities of Delay with (i) Quadratic Arrival Rate $\lambda(t) = 90 + 5t - 0.15t^2, 0 \leqslant t \leqslant 24$ (Left); and (ii) Piecewise Constant Arrival Rate Alternating Between 120 and 80 Every 5 Time Units (Right)



of S is

$$g(x) = \frac{1}{x\bar{\sigma}\sqrt{2\pi}} e^{-(\log x - \bar{\mu})^2/(2\bar{\sigma}^2)},$$

with mean, variance, and SCV

$$\mathrm{E}[S] = e^{\bar{\mu}+\bar{\sigma}^2/2}, \quad \mathrm{Var}(S) = (e^{\bar{\sigma}^2} - 1)e^{2\bar{\mu}+\bar{\sigma}^2},$$
$$\text{and} \quad c_s^2 = e^{\bar{\sigma}^2} - 1.$$

We fix the mean service time $\mathrm{E}[S] = 1$ and vary the SCV c_s^2 in three cases: (i) $c_s^2 = 0.25$ ($\bar{\mu} = -0.3466$ and $\bar{\sigma} = 0.8326$), (ii) $c_s^2 = 1$ ($\bar{\mu} = -0.1116$, $\bar{\sigma} = 0.4724$), and (iii) $c_s^2 = 4$ ($\bar{\mu} = -0.8047$, $\bar{\sigma} = 1.2686$). We use $LN(a, c^2)$ to denote an LN distribution with mean a and SCV c^2.

The cases of $c_s^2 = 0.25$ and 1 are of less interest because service times in many service systems are more volatile than exponential, exhibiting an SCV larger than 1, e.g., $c_s^2 = 1.52$ in financial call centers; see Brown et al. (2005). We now consider the more challenging case $c_s^2 = 4$. Figure 15 (an analog of Figure 3) shows that (i) the TPoD is stabilized at desired targets for the $H_2(t)/LN(1,4)/s_t + H_2$ model and (ii) other important performance functions all tend to agree with our approximating formulas, except for a short initial warm-up period. We give the simulations for the cases $c_s^2 = 0.25$ and 1 in the online appendix, because they are similar to (in fact better than) Figure 15. We have also conducted simulations in the online appendix for models with other types of nonexponential service distributions, such as H_2 distribution.

5. Proofs

We first review useful MSHT limit theorems from Liu and Whitt (2012b, c, 2014a) in Sections 5.1–5.2. We next prove the asymptotic stability of TTGA (Theorem 2) and asymptotic accuracy of the performance approximations (Theorem 3) in Sections 5.3–5.4. We give the proofs of other results in the online supplement.

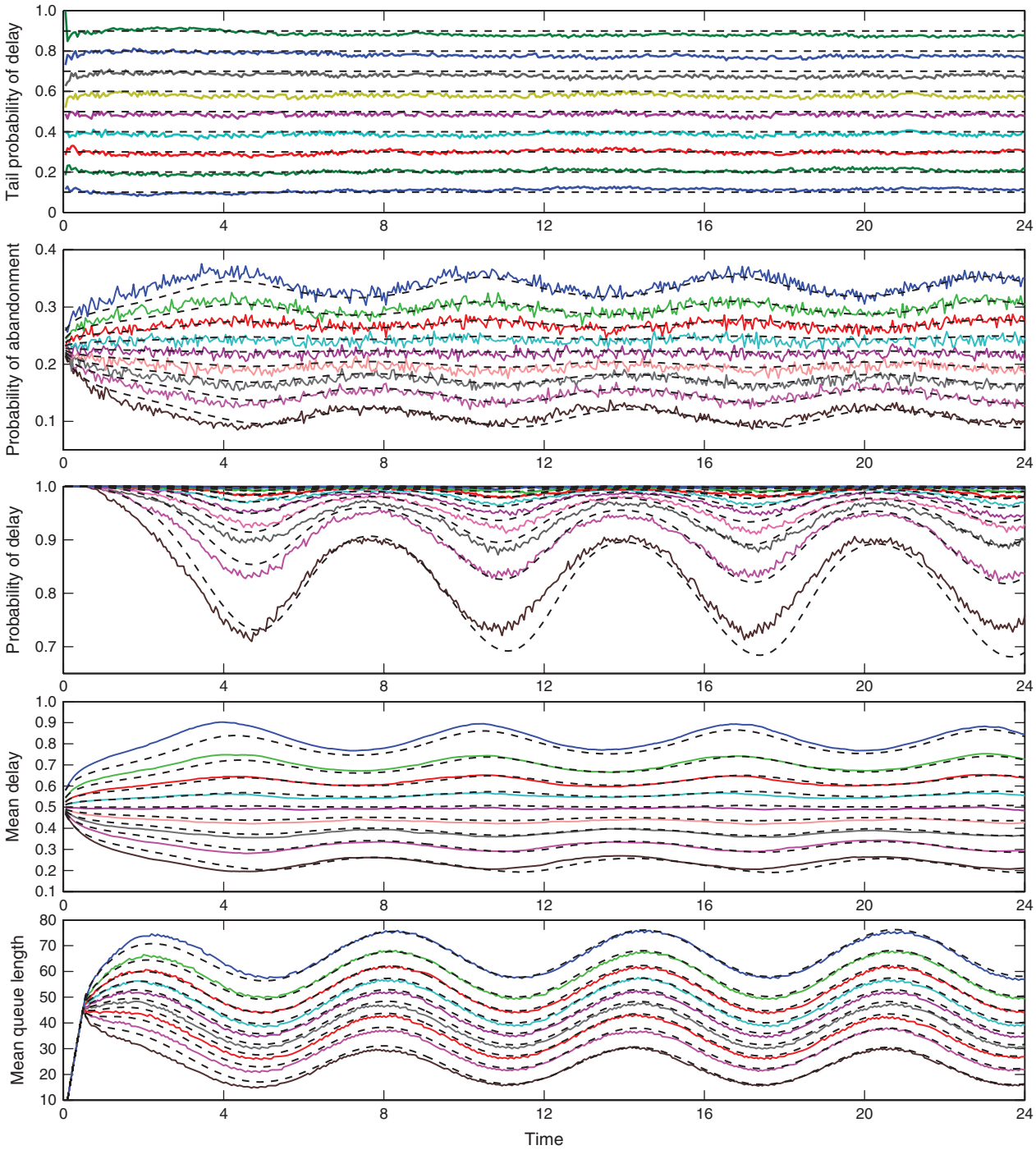
5.1. A Sequence of Queueing Models Indexed by n

We follow the convention by defining a sequence of $G_t/GI/s_t + GI$ queueing models indexed by n, with the nth model having a nonstationary non-Poisson arrival process $N_n(t)$, i.i.d. general service times with cdf G, i.i.d. general patience times with cdf F, and a time-varying staffing function

$$s_n(t) \equiv \lceil n\, s(t) + \sqrt{n} s_g(t) \rceil, \tag{16}$$

that is a generalized version of the SRS. Paralleling Section 2, we let $N_n(t) \equiv N^{(0)}(n\Lambda(t))$ where $N^{(0)}(t)$ is a rate-1 ERP with interrenewal SCV c_λ^2 and $\Lambda(t) \equiv \int_0^\infty \lambda(u)du$. We assume the base staffing and arrival-rate functions $s(t), s_g(t)$ and $\lambda(t)$ are all nonnegative piecewise smooth functions. We also assume the pdf's of service and patience times g and f exist. We let $\bar{G} \equiv 1 - G$ and $\bar{F} = 1 - F$ be the ccdf's of G and F, and let $h_F(x) \equiv f(x)/\bar{F}(x)$ be the patience-time hazard-rate function. Thus, the $G_t/GI/s_t + GI$ queueing model indexed by n is characterized by the six-tuple $(\lambda(t), c_\lambda^2, G, F, s(t), s_g(t))$ and the scaling factor $n \geqslant 1$.

Figure 15. (Color online) A Simulation Comparison: Estimated Time-Dependent TPoD, PoD, Mean Queue Length for the $H_2(t)/LN(1,4)/s_i + H_2$ Example with a Sinusoidal Arrival Rate $\lambda(t) = 100 + 20\sin(t)$, Service SCV $c_s^2 = 4$, $w = 0.5$ and $\alpha = 0.1, \ldots, 0.9$



Although we allowed the average arrival rate $\bar{\lambda}$ to increase in Theorems 1–3, we point out that is equivalent to allowing the scale n to increase, because we can always rewrite $\lambda_n(t) = n\bar{\lambda} \cdot (\lambda(t)/\bar{\lambda}) = \bar{\lambda}_n \cdot (\lambda(t)/\bar{\lambda})$ and let the average $\bar{\lambda}_n$ go to infinity. Considering the main example in Section 4, we may let $n = 100$ and the base arrival-rate function $\lambda(t) = 1 + 0.2\sin(t)$ so that the staffing function of the main example is $\lambda_n(t) = n\lambda(t) = 100 + 20\sin(t)$. We remark that the choice of n is not unique; we can alternatively let $n = 50$ and $\lambda(t) = 2 + 0.4\sin(t)$, because the resulting staffing function $s_n(t)$ will remain the same. This is why we allowed the average arrival to increase.

Here both s and s_g in the staffing function (16) are part of the model input. Although we have not yet set

the staffing function in the desired form of TTGA, we can see that the staffing function in (16) indeed has the correct orders (with n in the first term s and $\sqrt{n}$ in the second term s_g), as discussed in Remark 1.

For the nth queueing model, let $B_n(t)$ and $Q_n(t)$ be the total number of customers being served in service and waiting in queue at time t, respectively. Also let $X_n(t) \equiv B_n(t) + Q_n(t)$ be the total number of customers in system at time t. Let $V_n(t)$ be the *offered waiting time* at t, that is the virtual waiting time of an arrival at t assuming infinite patience; and let $W_n(t)$ be the *head-of-line waiting time* (HWT) at t, that is the elapsed delay of the head-of-line customer, if there is any. Define the time-dependent TPoD as

$$p_n(t, w) \equiv P(V_n(t) > w) \tag{17}$$

for any delay target $w > 0$.

5.2. MSHT Fluid and Diffusion Limits

We next provide useful MSHT *functional weak law of large numbers* (FWLLN) and FCLT for $G_t/M/s_t + GI$ queues with time-varying arrivals and staffing functions, and exponential service-time cdf $G(x) = 1 - e^{-\mu x}$, developed in Liu and Whitt (2012b, 2014a). We impose a special initial condition by assuming the system initially critically loaded; that is,

$$Q_n(0) = 0 \quad \text{and} \quad X_n(0) = B_n(0) = s_n(0), \quad \text{for all } n \geqslant 1. \tag{18}$$

For the system to be in the ED regime, we assume the queueing system is asymptotically overloaded, which requires the total input dominates the maximum output, namely,

$$\Lambda(t) > D(t) \equiv \int_0^t \mu s(u)\, du, \quad \text{for all } 0 \leqslant t \leqslant T. \tag{19}$$

Following Liu and Whitt (2012b, 2014a), the LLN-scaled processes are defined as

$$\bar{B}_n(t) \equiv n^{-1} B_n(t), \quad \bar{Q}_n(t) \equiv n^{-1} Q_n(t), \quad \bar{X}_n(t) \equiv n^{-1} X_n(t).$$

The next result is a special case of the FWLLN developed by Liu and Whitt (2012b, 2014a).

Theorem 4 (FWLLN Limits for the Overloaded $G_t/M/s_t + GI$ Queue). *For a sequence of $G_t/M/s_t + GI$ models defined in Section 5.1 with staffing function (16). If the assumptions in (18) and (19) are satisfied, then the LLN-scaled processes*

$$(\bar{B}_n, \bar{Q}_n, \bar{X}_n, W_n, V_n) \Rightarrow (s, Q, s + Q, w, v) \quad \text{in } \mathbb{D}^5, \quad \text{as } n \to \infty, \tag{20}$$

where $w(t)$ and $v(t)$ uniquely solve the ordinary differential equations (ODEs)

$$\dot{w}(t) = 1 - \frac{\mu s(t) + \dot{s}(t)}{\bar{F}(w(t))\lambda(t - w(t))} \quad \text{and}$$

$$\dot{v}(t) = \frac{\bar{F}(v(t))\lambda(t)}{\mu s(t + v(t)) + \dot{s}(t + v(t))} - 1,$$

with initial values $w(0) = v(0) = 0$,

$$Q(t) \equiv \int_{t - w(t)}^t \lambda(u)\bar{F}(t - u)\, du,$$

and $\mathbb{D}$ is the space of right-continuous functions having limits from the left.

Following Liu and Whitt (2014a), the CLT-scaled processes are

$$\hat{B}_n(t) \equiv n^{1/2}(\bar{B}_n(t) - s(t)), \quad \hat{Q}_n(t) \equiv n^{1/2}(\bar{Q}_n(t) - Q(t)),$$
$$\hat{X}_n(t) \equiv n^{1/2}(\bar{X}_n(t) - s(t) - Q(t)),$$
$$\hat{W}_n(t) \equiv n^{1/2}(W_n(t) - w(t)), \quad \hat{V}_n(t) \equiv n^{1/2}(V_n(t) - v(t)).$$

The next result is a special case of the FCLT developed by Liu and Whitt (2014a).

Theorem 5 (FCLT Limits for the Overloaded $G_t/M/s_t + GI$ Queue). *For a sequence of $G_t/M/s_t + GI$ models defined in Section 5.1 with staffing function (16). If the assumptions in (18) and (19) are satisfied, then the CLT-scaled processes*

$$(\hat{B}_n, \hat{Q}_n, \hat{X}_n, \hat{W}_n, \hat{V}_n) \Rightarrow (0, \hat{X}, \hat{X}, \hat{W}, \hat{V}) \quad \text{in } \mathbb{D}^5, \quad \text{as } n \to \infty, \tag{21}$$

where $\hat{X}$, $\hat{V}$ and $\hat{W}$ are continuous-path Gaussian processes. For $t \geqslant 0$, $\hat{X}(t) = \mathcal{N}(X_g(t), \sigma^2_{\hat{X}}(t))$, $\hat{V}(t) = \mathcal{N}(V_g(t), \sigma^2_{\hat{V}}(t))$, $\hat{W}(t) = \mathcal{N}(W_g(t), \sigma^2_{\hat{W}}(t))$ are Gaussian random variables with

$$X_g(t) \equiv \lambda(t - w(t))\bar{F}(w(t))W_g(t),$$

$$V_g(t) \equiv \frac{W_g(t + v(t))}{1 - \dot{w}(t + v(t))}, \quad W_g(t) \equiv -\int_0^t H(t, u)z(u)\, du$$

$$\sigma^2_{\hat{V}}(t) \equiv \frac{\sigma^2_{\hat{W}}(t + v(t))}{[1 - \dot{w}(t + v(t))]^2}, \quad \sigma^2_{\hat{W}}(t) \equiv \int_0^t H^2(t, u)I^2(u)\, du$$

$$\sigma^2_{\hat{X}}(t) \equiv \int_{t - w(t)}^t \lambda(u)\bar{F}(t - u)[(c^2_\lambda - 1)\bar{F}(t - u) + 1]\, du + (\lambda(t - w(t))\bar{F}(w(t)))^2\sigma^2_{\hat{W}}(t),$$

$$z(t) \equiv \frac{s_g(t)\mu + \dot{s}_g(t)}{\lambda(t - w(t))\bar{F}(w(t))},$$

$$I^2(t) \equiv \frac{\mu s(t) + [1 + (c^2_\lambda - 1)\bar{F}(w(t))](\mu s(t) + \dot{s}(t))}{[\lambda(t - w(t))\bar{F}(w(t))]^2}, \quad \text{and}$$

$$H(t, u) \equiv \exp\left\{\int_u^t (1 - \dot{w}(x))\left(\frac{-\dot{\lambda}(x - w(x))}{\lambda(x - w(x))} - h_F(w(x))\right) dx\right\}. \tag{22}$$

5.3. Proof of Theorem 2

We are now ready to show the asymptotic stability of our TTGA in Theorem 2 and the asymptotic accuracy of our approximating formulas in Theorem 3. In the new notation using the scaling factor n (that is, replacing

$\lambda(t)$ in (6) by $\lambda_n(t) = n\lambda(t)$ with $\lambda(t)$ being the base arrival rate), our TTGA staffing function for the nth model becomes

$$s_n(t) = \lceil n s_w^{(1)}(t) + \sqrt{n} s_{w,\alpha}(t) \rceil^+, \tag{23}$$

where the definitions of $s_w^{(1)}(t)$ and $s_{w,\alpha}^{(2)}(t)$ remain the same as in (5) and (7).

Our proofs are based on the following two lemmas.

Lemma 1. *For a sequence of $G_t/M/s_t + GI$ models defined in Section 5.1 with staffing function (23). If the assumptions in (18) and (19) are satisfied, then both of the fluid waiting times $w(t)$ and $v(t)$ in Theorem 4 are stabilized at w for all t.*

Proof. Because the second term in (23) is $O(\sqrt{n})$, then the staffing function in the associated limiting fluid model is $\lim_{n\to\infty} n^{-1} s_n(t) = s_w^{(1)}(t)$. Hence, the claimed result holds by Theorems 1 and 4 here, along with Theorem 8 in section 10 of Liu and Whitt (2012b). □

Lemma 2. *For a sequence of $G_t/M/s_t + GI$ models defined in Section 5.1 with staffing function (23). If the assumptions in (18) and (19) are satisfied, then the means and standard deviations of the Gaussian FCLT limit of delays $\hat{V}$ and $\hat{W}$ in Theorem 5 satisfy the following relations*:

$$V_g(t) = W_g(t+w), \quad \sigma_{\hat{V}}(t) = \sigma_{\hat{W}}(t+w),$$
$$W_g(t) = -z_\alpha \sigma_{\hat{W}}(t), \quad V_g(t) = -z_\alpha \sigma_{\hat{V}}(t), \quad t \geqslant 0.$$

Proof. Lemma 1 implies $w(t) = w$ ($\dot{w}(t) = 0$). The function $H(t,u)$ in (22) simplifies to

$$\begin{aligned} H(t,u) &= \exp\left(-\int_u^t \frac{\dot{\lambda}(x-w)}{\lambda(x-w)}\,dx - h_F(w)(t-u)\right) \\ &= \exp\left(-\int_u^t d\log(\lambda(x-w)) - h_F(w)(t-u)\right) \\ &= \frac{\lambda(u-w)}{\lambda(t-w)} e^{-h_F(w)(t-u)}. \end{aligned} \tag{24}$$

Replacing $s(t)$ and $s_g(t)$ in (22) by $s_w^{(1)}(t)$ and $s_{w,\alpha}^{(2)}(t)$, we have

$$\begin{aligned} \sigma_{\hat{W}}^2(t) &= \int_0^t \frac{\lambda^2(u-w)}{\lambda^2(t-w)} e^{-2h_F(w)(t-u)} \\ &\qquad \frac{((c_\lambda^2-1)\bar{F}(w)+2)(\mu s_w^{(1)}(u) + \dot{s}_w^{(1)}(u)) - \dot{s}_w^{(1)}(u)}{q^2(u,w)}\,du \\ &= \frac{e^{-2h_F(w)t}}{\lambda^2(t-v)(\bar{F}(w))^2} \int_0^t e^{2h_F(w)u}([(c_\lambda^2-1)F^c(w)+2] \\ &\quad \cdot(\mu s_w^{(1)}(u) + \dot{s}_w^{(1)}(u)) - \dot{s}_w^{(1)}(u))\,du. \\ &= \frac{e^{-2\mu t}}{\lambda^2(t-w)(\bar{F}(w))^2} Z(t), \end{aligned} \tag{25}$$

where the first equality holds by (24) and the last equality holds by the definition of $Z(t)$ in (8). Note that $s_w^{(1)}(t) = \dot{s}_w^{(1)}(t) = Z(t) = 0$ for $0 \leqslant t \leqslant w$. Next,

$$\begin{aligned} W_g(t) &= -\int_0^t \frac{\lambda(u-w)}{\lambda(t-w)} e^{-h_F(w)(t-u)} \frac{s_{w,\alpha}^{(2)}(u)\mu + \dot{s}_{w,\alpha}^{(2)}(u)}{q(u,w)}\,du \\ &= -\frac{e^{-h_F(w)t}}{\lambda(t-w)\bar{F}(w)} \int_0^t e^{h_F(w)u}(s_{w,\alpha}^{(2)}(u)\mu + \dot{s}_{w,\alpha}^{(2)}(u))\,du \\ &= -\frac{e^{-h_F(w)t}}{\lambda(t-w)\bar{F}(w)} \int_w^t e^{h_F(w)u}\left\{ z_\alpha e^{-\mu u}\left(Z(u) \right.\right. \\ &\qquad \left. -(\mu - h_F(w)) \int_w^u Z(x)\,dx \right)\mu \\ &\quad + z_\alpha e^{-\mu u}\left[-\mu\left(Z(u) - (\mu - h_F(w)) \int_w^u Z(x)\,dx \right)\right. \\ &\qquad \left.\left. + \dot{Z}(u) - (\mu - h_F(w))Z(u) \right]\right\} du \\ &= -\frac{z_\alpha e^{-h_F(w)t}}{\lambda(t-w)\bar{F}(w)} \\ &\quad \cdot \int_w^t e^{(h_F(w)-\mu)u}\left(\dot{Z}(u) - (\mu - h_F(w))Z(u)\right) du \\ &= -\frac{z_\alpha e^{-h_F(w)t}}{\lambda(t-w)\bar{F}(w)} e^{(h_F(w)-\mu)t} Z(t) \\ &= -\frac{z_\alpha e^{-\mu t}}{\lambda(t-w)\bar{F}(w)} Z(t). \end{aligned} \tag{26}$$

Combining (25) and (26) completes the proof. □

Finishing the proof of Theorem 2. Lemma 2 implies that under the TTGA staffing (23), the FCLT limit for $\hat{V}_n$ is a negative-mean Gaussian random variable

$$\hat{V}(t) = \mathcal{N}(-z_\alpha \sigma_{\hat{W}}(t), \sigma_{\hat{W}}(t)) = -z_\alpha \sigma_{\hat{W}}(t) + \sigma_{\hat{W}}(t)\mathfrak{Z},$$
$$\text{where } \mathfrak{Z} = \mathcal{N}(0,1), \tag{27}$$

Therefore, for all $0 < t \leqslant T$, the TPoD of the nth model

$$\begin{aligned} p_n^{TPoD}(t,w) &= \mathrm{P}(V_n(t) > w) = \mathrm{P}(\sqrt{n}(V_n(t) - v(t)) \\ &\quad > \sqrt{n}(w - v(t))) = \mathrm{P}(\hat{V}_n(t) > 0) \to \mathrm{P}(\hat{V}(t) > 0) \\ &= \mathrm{P}(\mathfrak{Z} > 0) = 1 - \Phi(z_\alpha) = \alpha, \quad \text{as } n \to \infty, \end{aligned} \tag{28}$$

where the convergence holds by Theorem 5, the third equality holds because $v(t) = w$ according to Lemma 1, and the fourth equality holds by (27). Because the FCLT states a convergence in space $\mathbb{D}$ for the whole process $\hat{V}_n$ and the FCLT diffusion limit $\hat{V}$ has continuous paths (see Liu and Whitt 2014a), the convergence in the J_1 metric is equivalent to the uniform metric. Accordingly, we can claim the uniform convergence $\sup_{0<t\leqslant T} |p_n^{TPoD}(t,w) - \alpha| \to 0$. □

5.4. Proof of Theorem 3

Next we prove the convergence results in Theorem 3 for performance functions including the mean delay, mean queue length, utilization, PoD, and PoA.

Mean delay. Under the TTGA staffing in (23), we know from Theorem 5 and (27) that

$$V_n(t) = v(t) + \frac{1}{\sqrt{n}}\hat{V}(t) + o\left(\frac{1}{\sqrt{n}}\right) = w - \frac{z_\alpha}{\sqrt{n}}\sigma_{\hat{V}}(t) + \frac{1}{\sqrt{n}}\sigma_{\hat{V}}(t)\mathcal{Z} + o\left(\frac{1}{\sqrt{n}}\right).$$

Define $\tilde{V}_n(t) \equiv w - z_\alpha n^{-1/2}\sigma_{\hat{V}}(t) + n^{-1/2}\sigma_{\hat{V}}(t)\mathcal{Z}$. Then for any $\epsilon > 0$ and each $t \in (0, T]$, $\mathrm{P}(|\tilde{V}_n(t) - \tilde{V}_n(t)^+| > \epsilon) \to 0$ because $|\tilde{V}_n(t) - \tilde{V}_n(t)^+| > 0$ if and only if $\tilde{V}_n(t) < 0$ and $\mathrm{P}(\tilde{V}_n(t) < 0) = \Phi(z_\alpha - \sqrt{n}v(t)/\sigma_{\hat{V}(t)}(t)) \to 0$. Therefore, $\mathrm{P}(|\tilde{V}_n(t) - \tilde{V}_n(t)^+ + o(1/\sqrt{n})| > \epsilon) \to 0$. For $t = 0$, $V_n(t) - \tilde{V}_n(t)^+ = o(1/\sqrt{n})$ so $\mathrm{P}(V_n(t) - \tilde{V}_n(t)^+ > \epsilon) \to 0$. So $\mathrm{P}(|V_n(t) - \tilde{V}_n(t)^+| > \epsilon) \to 0$ for all t. Convergence in probability implies convergence in distribution, so $V_n(t) - \tilde{V}_n(t)^+ \Rightarrow 0$. To Prove $\mathrm{E}[|V_n(t) - \tilde{V}_n(t)^+|] \to 0$, it suffices to show the *uniform integrability* (u.i.) of $|V_n(t) - \tilde{V}_n(t)^+|$, which will follow if $\sup_n \mathrm{E}[(V_n(t) - \tilde{V}_n(t)^+)^2] < \infty$. We show this next.

$$\begin{aligned}
&\mathrm{E}[(V_n(t) - \tilde{V}_n(t)^+)^2] \\
&= \mathrm{E}[(V_n(t) - \tilde{V}_n(t)^+)^2 \mid \tilde{V}_n(t) \geqslant 0]\mathrm{P}(\tilde{V}_n(t) \geqslant 0) \\
&\quad + \mathrm{E}[(V_n(t) - \tilde{V}_n(t)^+)^2 \mid \tilde{V}_n(t) < 0]\mathrm{P}(\tilde{V}_n(t) < 0) \\
&= o\left(\frac{1}{n}\right)\mathrm{P}(\tilde{V}_n(t) \geqslant 0) + \mathrm{E}\left[\left(\tilde{V}_n(t) + o\left(\frac{1}{\sqrt{n}}\right)\right)^2 \middle| \tilde{V}_n(t) < 0\right] \\
&\quad \cdot \mathrm{P}(\tilde{V}_n(t) < 0) \leqslant 1 + o\left(\frac{1}{n}\right)\mathrm{P}(\tilde{V}_n(t) < 0) + 2o\left(\frac{1}{\sqrt{n}}\right) \\
&\quad \cdot \mathrm{E}[\tilde{V}_n(t) \mid \tilde{V}_n(t) < 0]\mathrm{P}(\tilde{V}_n(t) < 0) + \mathrm{E}[\tilde{V}_n(t)^2 \mid \tilde{V}_n(t) < 0] \\
&\leqslant 2 + \mathrm{E}[\tilde{V}_n(t)^2] = 2 + \mathrm{Var}(\tilde{V}_n(t)) + \mathrm{E}[\tilde{V}_n(t)]^2 \\
&= 2 + \frac{\sigma^2_{\hat{V}}(t)}{n} + \left(w - z_\alpha\frac{\sigma_{\hat{V}}(t)}{\sqrt{n}}\right)^2,
\end{aligned}$$

which attains its maximal value for some $n^* \geqslant 1$. Thus, we have obtained a finite bound

$$\sup_{n\geqslant 1}\sup_{0<t\leqslant T}\mathrm{E}[(V_n(t) - \tilde{V}_n(t)^+)^2] < C_V < \infty.$$

Mean queue length. Under TTGA staffing, we know from Theorem 5 that

$$\bar{Q}_n(t) = Q(t) + \frac{1}{\sqrt{n}}X_g(t) + \frac{1}{\sqrt{n}}\sigma_{\hat{X}}(t)\mathcal{Z} + o\left(\frac{1}{\sqrt{n}}\right).$$

Define $\tilde{Q}_n(t) \equiv Q(t) + n^{-1/2}X_g(t) + n^{-1/2}\sigma_{\hat{X}}(t)\mathcal{Z}$. Similar to the proof for the convergence of the mean delay, we have the uniform convergence $\sup_{0<t\leqslant T}|\bar{Q}_n(t) - \tilde{Q}_n(t)| \Rightarrow 0$ and a uniform bound for the purpose of u.i.

$$\begin{aligned}
&\sup_{n\geqslant 1}\sup_{0<t\leqslant T}\mathrm{E}[(\bar{Q}_n(t) - \tilde{Q}_n(t)^+)^2] \\
&\quad \leqslant \sup_{n\geqslant 1}\sup_{0<t\leqslant T}\left(2 + (\sigma^2_{\hat{X}}(t) + X^2_g(t))\frac{1}{n} + 2Q(t)X_g(t)\frac{1}{\sqrt{n}}\right) \\
&\quad < C_Q < \infty,
\end{aligned}$$

which concludes the convergence $\mathrm{E}[|\bar{Q}_n(t) - \tilde{Q}_n(t)^+|] \to 0$, as $n \to \infty$.

Utilization. For the system utilization $u(t)$, essentially we need to prove

$$\mathrm{E}\left[\frac{X_n(t)}{s_n(t)} \wedge 1 - \frac{\mathcal{N}(nX(t) + \sqrt{n}X_g(t), n\sigma_{\hat{X}}(t))^+}{s_n(t)} \wedge 1\right] \to 0, \quad \text{as } n \to \infty.$$

We know that $X_n(t) = nX(t) + \sqrt{n}X_g(t) + \sqrt{n}\sigma_{\hat{X}}(t)\mathcal{Z} + o(\sqrt{n})$ from Theorem 5. Define $\tilde{X}_n(t) \equiv nX(t) + \sqrt{n}X_g(t) + \sqrt{n}\sigma_{\hat{X}}(t)\mathcal{Z}$. Now we need to prove $(\tilde{X}_n(t) + o(\sqrt{n}))/s_n(t) \wedge 1 - \tilde{X}_n(t)^+/s_n(t) \wedge 1) \Rightarrow 0$. It is easy to see that under TTGA staffing, as $n \to \infty$, $\mathrm{P}(\tilde{X}_n(t) \geqslant s_n(t)) \to 1$ because the system is always overloaded. Also notice that $s_n(t)$ is of order n. So $\mathrm{P}((\tilde{X}_n(t) + o(\sqrt{n}))/s_n(t) \wedge 1 = 1) \to 1$ and $\mathrm{P}(\tilde{X}_n(t)^+/s_n(t) \wedge 1 = 1) \to 1$. It follows that $(\tilde{X}_n(t) + o(\sqrt{n}))/s_n(t) \wedge 1 - \tilde{X}_n(t)^+/s_n(t) \wedge 1 \Rightarrow 0$. Clearly, $\mathrm{E}[((\tilde{X}_n(t) + o(\sqrt{n}))/s_n(t) \wedge 1 - \tilde{X}_n(t)^+/s_n(t) \wedge 1)^2]$ is bounded by 1. Then we have the desired result.

PoD and PoA. We have shown that $V_n(t) \Rightarrow \tilde{V}_n(t)^+$, which implies that the PoD $\mathrm{P}(V_n(t) > 0) \sim \mathrm{P}(\tilde{V}_n(t)^+ > 0) = \mathrm{P}(\tilde{V}_n(t) > 0) = \mathrm{P}(\mathcal{Z} > (z_\alpha - \sqrt{n}v(t)/\sigma_{\hat{V}}(t)) = \Phi(\sqrt{n}w/\sigma_{\hat{V}}(t) - z_\alpha)$. Conditioning on the patience time A, the proof of the convergence of the PoA $\mathrm{P}(V_n(t) > A)$ is similar. □

6. Conclusion

We have developed an analytic staffing formula (dubbed TTGA) to stabilize the tail probability of delay across a wide range of performance targets in the $G_t/GI/s_t + GI$ model with a non-NHPP arrival process and time-varying arrival rate and nonexponential service and patience distributions. Simulation experiments show that TTGA successfully stabilizes the tail probability of delay for both large systems (having at least 100 servers) and small systems (having no more than 4 servers). In addition to stabilizing the tail probability of delay, we have provided effective approximating formulas for other important performance measures, such as the probability of delay and probability of abandonment, mean delay, and mean queue length. Supporting heavy-traffic limit theorems are developed to show that the TTGA staffing method stabilizes performance as the scale increases and that the performance approximating formulas are accurate as the scale increases, for the special case of exponential service times.

Future directions. Motivated by the Canadian emergency department example, we plan to extend our TTGA method to models having multiple customer classes, each of which has its own TPoD targets w_i and α_i. For example, the CTAS guideline (Example 3 of Section 1) specifies five patient sickness levels, which can be represented by five classes. We plan to do this

extension in two steps: First, we will determine the total service capacity, that is the overall staffing level for all customer groups; next, we will develop the time-dependent service allocation policy.

Acknowledgments

The author thanks the department editor Andrew Schaefer and two referees for their helpful comments. Special thanks to Rouba Ibrahim, Song-Hee Kim, and Ward Whitt for providing constructive comments and to Ph.D. student Beixiang He for conducting simulation experiments.

References

Aksin Z, Armony M, Mehrotra V (2007) The modern call center: A multi-disciplinary perspective on operations management research. *Production Oper. Management* 16(6):665–688.

Armony M, Israelit S, Mandelbaum A, Marmor Y, Tseytlin Y, Yom-Tov G (2014) *Patient Flow in Hospitals: A Data-Based Queueing-Science Perspective* (New York University, New York).

Brown L, Gans N, Mandelbaum A, Sakov A, Shen H, Zeltyn S, Zhao L (2005) Statistical analysis of a telephone call center: A queueing science perspective. *J. Amer. Statist. Assoc.* 100(469):36–50.

Bullard MJ, Chan T, Brayman C, Warren D, Musgrave E, Unger B (2014) Revisions to the Canadian emergency department triage and acuity scale (CTAS) guidelines. *Canadian Assoc. Emergency Physicians* 16(6):485–489.

Cram P, Hillis SL, Barnett M, Rosenthal GE (2004) Effects of weekend admission and hospital teaching status on in-hospital mortality. *Amer. J. Medicine* 117(3):151–157.

Defraeye M, 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.

Donahue K (2013) Nurses: It's about staffing at St. Joseph Hospital, not raises. *Times-Standard* (March 16).

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

Gans N, Koole G, Mandelbaum A (2003) Telephone call centers: Tutorial, review, and research prospects. *Manufacturing Service Oper. Management* 5(2):79–141.

Garnett O, Mandelbaum A, Reiman M (2002) Designing a call center with impatient customers. *Manufacturing Servce Oper. Management* 4(3):208–227.

Green LV, Kolesar PJ, Whitt W (2007) Coping with time-varying demand when setting staffing requirements for a service system. *Production Oper. Management* 16(1):13–39.

He B, Liu Y, Whitt W (2016) Staffing a service system with non-Poisson nonstationary arrivals. *Probab. Engrg. Inform. Sci.* 30(4): 593–621.

Jennings OB, Mandelbaum A, Massey WA, Whitt W (1996) Server staffing to meet time-varying demand. *Management Sci.* 42(10):1383–1394.

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

Li A, Whitt W, Zhang J (2016) Staffing to stabilize blocking in loss models with time-varying arrival rates. *Probab. Engrg. Inform. Sci.* 30(2):185–211.

Liu R, Kulh M, Liu Y, Wilson JR (2018) Modeling and simulation of nonstationary non-Poisson processes. *INFORMS J. Comput.* Forthcoming.

Liu Y (2017) Online appendix: Stabilizing customer abandonment in many-server queues with time-varying arrivals. North Carolina State University. https://yunanliu.wordpress.ncsu.edu/files/2014/02/StabilizeTPoDAppendix12162016.pdf.

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

Liu Y, Whitt W (2012b) The $G_t/GI/s_t + GI$ many-server fluid queue. *Queueing Systems* 71(4):405–444.

Liu Y, Whitt W (2012c) A many-server fluid limit for the $G_t/GI/s_t + GI$ queueing model experiencing periods of overloading. *Oper. Res. Lett.* 40(5):307–312.

Liu Y, Whitt W (2014a) Many-server heavy-traffic limit for queues with time-varying parameters. *Ann. Appl. Probab.* 24(1):378–421.

Liu Y, Whitt W (2014b) Stabilizing performance in network of queues with time-varying arrival rates. *Probab. Engrg. Inform. Sci.* 28(4):419–449.

Liu Y, Whitt W (2017) Stabilizing performance in many-server queues with time-varying arrivals and customer feedback. *Eur. J. Oper. Res.* 256(2):473–486.

Nelson B, Gerhardt I (2011) Modeling and simulating nonstationary arrival processes to facilitate analysis. *J. Simulation* 5(1):3–8.

SEE Center, Technion. (2014) SEEStat database. Accessed December 1, 2016, http://seeserver.iem.technion.ac.il/see-terminal/.

Shi P, Chou M, Dai JG, Ding D, Sim J (2014) Models and insights for hospital inpatient operations: Time-dependent ED boarding time. *Management Sci.* 62(1):1–28.

Stanton MW (2004) Hospital nurse staffing and quality of care. *Res. Action* 14:1–9.

Whitt W (2002) *Stochastic-Process Limits* (Springer, New York).

Whitt W, Zhao J (2017) Many-server loss models with non-Poisson time-varying arrivals. *Naval Res. Logist.* 64(3):177–202.

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

Zeltyn S, Mandelbaum A (2005) Call centers with impatient customers: Many-server asymptotics of the $M/M/n + G$ queue. *Queueing Systems* 51(3–4):361–402.

Yunan Liu is an associate professor in the Fitts Department of Industrial and Systems Engineering at North Carolina State University. His research interests include queueing theory, stochastic modeling, applied probability, simulations, and their applications in call centers, healthcare, transportation, and manufacturing systems.