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

OPERATIONS RESEARCH
Vol. 65, No. 2, March–April 2017, pp. 514–536
ISSN 0030-364X (print), ISSN 1526-5463 (online)

A Two-Time-Scale Approach to Time-Varying Queues in Hospital Inpatient Flow Management

J. G. Dai,[a] **Pengyi Shi**[b]

[a] School of Operations Research and Information Engineering, Cornell University, Ithaca, New York 14853; [b] Krannert School of Management, Purdue University, West Lafayette, Indiana 47907

Contact: jd694@cornell.edu (JGD); shi178@purdue.edu (PS)

Received: August 22, 2014
Revised: June 13, 2016
Accepted: August 5, 2016
Published Online in Articles in Advance: February 2, 2017

Subject Classifications: queues: nonstationary, algorithm; healthcare: hospitals
Area of Review: Stochastic Models

https://doi.org/10.1287/opre.2016.1566

Abstract. We analyze a time-varying $M_{\text{peri}}/\text{Geo}_{\text{2timeScale}}/N$ queueing system. The arrival process is periodic Poisson. The service time of a customer has components in different time scales: length of stay (LOS) in days and departure time (h_{dis}) in hours. This queueing system has been used to study patient flows from the emergency department (ED) to hospital inpatient wards. In that setting, the LOS of a patient is simply the number of days she spends in a ward, and her departure time h_{dis} is the discharge hour on the day of her discharge.

We develop a new analytical framework that can perform exact analysis on this novel queueing system. This framework has two steps: first analyze the midnight customer count process and obtain its stationary distribution, then analyze the time-dependent customer count process to compute various performance measures. We also develop approximation tools that can significantly reduce the computational time. In particular, via Stein's method, we derive explicit expressions to approximate the stationary distribution of the midnight count. We provide error bounds for these approximations and numerically demonstrate that they are remarkably accurate for systems with various sizes and load conditions. Our theoretical and numerical analysis have produced a number of insights that can be used to improve hospital inpatient flow management. We find that the LOS term affects the overnight wait caused by the mismatch between daily arrivals and discharges, whereas the h_{dis} term affects the intraday wait caused by the nonsynchronization between the arrival and discharge time patterns. Thus, reducing LOS or increasing capacity can impact the daily average performance significantly; shifting the discharge timing to earlier times of a day can alleviate the peak congestion in the morning and mainly affects the time-dependent performance.

Funding: This research is supported in part by National Science Foundation [Grants CNS-1248117, CMMI-1335724, and CMMI-1537795].
Supplemental Material: The e-companion is available at https://doi.org/10.1287/opre.2016.1566.

Keywords: two-time-scale • time-varying queue • time-dependent performance • approximation • Stein's method • hospital inpatient operations

1. Introduction

Customer demand for service is often time dependent in various service systems. As a result, queueing systems with time-varying arrival process, or *time-varying queues*, have been widely used to model call centers, healthcare delivery systems, and many other service systems (Green et al. 2007). In this paper, we study a time-varying queue with a novel *two-time-scale* service time model that is strongly motivated by a healthcare application.

Specifically, we study a single-customer-class, single-server-pool system (or *single-pool system* in short), denoted as an $M_{\text{peri}}/\text{Geo}_{\text{2timeScale}}/N$ system. This system has N identical, parallel servers in the server pool. Customers arrive at the system following a time-varying periodic Poisson process (denoted by "M_{peri}"). We assume the arrival rate function $\lambda(\cdot)$ to be periodic with period $T > 0$, i.e., $\lambda(t) = \lambda(t + T)$ for any $t \geqslant 0$. For ease of exposition, we use T as the time unit. In this paper we set the period $T = 1$ and interpret it as one day.

Upon a customer arrival, if there is an idle server, the customer is admitted into service immediately; otherwise, she waits in a buffer that can hold infinitely many waiting customers. Upon a customer departure from a server, the just-freed server takes a customer from the buffer following a first-come, first-served (FCFS) rule if the buffer is not empty; otherwise, the server becomes idle. Once a customer is admitted into service, she occupies the server for a duration of S until departing from the system. This service time, S, follows a *two-time-scale* model (denoted by "$\text{Geo}_{\text{2timeScale}}$"):

$$S = \text{LOS} + h_{\text{dis}} - h_{\text{adm}}. \tag{1}$$

Here, LOS (length-of-stay) denotes the *number of midnights* that the customer occupies a server (which equals the departure day minus the admission day), and $h_{\text{adm}} \in (0,1)$ and $h_{\text{dis}} \in (0,1)$ represent the time-of-day when the customer is admitted and departs the system, respectively. The departure hour h_{dis} is sometimes referred as *discharge hour*. Note that a customer's service time could be shorter than her LOS. For example, if the customer is admitted later today and departs early tomorrow (with $h_{\text{adm}} > h_{\text{dis}}$), her LOS equals one day according to our definition, but she spends less than 24 hours (one day) in service. Our definition of LOS is consistent with the medical conventions on measuring inpatient stays U.S. Centers for Disease Control and Prevention (2010); see Section 1.1 for more details on the hospital background. Mathematically speaking,

$$\begin{aligned} \text{LOS} = \lfloor T_{\text{dis}} \rfloor - \lfloor T_{\text{adm}} \rfloor, \quad h_{\text{dis}} = T_{\text{dis}} - \lfloor T_{\text{dis}} \rfloor, \\ h_{\text{adm}} = T_{\text{adm}} - \lfloor T_{\text{adm}} \rfloor, \end{aligned} \tag{2}$$

where T_{adm} and T_{dis} denote the admission time and departure time, respectively, and $\lfloor x \rfloor$ denotes the largest integer that is less than or equal to the given real number x. In this paper we assume the LOS of each customer follows a *geometric* distribution that takes values on $1, 2, \ldots$ and excludes 0 from the possible value list. We assume the success probability of this geometric distribution is $\mu \in (0,1)$; equivalently, we assume the mean LOS to be $m = 1/\mu$, which is strictly bigger than 1. We assume that LOS and h_{dis} of customers form two independent and identically distributed (i.i.d.) sequences, and the two sequences are independent of each other. The sequence of random variables h_{dis} follows a general distribution on the interval $(0,1)$. The rationale for the independence assumption between LOS and h_{dis} will be explained in Section 1.1.

The service time in our single-pool system is no longer an exogenous factor but depends on h_{adm} and two exogenous factors: the LOS and the h_{dis}. Since the admission hours h_{adm} are ordered for customers admitted on the same day, it follows from (1) that the service times are not i.i.d. This is different from the popular i.i.d. service time assumption used in most existing queueing systems. Note that LOS is in the order of days and h_{dis} is in the order of hours. Because of these two different time scales, we call the service times represented in (1) the *two-time-scale service time model*.

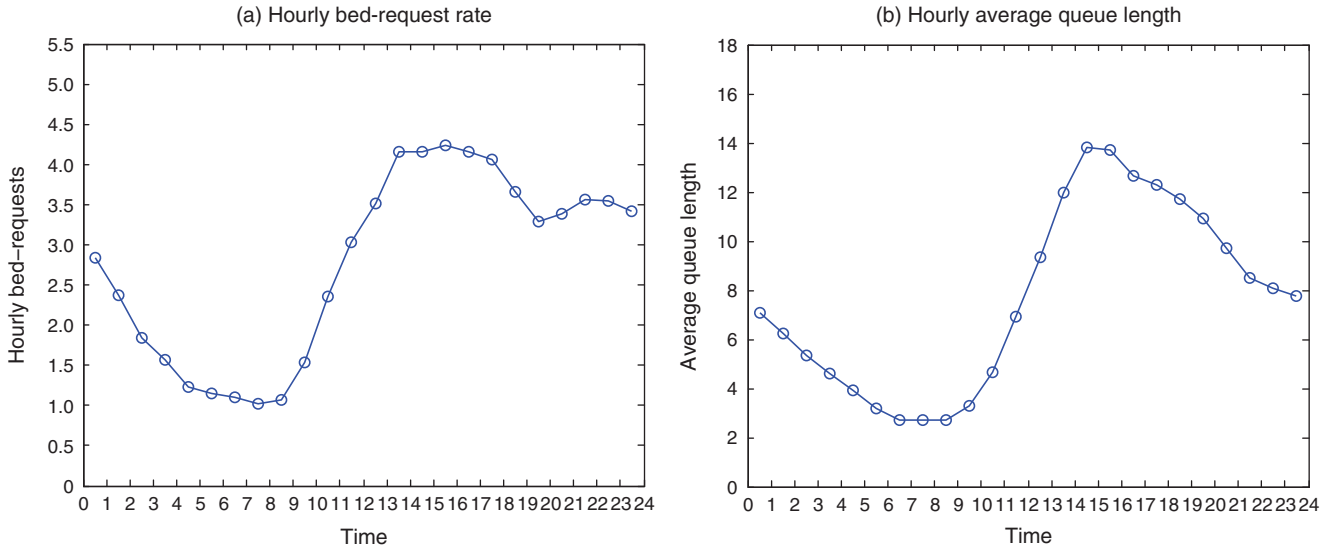
In this paper we analyze the $M_{\text{peri}}/\text{Geo}_{\text{2timeScale}}/N$ queueing system with a focus on predicting the steady-state time-dependent performance measures, including the time-dependent mean queue length $\mathbb{E}_\infty[Q(t)]$, time-dependent mean virtual waiting time $\mathbb{E}_\infty[W(t)]$, and time-dependent x-hour service level $\mathbb{P}_\infty(W(t) > x)$. We use $W(t)$ to denote the virtual waiting time at t, that is, the time that a virtual customer arriving at time t would have to wait until her service begins. In the rest of the paper, we use waiting time and virtual waiting time interchangeably when referring to $W(t)$. We use the subscript ∞ to denote the steady-state probability or expectation. These steady-state time-dependent performance measures are well defined given a certain initial distribution, which we will elaborate in Section 2.2.

1.1. Motivation to Study Systems with Two-Time-Scale Service Times

Our time-varying queueing system is motivated by studying patient flows from the emergency department (ED) to hospital inpatient wards (Shi et al. 2016, 2014). Patients who have received treatment in the ED and are waiting to be admitted to inpatient beds are modeled as customers, while the inpatient beds are modeled as servers. These patients are sometimes also referred as *boarding* patients (U.S. General Accounting Office 2003). The bed-request process of these patients naturally becomes the arrival process in our system. Empirical studies show that the bed-request process has time-varying rates (Shi et al. 2016, Armony et al. 2015b) and can be modeled by a time-nonhomogeneous Poisson process. See Figure 1(a) for an illustration of the time-varying rates using empirical data from a Singaporean hospital. Although we focus on the Poisson process setting, the methods developed in this paper allow analysis of systems with more general arrival processes that are not necessarily Poisson; see discussions in Section 5 of the electronic companion.

The service time of a customer models a patient's *inpatient stay*, the duration between her admission and discharge from the inpatient ward. It has been shown in Shi et al. (2016) that the service time taking the two-time-scale form in (1) is one of the critical features to capture hospital inpatient flow dynamics. Section 5.2 in Shi et al. (2016) discusses how the conventional $M_t/GI/N$ systems with i.i.d. service times fail to reproduce empirical performance at the hourly resolution. The rationale of this two-time-scale service time model is that a patient's inpatient stay is affected by different factors: first, the patient's medical conditions determine how many days she needs to spend in the hospital to recover, which is captured through the LOS term; then, when the patient is to be discharged on a day, the time-of-day h_{dis} when she leaves the hospital is driven by operational factors other than her medical conditions. Empirical studies show that in many hospitals, patients are "clustered" to leave in the afternoon, usually between 2 and 5 P.M.; see Figure 4(a) in Section 5.1 for an example. These common discharge patterns often result from the schedules and behaviors of medical staff such as physician's rounding time; see Armony et al. (2015b), Griffin et al. (2012), Chan

Figure 1. (Color online) Hourly bed-request rate and average queue length from empirical analysis using data from a Singaporean hospital (Shi et al. 2014, 2016)



et al. (2016) for relevant discussions. Thus, in the service time model (1), we use the LOS and the h_{dis} term, which are independent of each other, to decouple the impact of medical and operational factors on inpatient stays; see more empirical evidence supporting the independence assumption in Section 8.5 of Shi et al. (2014). Note that the empirical LOS distribution is usually *not* geometric. However, we find that the system performance is not very sensitive to the LOS distributions when the utilization is not extremely high; for example, the relative differences in the hourly mean waiting time between the geometric and empirical LOS distributions are less than 10% when $N = 505$ (with a 95% utilization); see Section 4.7 of Shi (2013). Therefore, we focus on the geometric LOS setting in this paper for tractability.

A customer's waiting time in our queueing system corresponds to the so-called ED *boarding time*, which is the duration between a patient's bed-request and admission to an inpatient bed. It is well known that ED boarding is a key contributor to ED overcrowding, a challenging problem faced by many healthcare systems worldwide (Bernstein et al. 2009, Hoot and Aronsky 2008). While waiting to be admitted, the boarding patients pose extra workloads on ED staff and can block new patients from receiving treatment in the ED. Moreover, prolonged boarding time can result in adverse patient outcomes (Liu et al. 2009, Singer et al. 2011) and lead to a significant increase in the hospital operational costs (Pines et al. 2011, Huang et al. 2010).

To alleviate ED boarding, as a first step, one needs to develop efficient tools to predict boarding-related performance. In particular, when the bed-request process has time-varying rates, it is not surprising to see that many empirical performance measures are also time-varying (Powell et al. 2011, Armony et al. 2015b, Shi et al. 2016). Figure 1(b) shows the time-dependent mean queue length from the empirical analysis in Shi et al. (2014), i.e., the average number of patients boarding in the ED at different times of a day. Predicting these performance measures allows us to understand how different factors such as the arrival pattern and discharge time affect the time-dependency, and then we can gain insights into the impact of various policies such as the discharge policy on the system performance.

1.2. Contributions

First, we are able to develop a simple analytical framework to perform *exact* analysis on the $M_{\text{peri}}/\text{Geo}_{\text{2timeScale}}/N$ system. This framework has two steps: (1) analyze the queueing dynamics day by day and obtain the stationary distribution of the *midnight* customer count process; (2) predict the time-dependent performance based on the midnight count distribution. The key advantage of this framework is to utilize the Markov property of the midnight customer count process so that we can predict the time-dependent performance from the most recent midnight without tracking all previous history. We call this framework a *two-time-scale framework*.

There are many existing methods developed for the conventional $M_t/GI/N$ systems, for which we give a detailed review in Section 1.4. However, it is difficult to apply these methods to our $M_{\text{peri}}/\text{Geo}_{\text{2timeScale}}/N$ system, not only because the service time model in our system is novel, but also because most of these existing methods rely on the assumption that the arrival

rate does not change drastically within a service completion (Green et al. 2007). This assumption is usually not valid in the hospital inpatient setting, since a typical patient service time is four to five days, during which the arrival process has gone through several cycles. Our two-time-scale analytical framework, however, explicitly captures the different time scales (day versus hours) raised in such settings and can predict the time-dependent performance when the service time is *extremely* long. The importance of capturing different time scales has also been discussed in other healthcare contexts (Armony et al. 2015b, Mandelbaum et al. 2012, Ramakrishnan et al. 2005, Zacharias and Armony 2016).

Second, besides the exact analysis, we develop efficient approximation tools to analyze the queueing system. For step (1) in the two-time-scale framework, we apply the Stein's method, developed in Braverman and Dai (2017) for steady-steady diffusion approximation of continuous-time queueing systems, to the discrete-time midnight count process in our setting. This method allows us to identify a continuous distribution with an explicit formula to approximate the stationary distribution of the midnight count and to establish an error bound for the approximation. In particular, we find that the error bound becomes small when the number of servers N and the mean LOS $1/\mu$ are *both* large. For step (2), we develop normal approximations to predict the time-dependent performance and establish Berry-Esseen type bounds. We demonstrate, via numerical experiments, that these approximation tools are remarkably accurate in predicting various performance (even when the system size is less than 20), but they require much less computational time. For example, normal approximations typically result in less than 1% relative error in predicting the time-dependent mean waiting time curve but only require a few minutes to compute the curve, whereas the exact analysis needs days to compute the same curve.

Third, through both the exact analysis and approximations, we quantify the impact of different policies and gain insights into the $M_{\text{peri}}/\text{Geo}_{\text{2timeScale}}/N$ system. In particular, we understand how the two terms in the service time model, LOS and discharge time h_{dis}, affect the customer wait and impact the system performance in different ways. We classify the customer wait into two types: (i) *overnight wait* occurs when there is a mismatch between the daily number of arrivals and discharges so that a fraction of customers need to wait overnight, and (ii) *intraday wait* occurs when the arrival and discharge patterns are nonsynchronized so that the morning arrivals need to wait until the afternoon when most discharges occur. We find that reducing LOS or equivalently, increasing capacity can reduce the daily mismatch and thus reduce the overnight wait, whereas shifting h_{dis} to earlier times of a day (early discharge) does not affect the overnight wait but can reduce the intraday wait by eliminating the nonsynchronization; see Section 3.1. As a result, increasing capacity mainly impacts the system daily performance, and early discharge mainly impacts the time-dependent performance. When the system load is high and most customers experience overnight wait, early discharge brings a very limited impact. We confirm these insights via numerical experiments in Section 5.

As a remark, caution should be taken when applying these insights to a hospital setting because the $M_{\text{peri}}/\text{Geo}_{\text{2timeScale}}/N$ system studied in this paper is limited by having a single pool of servers and other simplified assumptions, and we do not expect it can fully capture the actual inpatient flows. As demonstrated in Shi et al. (2016), to accurately replicate empirical performance curves such as the one in Figure 1(b), a high-fidelity hospital model needs to incorporate not only the two-time-scale service time but also other important features including multiple pools of servers and *allocation delays* caused by secondary bottlenecks other than bed unavailability; the latter two features are not considered in this paper. Nevertheless, the insights gained from the single-pool system in this paper are consistent with those discovered in Shi et al. (2016), where the authors simulated the high-fidelity model to evaluate the impact of early discharge and capacity increase. The analytical framework and the two types of wait found in this paper provide a systematic way to explain those simulation findings. The computational efficiency of our developed approximations also allows one to use the single-pool system to identify a range of policies that satisfy certain desired objectives prior to conducting a full-scale simulation. More importantly, we believe this paper represents an important first step toward building a *framework* to analyze time-varying systems with the two-time-scale service times, and it sets the stage for subsequent research to analyze high-fidelity hospital models with more realistic features such as multiple pools of servers and allocation delays.

1.3. Outline and Conventions

The remainder of this paper is organized as follows. In Section 1.4 we review the relevant literature. In Section 2 we demonstrate the basic idea of the two-time-scale analytical framework to analyze the single-pool system. In Section 3 we use this framework to derive structure properties for the single-pool system. In particular, we classify the patient wait into two types and characterize the impact of early discharge on the waiting time. In Section 4 we develop approximation tools to efficiently compute the midnight count stationary distribution and time-dependent performance measures. In Section 5 we show numerical results from analyzing the single-pool system and summarize insights. Finally, we conclude this paper in Section 6.

This paper uses (i) servers and beds, (ii) customers and patients, (iii) arrival and bed-request, and (iv) departure and discharge, interchangeably. For notational simplicity, we assume that there is no arrival or discharge at the exact point of midnight each day.

1.4. Literature Review

Many works have developed methods to analyze time-varying queues under the conventional $G_t/GI/N$ framework with i.i.d. service times. In the simplest $M_t/M/N$ setting, one can directly solve the Chapman-Kolmogorov forward equations (a system of ordinary differential equations) to obtain exact numerical solutions, which is usually computationally intensive but can serve as a benchmark (Green and Kolesar 1991). To reduce the computational time and to analyze more general service time distributions, one often resorts to approximations. The commonly used approximate approaches include closure approximation (Rothkopf and Oren 1979, Clark 1981), pointwise stationary approximation (PSA) (Green and Kolesar 1991, Whitt 1991), lagged PSA (Green and Kolesar 1997), modified offered-load approximation (Massey and Whitt 1994, Yom-Tov and Mandelbaum 2014), infinite-server approximation (Jennings et al. 1996), and iteration algorithms (Choudhury et al. 1997, Feldman et al. 2008). See Green et al. (2007), Ingolfsson et al. (2007) for comprehensive surveys and comparisons on these approximation methods. Recently, Liu and Whitt (2011c, 2012a, b) developed fluid models that can alternate between over- and under-loaded regimes to approximate time-varying queues in a general $G_t/GI/s_t+GI$ setting ($+GI$ denotes customer abandonment). As mentioned in Section 1.2, these methods do not apply well to analyze our $M_{\text{peri}}/\text{Geo}_{\text{2timeScale}}/N$ system.

The two works most closely related to our analytical framework are Ramakrishnan et al. (2005) and Powell et al. (2011). The discrete-time Markov chain (DTMC) considered in Ramakrishnan et al. (2005) is similar to the one we develop for the midnight count process; see Section 2.1. However, Ramakrishnan et al. (2005) focused on operations *within* the ED and used the DTMC to support calculations of ED-related performance such as ED visit blocking probability. The model studied in our paper is motivated by the patient flow from ED to the inpatient wards, and we focus on understanding the impact of *inpatient* discharge timing and bed capacity on the boarding time performance. Powell et al. (2011) developed a deterministic fluid-type model to predict the mean hourly customer count, which is similar to (9) in Section 2.2. The authors of Powell et al. (2011) assumed that all servers are occupied at 8 A.M. each day. Because their model did not incorporate random fluctuations, it cannot predict performance that needs the information of the entire distribution of the hourly customer count; for example, the time-dependent mean waiting time and x-hour service level.

Finally, the approximations we develop in Section 4 use the assumption that the number of servers N follows the square-root safety staffing rule, which is known to lead systems operating in the quality-and-efficiency-driven (QED) regime. The QED regime was first mathematically formalized in Halfin and Whitt (1981) and has been widely considered in call center research (Gans et al. 2003). Recent studies (Armony et al. 2015b, Mandelbaum et al. 2012) have justified the relevance of QED regime in hospital inpatient operations, where the number of beds is large, the average bed utilization is high (e.g., more than 90%), while the mean waiting time is only a small proportion of the mean service time. See Mandelbaum et al. (2012), Yom-Tov and Mandelbaum (2014), Armony et al. (2015a) for some examples on analyzing systems in the QED regime that are motivated by hospital operations. Moreover, the Stein's method we use in Section 4.3 to approximate the midnight count stationary distribution is based on the framework initiated in Gurvich (2014) and later systematically developed in Braverman and Dai (2017); these two works considered steady-state approximations of many-server queues in call center operations.

2. A Two-Time-Scale Approach for the Single-Pool System

In this section we introduce a two-time-scale analytical framework to analyze the $M_{\text{peri}}/\text{Geo}_{\text{2timeScale}}/N$ system. As mentioned in the introduction, this framework has two steps: first analyze the midnight customer count process, and then analyze the time-dependent customer count based on the midnight count. In Sections 2.1 and 2.2 we describe these two steps. Once we get the stationary distributions of the midnight and time-dependent customer counts, we can compute various time-dependent performance measures, and we demonstrate how to do so in Section 2.3.

2.1. A Two-Time-Scale Approach: Step 1, Midnight Dynamics

Let X_k denote the number of customers in the system at the midnight (zero hour) of day k, i.e., the *midnight customer count*. Let A_k and D_k denote the total number of arrivals and discharges within day k, respectively. The relationship between X_k and X_{k+1} is

$$X_{k+1} = X_k + A_k - D_k, \quad k = 0, 1, \ldots, \tag{3}$$

and this leads to the following proposition.

Proposition 1. *The midnight customer count process* $\{X_k, k = 0, 1, \ldots\}$ *is an irreducible discrete-time Markov chain*

(DTMC) on state space $\mathbb{Z}_+ = \{0, 1, \ldots, \}$. *This DTMC is positive recurrent with a unique stationary distribution* π *if*

$$\rho = \frac{\Lambda}{N\mu} < 1, \tag{4}$$

where

$$\Lambda = \int_0^1 \lambda(t)\, dt, \tag{5}$$

and $\lambda(\cdot)$ *is the periodic arrival rate function.*

Proof. We first show that $\{X_k, k = 0, 1, \ldots\}$ is a DTMC. Under our arrival process assumption, A_k is a Poisson random variable with mean Λ. Since there is no same-day discharge (LOS is at least one day), the number of discharges within day k, D_k only depends on the number of customers admitted before day k. Recall that LOS follows a geometric distribution, which can be seen as the number of independent coin tosses needed to get the first success. Thus, it is equivalent to think that we toss a coin for each busy server at the midnight of day k to determine whether its customer being served leaves on day k or not. Consequently, D_k is the sum of the outcomes of these coin tosses, and thus it follows a binomial distribution with parameters $(z(n), \mu)$ when conditioning on $X_k = n$, where

$$z(n) = \min(n, N) \tag{6}$$

is the number of busy servers at the midnight of day k. This coin-toss argument is elaborated in Section 1.1 of the electronic companion, where we construct a revised system with the coin toss scheme mentioned above and we prove that this revised system is equivalent to the original system in distribution. Also, see a precise representation of D_k in Section 1.2 of the electronic companion.

As a result, we know that D_k only depends on X_k, and A_k is independent of (X_k, D_k). Then, from (3) we can see that $\{X_k : k = 0, 1, \ldots\}$ forms a DTMC.

This Markov chain is irreducible. Moreover, we can prove that it is positive recurrent under condition (4) by checking that the Foster-Lyapunov criterion holds (Bramson 2008), i.e., when $x > N$, we have

$$\mathbb{E}[X_{k+1} - X_k \mid X_k = x] = \mathbb{E}[A_k - D_k \mid X_k = x] = \Lambda - N\mu < 0. \quad \square$$

From the above argument, the transition probability from state i to state j for the DTMC is

$$P_{ij} = \sum_{k=(i-j)^+}^{z(i)} g_1(z(i), k) \cdot f_1(k + j - i) \quad \text{for } i, j \in \mathbb{Z}_+. \tag{7}$$

Here, $f_1(k) = (\Lambda^k/k!)e^{-\Lambda}$ is the probability mass function (pmf) at point k for a Poisson distribution with mean Λ, $g_1(i, k) = (i!/(k!(i-k)!))\mu^k(1-\mu)^{i-k}$ is the pmf at point k for a binomial distribution with parameters (i, μ), $a^+ = \max(a, 0)$ for $a \in \mathbb{R}$, and k starting from $(i-j)^+$ ensures $k + j - i \geqslant 0$.

Under condition (4), the stationary distribution π, viewed as a row vector, is the unique solution to

$$\pi P = \pi, \quad \pi_k \geqslant 0, \quad \text{and} \quad \sum_k \pi_k = 1, \tag{8}$$

where $P = (P_{ij})$ is an infinite size, irreducible stochastic matrix. We numerically compute the vector π by appropriately truncating the matrix P to a finite matrix whose size depends on the load condition ρ. One may also apply the transform method to solve π as shown in Gao et al. (2004).

2.2. A Two-Time-Scale Approach: Step 2, Time-of-Day Dynamics

For $t \geqslant 0$, let $X(t)$ be the total number of customers in the system at time t, i.e., the *time-dependent customer count*. Similar to (3), for $t \geqslant 0$, $X(t)$ can be expressed as

$$X(t) = X(0) + A_{(0,t]} - D_{(0,t]}, \tag{9}$$

where $A_{(0,t]}$ denotes the cumulative number of arrivals in the period $(0, t]$, and $D_{(0,t]}$ denotes the cumulative number of discharges in the period $(0, t]$. Since the arrival and discharge occurring at time t (if any) are included in $A_{(0,t]}$ and $D_{(0,t]}$, respectively, $X(\cdot)$ is right continuous. When $X(0) = X_0$, this continuous customer count process $X(\cdot)$ coincides with the midnight count process at integer points, i.e., $X(k) = X_k$ for $k = 0, 1, \ldots$.

We first state the following proposition on the periodicity of the customer count process $X = \{X(t), t \geqslant 0\}$, the queue length process $Q = \{Q(t), t \geqslant 0\}$, and the virtual waiting time process $W = \{W(t), t \geqslant 0\}$. Here, for a given $t \geqslant 0$, the queue length and virtual waiting time are defined as

$$Q(t) = (X(t) - N)^+, \tag{10}$$

$$W(t) = \inf_{x \geqslant 0}\{D_{(t,t+x]} > X(t) - N\}. \tag{11}$$

Proposition 2. *When* $X(0)$ *follows the stationary distribution* π, *each of the processes* X, Q, *and* W *is periodic in distribution with one day as a period.*

Proof. See Section 1.4 of the electronic companion. $\square$

This proposition makes it sufficient for us to focus on the dynamics of $X(t)$, $Q(t)$, and $W(t)$ for $t \in [0, 1)$. Moreover, it indicates that the system is in a *periodic steady state* (PSS) given that the distribution of $X(0)$ is π (see the formal definition and more discussions on PSS in Liu and Whitt 2011a, b). Thus, in the remainder of this paper we assume the system is in such a PSS when we refer to the stationary distribution of $X(t)$, $Q(t)$, or $W(t)$, and we use $\pi(n) = \mathbb{P}_\infty(X(0) = n)$ to denote the initial distribution.

Using (9) and the convolution technique, we obtain the stationary distributions of $X(t)$ and $Q(t)$ for any given $t \in [0,1)$ as in Proposition 3 below. The stationary distribution of $W(t)$ is more complicated, and we leave the details to Section 2.3.

Proposition 3. *Fix $0 \leqslant t < 1$. For $m \in \mathbb{Z}_+$,*

$$\mathbb{P}_\infty(X(t)=m)=\sum_{n=0}^{\infty}\left(\sum_{k=(n-m)^+}^{z(n)} f_t(k+m-n)g_t(z(n),k)\right)\pi(n). \tag{12}$$

$$\mathbb{P}_\infty(Q(t)=m)=\mathbb{P}_\infty(X(t)=m+N). \tag{13}$$

Here, $z(n)$ is defined in (6),

$$f_t(k)=\frac{(\Lambda G(t))^k}{k!}e^{-\Lambda G(t)} \quad \text{and} \qquad g_t(i,k)=\frac{i!}{k!(i-k)!}(\mu H(t))^k(1-\mu H(t))^{i-k} \tag{14}$$

are the pmf for a Poisson distribution with mean $\Lambda G(t)$ and the pmf for a binomial distribution with parameters $(i,\mu H(t))$ evaluated at point k, respectively, and

$$G(t)=\frac{1}{\Lambda}\int_0^t \lambda(s)\,ds, \quad H(t)=\mathbb{P}(h_{\text{dis}}\leqslant t) \quad \text{for } 0\leqslant t<1. \tag{15}$$

We call $G(\cdot)$ and $H(\cdot)$ the cumulative distribution functions (cdf) associated with the arrival rate function $\lambda(\cdot)$ and the discharge time h_{dis}, respectively. To prove this proposition, we need the following lemma, which will be proved in Section 1.2 of the electronic companion using the revised coin-toss system.

Lemma 1. *For any $t \geqslant 0$, conditioning on $X(k_t)=n$, $D_{(k_t,t]}$ follows a binomial distribution with parameters $(z(n),\mu H(t-k_t))$, where $k_t=\lfloor t \rfloor$.*

Proof of Proposition 3. Since $Q(t)=(X(t)-N)^+$ for each $t \geqslant 0$, (13) follows immediately from (12). When $0 \leqslant t < 1$, $k_t = 0$. Applying Lemma 1 and the fact that $A_{(0,t]}$ follows a Poisson distribution with mean $\Lambda G(t)$ and is independent of $(X(0), D_{(0,t]})$, we have

$$\begin{aligned}
&\mathbb{P}_\infty(X(t)=m \mid X(0)=n)\\
&=\sum_{k=(n-m)^+}^{z(n)} \mathbb{P}(A_{(0,t]}=k+m-n \mid D_{(0,t]}=k, X(0)=n)\\
&\quad\cdot \mathbb{P}(D_{(0,t]}=k \mid X(0)=n)\\
&=\sum_{k=(n-m)^+}^{z(n)} f_t(k+m-n)g_t(z(n),k),
\end{aligned}$$

from which (12) follows. Here, k starts from $(n-m)^+$ to ensure $k+m-n \geqslant 0$. □

2.3. Predicting Time-Dependent Queue Length and Waiting Time Performance

We focus on predicting the following steady-state time-dependent performance measures: mean queue length $\mathbb{E}_\infty[Q(t)]$ (similar to the curve in Figure 1(b)), mean waiting time $\mathbb{E}_\infty[W(t)]$, and six-hour service level $\mathbb{P}_\infty(W(t) > 6/24)$. Note that we pick this particular service level ($x = 6$ hours) because it is an important performance measure monitored in hospitals (Shi et al. 2016). However, all the methodologies developed below can be applied to a general x-hour service level. We focus on predicting these performance measures for $t \in [0,1)$ because of the periodicity shown in Proposition 2.

We can compute the mean queue length $\mathbb{E}_\infty[Q(t)]$ from the stationary distribution of $Q(t)$ given by (13). For the mean waiting time and the six-hour service level, the key step is to evaluate the probability $\mathbb{P}_\infty(W(t) > x)$ for a given $x \in \mathbb{R}_+$. Once we have this probability, we obtain the mean waiting time by

$$\mathbb{E}_\infty[W(t)]=\int_0^\infty \mathbb{P}_\infty(W(t)>x)\,dx. \tag{16}$$

The six-hour service level $\mathbb{P}_\infty(W(t) > 6/24)$ is trivial by letting $x = 6/24$.

Next, we evaluate $\mathbb{P}_\infty(W(t) > x)$. We first state the following proposition for this probability when $0 \leqslant t + x < 1$.

Proposition 4. *For a given $t \in [0,1)$ and $x \in \mathbb{R}_+$, if $0 \leqslant t+x<1$,*

$$\begin{aligned}
&\mathbb{P}_\infty(W(t)>x)\\
&\quad=\sum_{n=0}^{\infty}\left(\sum_{a=(N-n)^+}^{\infty} J_{t+x}(z(n),a+n-N)f_t(a)\right)\pi(n),
\end{aligned} \tag{17}$$

where $f_t(a)$ is defined in (14), and $J_{t+x}(i,k)$ is the cdf of a binomial distribution with parameters $(i,\mu H(t+x))$ evaluated at point k.

We give an outline of the proof for Proposition 4 below. We leave the complete proof to Section 1.3 of the electronic companion, where a more general version of the proposition will be stated and proved. The general version covers all cases of $t + x \geqslant 0$.

From (11), we rewrite $\mathbb{P}_\infty(W(t) > x)$ as the following:

$$\begin{aligned}
&\mathbb{P}_\infty(W(t)>x)\\
&=\mathbb{P}_\infty(X(0)+A_{(0,t]}-N \geqslant D_{(0,t+x]}) &(18)\\
&=\sum_{n=0}^{\infty}\mathbb{P}_\infty(D_{(0,t+x]}-A_{(0,t]}\leqslant n-N \mid X(0)=n)\pi(n). &(19)
\end{aligned}$$

Intuitively, (18) results from the fact that if a virtual customer arriving at time t still has to wait at time $t+x$, then the cumulative number of discharges from 0 to time $t+x$, $D_{(0,t+x]}$, cannot clear the queue in front of

this customer. It is easy to check that this queue length equals $(X(0)-N)+A_{(0,t]}$. Once we have (18), the rest of proof for Proposition 4 is similar to the proof of (12), since when $0 \leqslant t+x<1$, it follows from Lemma 1 that conditioning on $X(0)=n$, $D_{(0,t+x]}$ follows a binomial distribution with parameters $(z(n), \mu H(t+x))$.

Note that (17) does *not* apply to the scenarios when $t+x \geqslant 1$. The reason is that when $t+x$ is large, $D_{(0,t+x]}$ is the total number of discharges among multiple days, and thus it becomes the sum of multiple random variables. As a result, we need to use more levels of convolution to evaluate $\mathbb{P}_\infty(W(t)>x)$ for $t+x \geqslant 1$; see Section 1.3 of the electronic companion for the details.

3. Impact of Shifting the Discharge Time

The unique feature of our single-pool system is the two-time-scale service time model. The two exogenous terms in the service time representation (1), LOS and h_{dis}, are on two different time scales. A major task in this paper is to explore how these two terms affect the system performance. In this section we focus on the discharge time h_{dis} and use the analytical framework in Section 2 to investigate how changing h_{dis} affects the waiting time performance. First, in Section 3.1, we classify a customer's wait into two types: intraday wait and overnight wait, and show that changing h_{dis} only affects the intraday wait. Then, in Section 3.2, we further quantify the impact of shifting h_{dis} to earlier times of the day on the time-dependent waiting time $W(t)$.

Note that the impact of h_{dis} on the mean queue length can be easily understood from (9), and we will demonstrate this impact via numerical examples in Section 5. For the impact of the LOS term (or equivalently, the capacity N), we also leave the detailed discussion to Section 5.

3.1. Two Types of Wait

Consider a virtual customer arrives at time $t \in [0,1)$. The probability that this virtual customer needs to wait overnight is $\mathbb{P}_\infty(W(t)>1-t)$. Applying (18) with $x=1-t$, we get

$$\mathbb{P}_\infty(W(t)>1-t)=\mathbb{P}_\infty(X(0)+A_{(0,t]}-N \geqslant D_{(0,1]}). \quad (20)$$

The above says that if $X(0)+A_{(0,t]}-N \geqslant D_{(0,1]}$, this virtual customer cannot be admitted on the same day of her arrival, and she has to wait until the next day or even later to be admitted. We call the customer experiences an *overnight wait* in this situation.

If $X(0)+A_{(0,t]}-N<D_{(0,1]}$, this customer can be admitted on the same day of arrival: she is either (i) admitted immediately upon arrival if $X(t)=X(0)+A_{(0,t]}-D_{(0,t]}<N$, or (ii) admitted at $t+w$ for some $0<w \leqslant 1-t$, where w is the first time such that $X(0)+A_{(0,t]}-D_{(0,t+w]}<N$. We call the customer experiences an *intraday wait* in the latter case (ii).

Recall that $X(0)$ and $D_{(0,1]}$ depend on Λ, N, and the mean LOS $1/\mu$, but *not* on the discharge time h_{dis}. As a result, (20) indicates that changing h_{dis} does not affect the overnight wait probability $\mathbb{P}_\infty(W(t)>1-t)$ for each $t \in [0,1)$. Consequently, we have the following proposition.

Proposition 5. *The average fraction of customers experiencing overnight wait*

$$\int_0^1 \mathbb{P}_\infty(W(t)>1-t)\,dt$$

does not depend on h_{dis}.

Although not affecting the overnight wait, changing h_{dis} can affect the intraday wait, because $D_{(0,t]}$ depends on the discharge time distribution. Indeed, we can prove that an extreme "midnight discharge" distribution (customers are discharged at the beginning of each day) can fully eliminate the intraday wait: with probability 1, each customer is either admitted without any delay or she has to wait overnight. However, this midnight discharge distribution is hardly achievable in practice. Thus, in the following section we consider shifting a given discharge distribution h hours earlier.

3.2. The Impact of Shifting the Discharge Distribution

In this section we show the impact of shifting the discharge distribution h hours earlier on the waiting time $W(t)$. For a fixed h such that $0<h<24$, we impose the following condition on the given discharge distribution with the cdf $H(\cdot)$ defined in (15):

$$H(h/24)=0. \quad (21)$$

Condition (21) says that no discharge occurs between midnight and hour h under the given discharge distribution $H(\cdot)$. Under this condition, shifting $H(\cdot)$ by h hours earlier would not result in customers leaving one night earlier, and thus the LOS would not be affected. Condition (21) is reasonable because in practice h can only be a few hours (typically between zero and four according to our communications with hospital managers), and rarely does a patient discharge between midnight and 4 A.M. As shown in Shi et al. (2014), after an "early discharge campaign" in a Singaporean hospital, a new discharge peak emerged around 11 A.M. to noon, three hours earlier than the old 2–3 P.M. peak.

Let $W^{(-h)}(t)$ denote the virtual waiting time at t under the h-hour shifted discharge distribution $H^{(-h)}(\cdot)$, where

$$H^{(-h)}(t)=\begin{cases} H(t+h/24) & \text{for } t \in [0,1-h/24), \\ 1 & \text{for } t \in [1-h/24,1). \end{cases} \quad (22)$$

The following proposition connects the distributions of $W(t)$ and $W^{(-h)}(t)$.

Proposition 6. *Fix an h such that $0 < h < 24$. Under assumption (21), for a given $t \in [0,1)$ and $x \geqslant h/24$,*

$$\mathbb{P}_\infty(W^{(-h)}(t) > x - h/24) = \mathbb{P}_\infty(W(t) > x). \tag{23}$$

This proposition is intuitive, and we leave its proof to Section 1.5 of the electronic companion. Setting $x = (6+h)/24$ in (23), we have the following corollary.

Corollary 1. *Under assumption (21), for $t \in [0,1)$, if $W(t)$ does not take values between 6 hours and $6+h$ hours, i.e.,*

$$\mathbb{P}_\infty(6/24 < W(t) \leqslant (6+h)/24) = 0, \tag{24}$$

then

$$\mathbb{P}_\infty(W^{(-h)}(t) > 6/24) = \mathbb{P}_\infty(W(t) > 6/24).$$

This corollary says that shifting the discharge distribution h hour earlier has no impact at all on the six-hour service level for time t that satisfies condition (24). This condition is satisfied, for example, when $h = 1$ hour and no discharge occurs between $t + 6/24$ and $t + 7/24$, in which case the waiting time is either less than six hours or longer than seven hours for customers arriving at t.

Next, we use Proposition 6 to show the following property on the mean waiting time $\mathbb{E}[W(t)]$.

Proposition 7. *Fix an h such that $0 < h < 24$. Assume (21) holds and*

$$\mathbb{P}_\infty(0 < W(t) \leqslant h/24) = 0. \tag{25}$$

Then,

$$\mathbb{E}_\infty[W(t)] - \mathbb{E}_\infty[W^{(-h)}(t)] = \frac{h}{24}\mathbb{P}_\infty(W(t) > 0), \tag{26}$$

where $W^{(-h)}(t)$ is again the waiting time at t under the discharge distribution $H^{(-h)}(\cdot)$ in (22).

Condition (25) is similar to (24); the discussions surrounding (24) can be extended similarly to (25). Proposition 7 says that under conditions (21) and (25), the reduction in the mean waiting time is *linear* in $h/24$ when the discharge distribution is shifted h hours earlier. The slope of the reduction is given by $\mathbb{P}_\infty(W(t) > 0)$, the delay probability at t. Even when condition (25) is violated, we use numerical examples in Section 5.3 to show that the linear reduction may still approximately hold during certain times of a day. The proof of Proposition 7 is in Section 1.6 of the electronic companion.

4. Efficient Numerical Algorithms to Compute Time-Dependent Performance

In this section we explore approximation tools to devise efficient numerical algorithms to analyze the single-pool system. We develop normal approximations to compute the time-dependent customer count distribution and time-dependent waiting time performance in Sections 4.1 and 4.2, respectively. We use Stein's method to approximate the stationary distribution of the midnight customer count process in Section 4.3.

Our motivation for exploring these approximation tools stems from two aspects. First, the exact analysis from Section 2 is not efficient. When the midnight count stationary distribution π is given, to numerically evaluate $\mathbb{E}_\infty[W(t)]$ using (16), we need to evaluate $\mathbb{P}_\infty(W(t) > x)$ for a number of points x (to ensure the accuracy of the integral), which involves two to four summations depending on the values of $t + x$. When $N = 500$, just to compute $\mathbb{E}[W(t)]$ for one given t requires several hours. Furthermore, when N is large and the utilization ρ is close to 1, getting π from the Markov chain analysis (8) also becomes computationally intensive, because generating the entries in the transition matrix P requires calculating the convolution between the daily number of arrivals and the daily number of discharges, where the latter depends on the current number of customers in system; see the form of g_1 in (7). For a system with $N = 977$ and $\rho = 98\%$, which is a realistic setting for a hospital, it requires 1.014 hours to get π (using a truncation of 3,000 for the matrix P). In contrast, using the approximations developed in this section, we just need the first two moments of the daily and hourly number of arrivals and discharges. For $N = 500$, we need less than three minutes to produce the entire mean waiting time curve (the same curve needs days of computation with the exact analysis). Second, in addition to the computational advantages, the explicit formulas we derive for these approximations allow us to express π and other performance measures in closed form, which makes it potentially easier for practitioners to use our results and could generate direct insights into how various parameters affect the system performance. Also see Section 4 of the electronic companion for a detailed complexity analysis and description of the computer platform used for our computations.

4.1. Normal Approximation for the Distribution of the Time-Dependent Customer Count

We first propose an approximation for the distribution of $X(t)$ in Section 4.1.1, then show numerical results on the accuracy of the approximation in Section 4.1.2, and establish an error bound for the approximation in Section 4.1.3.

4.1.1. Approximation Formulas. From Section 2.2, the time-dependent customer count $X(t)$ at time t ($0 \leqslant t < 1$) has representation (9).

We propose the following Approximation 1 for the stationary distribution of $X(t)$. The intuition is to replace $A_{(0,t]}$ and $D_{(0,t]}$ by two normal random variables, so that their difference is still a normal random variable.

Recall that $z(n) = \min(n, N)$ is defined in (6), and we define

$$M(t,m,n) = \frac{0.5 + m - (n + \Lambda G(t) - z(n)\mu H(t))}{\sqrt{\Lambda G(t) + z(n)\mu H(t)(1-\mu H(t))}}, \quad (27)$$

where $G(\cdot)$ and $H(\cdot)$ are the cdf associated with the arrival rate and discharge time given in (15).

Approximation 1. *Fix $t \in [0,1)$. For $0 \leqslant m < \infty$,*

$$\mathbb{P}_\infty(X(t) \leqslant m) \approx \sum_{n=0}^{\infty} \Phi(M(t,m,n))\pi(n), \quad (28)$$

where $\Phi(\cdot)$ is the cdf of the standard normal distribution, and $\pi(\cdot)$ is the stationary distribution of the midnight count process.

Because μ is assumed to be in $(0,1)$, $1-\mu H(t) > 0$ for all $t \in [0,1)$. However, it is possible that $G(t) + H(t) = 0$ or $G(t) + n = 0$, resulting the denominator of M in (27) being zero. In such cases, we interpret $M(t,m,n)$ to be ∞ when $n \leqslant m$ and $-\infty$ when $n > m$. We adopt the convention that $\Phi(\infty) = 1$ and $\Phi(-\infty) = 0$. For example, if $G(t) + H(t) = 0$, when $n > m$, $\Phi(M(t,m,n)) = 0$ according to our convention, which gives $\mathbb{P}_\infty(X(t) \leqslant m \mid X(0) = n) = 0$. This is consistent with the fact that $X(t)$ must equal $X(0)$ with probability 1, because $G(t) + H(t) = 0$ indicates that no arrival or discharge occurs between 0 and t with probability 1.

4.1.2. Numerical Results. Once we get the approximate stationary distribution of $X(t)$ from (28), we can get the approximate stationary distribution of $Q(t)$ using (13). The solid (blue) and dotted (red) curves in Figure 2(a) are time-dependent mean queue length $\mathbb{E}_\infty[Q(t)]$ calculated from two methods: (i) simulation estimates, which serve as the "benchmark values" (we do not use exact analysis because its computational time is extremely long as described at the beginning of Section 4); (ii) normal approximations along with π solved from the exact Markov chain analysis (8). We can see that the performance curves predicted from the normal approximations are almost identical to those curves from simulation estimates for $N = 500$. The relative differences in $\mathbb{E}_\infty[Q(t)]$ for each t are less than 0.25%. The dash-dotted (green) curve in the figure will be explained in Section 4.3.

4.1.3. Error Bound. The following theorem provides a Berry-Esseen type bound, justifying the approximation in (28) when N is large and ρ is close to 1.

Theorem 1. *Fix $t \in [0,1)$. Then*

$$\sup_{m \in \mathbb{Z}_+} \left| \mathbb{P}_\infty(X(t) \leqslant m) - \sum_{n=0}^{\infty} \Phi(M(t,m,n))\pi(n) \right|$$
$$\leqslant 0.4785 \left(\mathbb{1}_{\{G(t)>0\}} \frac{1}{\sqrt{\Lambda G(t)}} + \mathbb{1}_{\{H(t)>0\}} \right.$$
$$\left. \cdot \frac{(\mu H(t))^2 + (1-\mu H(t))^2}{\sqrt{\mu H(t)(1-\mu H(t))}} \left((1-\rho) + \frac{\sqrt{2\mu}}{\sqrt{\Lambda}} \right) \right), \quad (29)$$

where, for a set A, $\mathbb{1}_A$ denotes the indicator function of A.

Both terms in the right-hand side of (29) converge to 0 at rate $1/\sqrt{N}$ under the following three conditions: (i) μ is fixed as $N \to \infty$; (ii) N, μ, and Λ satisfies a square-root safety staffing rule, namely, there exists a $\beta > 0$ such that

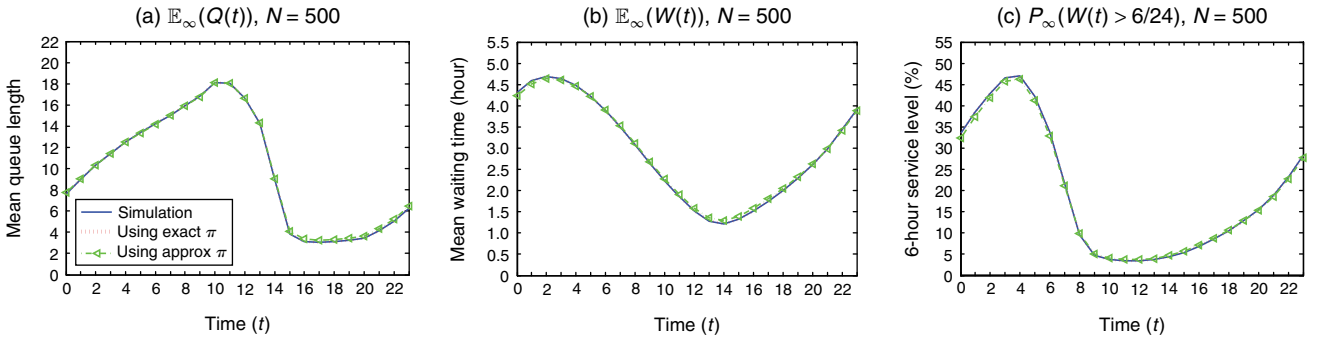
$$\frac{\Lambda}{\mu} = N - \beta\sqrt{N}; \quad (30)$$

and (iii) β is fixed as $N \to \infty$. Since

$$\sqrt{\mu}/\sqrt{\Lambda} = (1/\sqrt{1 - \beta/\sqrt{N}})/\sqrt{N}$$

and $1 - \rho = \beta\sqrt{N}/N = \beta/\sqrt{N}$, it is easy to check that the convergence rates of both terms in (29) are $1/\sqrt{N}$. The staffing rule in (30) is known to lead systems into the quality-and-efficiency-driven (QED) regime in many-server queues modeling customer call center operations (Gans et al. 2003); see discussions on the QED regime in Section 1.4.

Figure 2. (Color online) Time-dependent performance curves from simulation and approximations



Notes. Here, $\Lambda = 90.95$ for $N = 500$, the mean LOS equals 5.30 days, and we use the baseline discharge distribution. The three performance curves in each subfigure are from using (i) simulation, (ii) normal approximations and exact midnight count π, and (iii) normal approximations and approximate π from (40), respectively.

We now outline the proof for Theorem 1. Fix an $m \geqslant 0$. Clearly,

$$\begin{aligned}
&\mathbb{P}(X(t) \leqslant m)\\
&= \sum_{n=0}^{\infty} \mathbb{P}(X(t) \leqslant m \mid X(0) = n)\pi(n)\\
&= \sum_{n=0}^{\infty} \mathbb{P}(X(t) \leqslant m + 0.5 \mid X(0) = n)\pi(n)\\
&= \sum_{n=0}^{\infty} \mathbb{P}(n + A_{(0,t]} - D_{(0,t]} \leqslant m + 0.5 \mid X(0) = n)\pi(n).
\end{aligned}$$

where, the last equality follows from (9), and the 0.5 term in the second equality is added as a continuity correction factor.

When $G(t) + H(t) = 0$, both $A_{(0,t]} = 0$ and $D_{(0,t]} = 0$ with probability one. In this case, one can easily check that the expression in the left-hand side of (29) is equal to zero. Therefore, (29) is satisfied as an equality, proving Theorem 1. Now assume $G(t) + H(t) > 0$ and $n \geqslant 1$ (the $n = 0$ case can be treated in the same way as $H(t) = 0$). Then, Theorem 1 follows from

$$\begin{aligned}
&\mathbb{P}(n + A_{(0,t]} - D_{(0,t]} \leqslant m + 0.5 \mid X(0) = n) \qquad (31)\\
&\quad = \mathbb{P}\left(\frac{(A_{(0,t]} - \Lambda G(t)) - (D_{(0,t]} - z(n)\mu H(t))}{\sqrt{\Lambda G(t) + z(n)\mu H(t)(1 - \mu H(t))}} \right.\\
&\qquad \left. \leqslant M(t, m, n) \mid X(0) = n \right) \qquad (32)
\end{aligned}$$

and the following lemma.

Lemma 2. (a) *Assume that* $\mu \in (0,1)$ *and* $G(t) + H(t) > 0$. *Then, for* $n \geqslant 1$,

$$\begin{aligned}
&\sup_{x \in \mathbb{R}} \left| \mathbb{P}\left(\frac{(A_{(0,t]} - \Lambda G(t)) - (D_{(0,t]} - z(n)\mu H(t))}{\sqrt{\Lambda G(t) + z(n)\mu H(t)(1 - \mu H(t))}} \right.\right.\\
&\quad \left.\left. \leqslant x \mid X(0) = n \right) - \Phi(x) \right| \leqslant 0.4785 \left(\mathbb{1}_{\{G(t)>0\}} \frac{1}{\sqrt{\Lambda G(t)}} \right.\\
&\quad \left. + \mathbb{1}_{\{H(t)>0\}} \frac{1}{\sqrt{z(n)}} \frac{(\mu H(t))^2 + (1 - \mu H(t))^2}{\sqrt{\mu H(t)(1 - \mu H(t))}} \right). \qquad (33)
\end{aligned}$$

(b) *Assume condition* (4). *Then*

$$\sum_{n=1}^{\infty} \frac{1}{\sqrt{z(n)}} \pi(n) \leqslant (1 - \rho) + \frac{\sqrt{2\mu}}{\sqrt{\Lambda}}.$$

Section 2 of the electronic companion details the proof of Lemma 2, and the constant 0.4785 comes from Theorem 1.1 of Ross (2011).

4.2. Normal Approximation for the Time-Dependent Waiting Time Performance

We first propose an approximation for $\mathbb{P}_\infty(W(t) > x)$ in Section 4.2.1, then show numerical results on the accuracy of the approximation in Section 4.2.2, and establish an error bound for the approximation in Section 4.2.3. We also make some comments on the accuracy of normal approximations for small systems in Section 4.2.4.

4.2.1. Approximation Formulas. To approximate the time-dependent mean waiting time and six-hour service level, the key step is to approximate the probability $\mathbb{P}_\infty(W(t) > x)$, which we state below.

Approximation 2. *Fix* $0 \leqslant t < 1$. *When* $0 \leqslant t + x < 1$,

$$\begin{aligned}
\mathbb{P}_\infty(W(t) > x) &\approx \sum_{n=0}^{\infty} \pi(n)\\
&\cdot \Phi\left(\frac{0.5 + n + \Lambda G(t) - N - z(n)\mu H(t + x)}{\sqrt{\Lambda G(t) + z(n)h(t,x)}} \right). \qquad (34)
\end{aligned}$$

When $t + x \geqslant 1$,

$$\begin{aligned}
\mathbb{P}_\infty(W(t) > x) \approx \sum_{n=0}^{\infty} \Phi\Big[&(0.5 + n + \Lambda G(t) - N - (z(n)\mu\\
&+ (k_{t,x} - 1)N\mu + N\mu H(t + x - k_{t,x})))\\
\cdot (\Lambda G(t) + (z(n) + (k_{t,x} - 1)N)\mu(1 - \mu) + Nh(t,x))^{-1/2} \Big] &\pi(n). \qquad (35)
\end{aligned}$$

Here,

$$\begin{aligned}
z(n) &= \min(n, N), \quad k_{t,x} = \lfloor t + x \rfloor,\\
h(t,x) &= \mu H(t + x - k_{t,x})(1 - \mu H(t + x - k_{t,x})).
\end{aligned}$$

Unfortunately, we cannot combine (34) and (35) into a single form, even though they are continuous at $t + x = 1$. Because when $k_{t,x} = 0$ and $z(n) = n$, the last term in the numerator in (35), $z(n)\mu + (k_{t,x} - 1)N\mu + N\mu H(t + x - k_{t,x})$, does not equal to $z(n)\mu H(t + x)$, which is the last term in the numerator in (34).

The intuition for Approximation 2 is still to replace $A_{(0,t]}$ and $D_{(0,t+x]}$ with appropriate normal random variables. The reason why we need to consider different scenarios of $t + x$ in this approximation follows the discussion in Section 2.3; that is, $D_{(0,t+x]}$ takes different distributions (with different means and standard deviations) depending on the value of $t + x$.

4.2.2. Numerical Results. The solid (blue) and dotted (red) curves in Figures 2(b) and 2(c) show time-dependent mean waiting time and six-hour service level calculated from two methods: (i) simulation estimates ("benchmark values"), and (ii) normal approximations along with π solved from the exact Markov chain analysis (8). Similar to what we observed on the mean queue length, the waiting time performance curves predicted from the normal approximations are almost identical to those benchmark ones for $N = 500$, with the relative differences less than 0.5%.

4.2.3. Error Bound. Theorem 2 provides an error bound for the approximation in (34). We focus on the scenario when $0 \leqslant t+x<1$; the scenarios when $t+x \geqslant 1$ can be adapted accordingly.

Theorem 2. *Fix $t \in [0,1)$ and x such that $0 \leqslant t+x<1$. Then*

$$\begin{aligned}
&\left|\mathbb{P}_\infty(W(t)>x)\right. \\
&\quad \left.-\sum_{n=0}^{\infty}\Phi\left(\frac{0.5+n+\Lambda G(t)-N-z(n)\mu H(t+x)}{\sqrt{\Lambda G(t)+z(n)h(t,x)}}\right)\pi(n)\right| \\
&\leqslant 0.4785\left(\mathbb{1}_{\{G(t)>0\}}\frac{1}{\sqrt{\Lambda G(t)}}+\mathbb{1}_{\{H(t+x)>0\}}\right. \\
&\quad \left.\cdot\frac{(\mu H(t+x))^2+(1-\mu H(t+x))^2}{\sqrt{h(t,x)}}\left((1-\rho)+\frac{\sqrt{2\mu}}{\sqrt{\Lambda}}\right)\right).
\end{aligned}$$

The proof for Theorem 2 is similar to that of Theorem 1. Recall that conditioning on $X(0)=n$, $D_{(0,t+x]}$ follows a binomial distribution with parameters $(z(n), \mu H(t+x))$ when $0 \leqslant t+x<1$ (see Lemma 1). Using (18) and a similar transformation as in (32), we have

$$\begin{aligned}
&\mathbb{P}_\infty(W(t)>x \mid X(0)=n) \\
&=\mathbb{P}_\infty(0.5+n+A_{(0,t]}-N \geqslant D_{(0,t+x]} \mid X(0)=n) \\
&=\mathbb{P}_\infty\left(\frac{0.5+n+\Lambda G(t)-N-z(n)\mu H(t+x)}{\sqrt{\Lambda G(t)+z(n)h(t,x)}}\right. \\
&\geqslant \left.\frac{(D_{(0,t+x]}-z(n)\mu H(t+x))-(A_{(0,t]}-\Lambda G(t))}{\sqrt{\Lambda G(t)+z(n)h(t,x)}}\,\middle|\, X(0)=n\right).
\end{aligned} \tag{36}$$

Then, applying Lemma 2 with $D_{(0,t+x]}$ replacing $D_{(0,t]}$ and $H(t+x)$ replacing $H(t)$, we can prove Theorem 2. The 0.5 term in (36) is again added as a continuity correction factor.

4.2.4. Small Systems. Supported by Theorems 1 and 2, the normal approximations we develop in Sections 4.1 and 4.2 are more accurate when N is large and ρ is close to 1. However, our numerical results show that the normal approximations still work remarkably well when N is small and ρ is only about 90%; see Figures B.1 and B.2 in Appendix B for plots of systems with $N=66$ and $N=18$. Indeed, among all the numerical experiments we have tested, the relative differences in the time-dependent performance between normal approximations and benchmark values are less than 3%, typically less than 1%; see Section 5.1 for details on the numerical experimental settings, including the wide range of N and utilization ρ we have tested.

4.3. Stein's Method to Approximate the Midnight Count Distribution

In this section we apply the *Stein method* framework to identify a continuous density to approximate the midnight distribution π and establish an error bound of the approximation. We first specify the approximation formulas in Section 4.3.1, and then show numerical results in Section 4.3.2. We detail the procedure of applying Stein's method and prove Theorem 3 to justify the approximation in Section 4.3.3. The Stein method framework is demonstrated in Braverman and Dai (2017) for steady-state diffusion approximation in the continuous-time setting; see also Gurvich (2014) that has inspired this line of research.

4.3.1. Approximation Formulas. Let X_∞ be the steady-state number of customers in the system at the *midnight*. In other words, the distribution of X_∞ is π. For any arbitrary positive number $\delta>0$, let

$$\tilde{X}_\infty=\delta(X_\infty-N) \quad \text{and} \quad x=\delta(n-N) \quad \text{for } n \in \mathbb{Z}_+.$$

For $x \in \mathbb{R}$, define

$$b(x) = \delta(\Lambda-N\mu)+\mu x^-, \tag{37}$$
$$\sigma^2(x) = b^2(x)-\delta(1-\mu)b(x)+\delta^2(2-\mu)\Lambda, \tag{38}$$

where $x^-=-\min(x,0)$ for $x \in \mathbb{R}$. We define the following function $p\colon \mathbb{R} \to \mathbb{R}_+$

$$p(x)=C_1\frac{1}{\sigma^2(x)}\exp\left(\int_0^x\frac{2b(y)}{\sigma^2(y)}\,dy\right), \tag{39}$$

where $C_1=C_1(\Lambda,N,\mu)>0$ is the normalizing constant such that $\int_{\mathbb{R}}p(x)\,dx=1$. Note that $\sigma^2(x)>0$ for all $x \in \mathbb{R}$ since its form is derived from the second moment of a random variable; see (53) in Section 4.3.3. Thus, $p(x)$ is well defined. At the end of this section, we spell out the explicit form for $p(\cdot)$, and show $p(\cdot)$ is indeed integrable over $\mathbb{R}$. Although the following fact is not needed in our paper, we note that $p(x)$ is the stationary density of a diffusion process with infinitesimal variance $\sigma^2(x)$ and drift $b(x)$. Let Y_∞ denote a continuous random variable that has the density function $p(\cdot)$. Theorem 3 in Section 4.3.3 suggests that Y_∞ is "close" to the scaled steady-state midnight count $\tilde{X}_\infty$ in distribution, which motivates the following approximation for $\pi(\cdot)$.

Approximation 3. *For given Λ, N and μ such that (4) is satisfied, we approximate the stationary distribution of the midnight customer count process $\pi(\cdot)$ by*

$$\begin{aligned}
\pi(n) &= \mathbb{P}(X_\infty=n) \\
&= \mathbb{P}(\delta(n-N-0.5)<\tilde{X}_\infty<\delta(n-N+0.5)) \\
&\approx \int_{\delta(n-N-0.5)}^{\delta(n-N+0.5)} p(s)\,ds.
\end{aligned} \tag{40}$$

The 0.5 term in (40) is added as a continuity correction factor, similar to those in (27) and (34).

We now specify the explicit formulas for $p(\cdot)$. When

$$4(2-\mu)\Lambda-(1-\mu)^2>0, \tag{41}$$

plugging (37) and (38) to (39) gives us

$$p(x)=\begin{cases} p_+(x)=C_+\exp(2b(0)x/\sigma^2(0)), & x\geqslant 0,\\ p_-(x)=C_-(1+\zeta^2(x))^{-1-1/\mu} & \\ \qquad\cdot\exp(\nu\tan^{-1}(\zeta(x))), & \\ & x<0, \end{cases} \tag{42}$$

where

$$\nu=\frac{2\delta(1-\mu)}{\mu\eta},\quad \eta=\delta\sqrt{4(2-\mu)\Lambda-(1-\mu)^2},$$
$$\zeta(x)=\frac{2(\mu x-b(0))+\delta(1-\mu)}{\eta}.$$

The normalizing constants C_+ and C_- can be calculated using the following two conditions: (i) the density $p(x)$ is continuous at $x=0$, and (ii) $\int_{-\infty}^{\infty}p(x)\,dx=1$. Note that $p(\cdot)$ has an exponential distribution form on the positive part (because $b(0)=\delta(\Lambda-N\mu)<0$), and has a *Pearson type IV* distribution form on the negative part. When (41) is violated, the negative part of $p(x)$ has the form of a different type of Pearson distribution, which we do not specify here since for most systems, (41) is satisfied, e.g., when $\Lambda\geqslant 1/4$. In either case, $p(\cdot)$ is integrable over $\mathbb{R}$. Moreover, one can check that δ does not affect the approximation values in the right-hand side of (40) since $x=\delta(n-N)$ for $n\in\mathbb{Z}_+$. Thus, one can choose any arbitrary δ when using the approximation in (40), for example, setting $\delta=1$.

4.3.2. Numerical Results. The dash-dotted (green) curves in the three subplots of Figure 2 are time-dependent performance calculated from normal approximations, but using π approximated by (40). Figures B.1 and B.2 in Appendix B show similar plots for smaller systems with $N=66$ and 18. Comparing with the other two curves in each of these plots, we can see that using the approximate π gives reasonably good predictions on the time-dependent performance. The relative differences in the mean queue length fluctuate between 0.5% and 7% during different time periods.

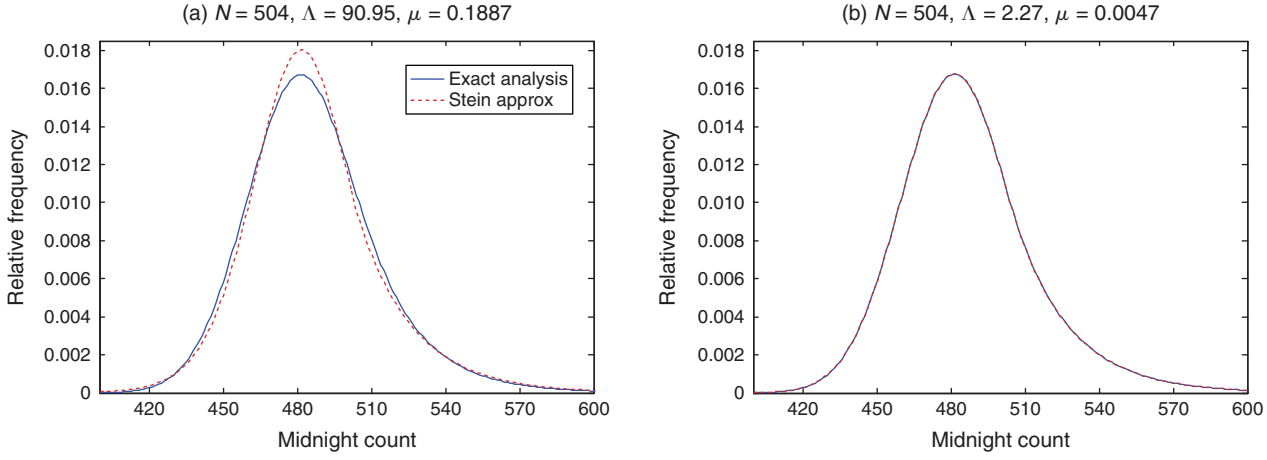
Impact of μ. Different from the normal approximations, we find that the relative difference between the exact analysis (for π) and the approximation in (40) does *not* decrease when N increases (with μ fixed); for example, see Table B.1 in Appendix B for comparisons on the midnight queue length $\mathbb{E}[(X_\infty-N)^+]$ when $\mu=1/5.3$. It turns out that if we decrease μ, or equivalently, increase the mean LOS, the relative difference between the exact analysis and approximation decreases. Figures 3(a) and 3(b) compare the exact and approximate midnight count distributions when $\mu=1/5.3=0.1887$ and $\mu=0.0047$, respectively (with $N=504$ in both figures). We can see that the approximation quality improves significantly when μ decreases. Also see Table B.2 in Appendix B for similar findings when μ decreases as N increases. These numerical results suggest that μ plays an important role in the approximation quality for the midnight count distribution, a finding that is supported by Theorem 3 below.

4.3.3. Error Bound. We now detail the procedure of applying Stein's method to identify the continuous density $p(\cdot)$ and prove Theorem 3 to support Approximation 3. Motivated by the numerical results in Section 4.3.2 that the approximation quality improves when μ decreases, we consider the following setting:

$$\mu=\delta=1/\sqrt{N}, \tag{43}$$
$$\Lambda=\sqrt{N}-\beta\quad\text{for some }\beta>0. \tag{44}$$

Under this setting, μ converges to 0 as $N\to\infty$, while the square-root staffing rule in (30) is still satisfied. Next, we state the main theorem.

Figure 3. (Color online) Stationary distribution of the midnight customer count from exact Markov chain analysis and approximation in (40) from Stein's method



Theorem 3. *Fix a $\beta > 0$. There exists a constant $C_2 = C_2(\beta)$ such that*

$$d_W(\tilde{X}_\infty, Y_\infty) \leqslant C_2(\beta)\delta^{1/2} \quad \text{for all } \Lambda, \mu, N \text{ satisfying (43) and (44).} \tag{45}$$

Here, $d_W(U,V)$ denotes the Wasserstein distance for two random variables U and V, which is defined as

$$d_W(U,V) = \sup_{h\in \mathrm{Lip}(1)} |\mathbb{E}[h(U)] - \mathbb{E}[h(V)]|, \tag{46}$$

with

$$\mathrm{Lip}(1) = \{h\colon \mathbb{R} \to \mathbb{R}, |h(x) - h(y)| \leqslant |x-y|\}, \quad \text{for all } x, y \text{ in } \mathbb{R}.$$

The continuous random variable Y_∞ has a density in the form of (39) with the drift term $b(x) = \delta(-\beta + x^-)$ and a constant variance term $\sigma_0^2 = 2\delta$.

Since $\delta = 1/\sqrt{N}$, the convergence rate in (45) is $N^{-1/4}$. Note that the drift term $b(x) = \delta(-\beta + x^-)$ for Y_∞ is the same as (37) when plugging (43) and (44), but the variance term σ_0^2 is different from $\sigma^2(x)$; the latter comes from (53) below and is state-dependent. However,

$$\begin{aligned}\sigma^2(x) &= b^2(x) - \delta(1-\mu)b(x) + (2-\mu)\delta^2\Lambda \\ &= 2\delta + \delta^2(-1-\beta(2-\delta) + (1-\delta)(\beta - x^-) + (\beta - x^-)^2) \\ &\approx 2\delta\end{aligned}$$

when δ is small so that we can ignore terms in the order of δ^2. We use $\sigma^2(x)$ instead of σ_0^2 in Approximation 3 to achieve a better approximation quality when N is only moderately large and μ is only moderately small.

Procedure of Stein's method and proof for Theorem 3. For a twice (continuous) differentiable function f, define

$$G_Y f(x) = \tfrac{1}{2}\sigma_0^2 f''(x) + b(x) f'(x), \quad x \in \mathbb{R}. \tag{47}$$

Next, consider the generator of the *scaled* midnight count process $\{\tilde{X}_k = \delta(X_k - N)\colon k = 0, 1, 2, \ldots\}$:

$$G_{\tilde{X}} f(x) = \mathbb{E}_n[f(x + \delta(A_0 - D_0)) - f(x)] \quad \text{for } x = \delta(n-N) \text{ and } n \in \mathbb{Z}_+, \tag{48}$$

Here, $\mathbb{E}_n$ is the expectation under $\mathbb{P}_n$, the conditional probability distribution given that the starting midnight count equals n with $x = \delta(n-N)$, A_0 is a Poisson random variable with mean Λ, and D_0 (under $\mathbb{P}_n$) is a binomial random variable with parameters $(\min(n,N), \mu)$ and is independent of A_0. Using the basic adjoint relation (which is verified in Section 3.3.1 of the electronic companion), we have

$$\mathbb{E}[G_{\tilde{X}} f(\tilde{X}_\infty)] = 0. \tag{49}$$

Now, we do a *generator coupling* using the Poisson equation. Fix an $h \in \mathrm{Lip}(1)$. Let $f = f_h$ be one solution to the Poisson equation

$$G_Y f(x) = h(x) - \mathbb{E}[h(Y_\infty)], \quad x \in \mathbb{R}. \tag{50}$$

From (50), we have

$$\begin{aligned}\mathbb{E}[h(\tilde{X}_\infty)] - \mathbb{E}[h(Y_\infty)] &= \mathbb{E}[G_Y f(\tilde{X}_\infty)] \\ &= \mathbb{E}[G_Y f(\tilde{X}_\infty) - G_{\tilde{X}} f(\tilde{X}_\infty)],\end{aligned} \tag{51}$$

where the second equality follows from (49). Although random variables $h(\tilde{X}_\infty)$ and $h(Y_\infty)$ in the left-hand side of (51) can be defined on two different probability spaces, random variables $G_Y f(\tilde{X}_\infty)$ and $G_{\tilde{X}} f(\tilde{X}_\infty)$ on the right-hand side are defined on the same probability space, i.e., being "coupled." This is why Equation (51) is referred as generator coupling; see Braverman and Dai (2017). Thus, the remaining task is to bound $G_{\tilde{X}} f(x) - G_Y f(x)$ for all $x = \delta(n-N)$. Doing Taylor expansion for $G_{\tilde{X}} f(x)$ for each x gives

$$\begin{aligned}G_{\tilde{X}} f(x) &= \mathbb{E}_n[f(x + \delta(A_0 - D_0)) - f(x)] \\ &= f'(x)\delta\mathbb{E}_n(A_0 - D_0) + \tfrac{1}{2}f''(x)\delta^2\mathbb{E}_n[(A_0 - D_0)^2] \\ &\quad + \tfrac{1}{6}\delta^3\mathbb{E}_n[f'''(\xi)(A_0 - D_0)^3] \\ &= G_Y f(x) + \tfrac{1}{2}\delta^2[-1 - \beta(2-\delta) + (1-\delta)(\beta - x^-) \\ &\quad + (\beta - x^-)^2]f''(x) + \tfrac{1}{6}\delta^3\mathbb{E}_n[f'''(\xi)(A_0 - D_0)^3],\end{aligned} \tag{52}$$

where

$$|\xi - x| \leqslant \delta|A_0 - D_0|,$$

and equality (52) follows from the following two facts. First,

$$\delta\mathbb{E}_n(A_0 - D_0) = \delta(\Lambda - \min(n,N)\mu) = -\mu(\beta - x^-) = b(x).$$

Second,

$$\begin{aligned}\delta^2\mathbb{E}_n[(A_0 - D_0)^2] &= \delta^2\mathrm{Var}_n(A_0 - D_0) + \delta^2(\mathbb{E}_n(A_0 - D_0))^2 \\ &= \delta^2(\Lambda + \min(n,N)\mu(1-\mu)) + b^2(x) \\ &= \delta^2\Lambda(2-\mu) - \delta(1-\mu)b(x) + b^2(x) \\ &= \sigma_0^2 + \delta^2(-1 - \beta(2-\delta)) \\ &\quad + \delta^2(1-\delta)(\beta - x^-) + \delta^2(\beta - x^-)^2.\end{aligned} \tag{53}$$

To get the third equality above, we have used the fact that

$$\min(n,N)\mu = N\mu - (N-n)^+\mu = \Lambda + \mu\beta\sqrt{N} - (N-n)^+\mu.$$

Also note that Λ in the second equality in (53) comes from the fact that A_0 is a Poisson random variable and its variance $\mathrm{Var}_n(A_0)$ equals the mean Λ. If the arrival process is not Poisson, we need to substitute Λ

with the variance of A_0; see Section 5 of the electronic companion.

Then, from (51) and (52), we have that

$$\begin{aligned}&|\mathbb{E}[h(\tilde{X}_\infty)]-\mathbb{E}[h(Y_\infty)]|\\&\quad\leqslant \tfrac{1}{2}\delta^2\|f''\|[(1+\beta(2-\delta))+(1-\delta)\mathbb{E}|\beta-\tilde{X}^-_\infty|\\&\qquad\qquad+\mathbb{E}[(\beta-\tilde{X}^-_\infty)^2]]+\tfrac{1}{6}\delta^3\|f'''\|\mathbb{E}[\mathbb{E}_n|A_0-D_0|^3],\end{aligned}$$

where $\|g\|=\sup_{x\in\mathbb{R}}|g(x)|$ is the supremum norm for a given function g defined on $\mathbb{R}$, and

$$\mathbb{E}[\mathbb{E}_n|A_0-D_0|^3]=\sum_{n=0}^{\infty}\mathbb{E}_n|A_0-D_0|^3\cdot\mathbb{P}(\tilde{X}_\infty=\delta(n-N)).$$

The rest of the proof follows from Lemmas 3 to 5 below. □

Lemma 3. *There exists a constant $C_3=C_3(\beta)>0$ such that for any $h\in\mathrm{Lip}(1)$, there exists a solution f_h to Poisson equation (50) that satisfies*

$$\|f_h''\|\leqslant C_3/\delta\quad\text{and}\quad\|f_h'''\|\leqslant C_3/\delta. \tag{54}$$

Lemma 4. *There exists a constant $C_4=C_4(\beta)>0$ such that*

$$\mathbb{E}|\beta-\tilde{X}^-_\infty|\leqslant C_4\quad\text{and}\quad\mathbb{E}[(\beta-\tilde{X}^-_\infty)^2]\leqslant C_4. \tag{55}$$

Lemma 5. *There exists a constant $C_5=C_5(\beta)>0$ such that*

$$\mathbb{E}[\mathbb{E}_n|A_0-D_0|^3]\leqslant C_5\delta^{-3/2}. \tag{56}$$

The proofs for Lemmas 3 to 5 are detailed in Sections 3.1 and 3.2 of the electronic companion.

Remark. In (43), $\mu=1/\sqrt{N}$ is a special case. Theorem 3 and its proof can be easily extended if we assume $\mu=1/N^\epsilon$, $\epsilon\in[1/2,1)$, where the convergence rate will change from $N^{-1/4}$ to $N^{-\epsilon/2}$. Theorem 3 justifies Approximation 3 when N is large and μ is close to 0. In particular, it gives us confidence to use Approximation 3 in hospital inpatient settings, because the typical patient average LOS is five to six days and the resulting μ is relatively small. Comparing to the steady-state approximation for continuous-time queueing systems studied in Braverman and Dai (2017), the midnight count process we analyze here is a discrete-time process. In their continuous-time systems, the customer count only increases or decreases by 1 (from an arrival or a discharge) at each transition, whereas in our discrete-time process, a number of arrivals and discharges could occur at each transition. The latter makes the proof of the moment bound complicated; see the proof of Lemma 5 in Section 3.2 of the electronic companion. Indeed, when trying to prove the moment bound we discovered the critical role that the parameter μ plays in the discrete-time process as explained in Section 4.3.2.

5. Numerical Results and Insights

In this section we conduct an extensive numerical study using the algorithms described before and summarize insights observed from the numerical results. We first describe the experimental settings in Section 5.1. Then in Section 5.2, we compare the time-dependent performance measures under two sets of scenarios: increasing capacity N versus shifting discharge timing (early discharge). We demonstrate their different impact on the daily and time-dependent performance, and we provide intuitive explanation using the properties shown in Section 3. In Section 5.3 we fix the arrival time t and compare a linear and a nonlinear effect in the mean waiting time reduction between the early discharge and capacity increase scenarios. The numbers reported in this section are obtained from normal approximations with π solved from exact Markov chain analysis.

5.1. Experimental Setting

Since the single-pool system we study in this paper is motivated by hospital inpatient operations, we choose most experimental parameter values based on the empirical hospital data reported in Shi et al. (2014).

5.1.1. Capacity, Utilization, and Arrival Rate. In the numerical study we test three types of systems: large systems with N around 500, medium systems with N around 120, and small systems with N around 60. The large systems correspond to pooling all inpatient beds from the entire hospital, and the small and medium systems correspond to a ward or a group of similar wards such as surgery wards that are allocated to a particular medical specialty.

We set the mean LOS to be 5.30 days and set the daily arrival rate Λ to be 90.95 for large systems, 22.74 for medium systems, and 11.37 for small systems, respectively. We vary the value of N so that the utilization ρ changes between 88% to 96%; most wards reported in Shi et al. (2014) have bed utilization in this range. For example, for large systems, the largest and the smallest N we test are 545 and 500, respectively, which correspond to 88% and 96% utilization, respectively.

We assume that the arrival rate function $\lambda(t)$ is constant in each hour. This function has a shape similar to the curve in Figure 1(a), while we proportionally adjust the actual value of $\lambda(t)$ in each hour to be consistent with the value of Λ.

5.1.2. Discharge Distributions. We consider a special type of discharge distribution, specifically, the random discharge time has the following representation

$$h_{\mathrm{dis}}=t_h+U_h,$$

where t_h follows a discrete distribution, taking values on the hour point $(1/24,2/24,\ldots,1)$ with probability

$p_{t_1}, p_{t_2}, \ldots, p_{t_{24}}$, respectively, U_h follows a uniform distribution on $(-1/24, 0)$, a one-hour interval, and t_h and U_h are independent. Note that t_h represents the discharge hour of a patient. When $t_h = i/24$, the patient discharges in the ith hour; because of the term U_h, the exact time that the patient leaves the system, h_{dis}, will be between time $(i-1)/24$ and time $i/24$, following a uniform distribution. In this paper we always use *discharge distribution* to refer to the distribution of h_{dis}, not t_h.

We first construct a *baseline* discharge distribution. We use the empirical distribution of patients' discharge hours to estimate the distribution of t_h in the baseline. Figure 4(a) shows a plot of the empirical distribution. To simplify the analysis, we further assume that $p_{t_i} = 0$ if less than 3% patients discharged in the ith hour from the empirical data, and we re-normalize the remaining probabilities. Eventually, t_h has positive probabilities on 10 points: $11/24, 12/24, \ldots, 20/24$, i.e., from hour 11 to hour 20. But because of the U_h term, patients can start to leave the hospital as early as 10 A.M. each day. We use t_{start} to denote the first time that patients can discharge each day, and set t_{start} to be 10 A.M. in the baseline discharge distribution.

To investigate the impact of changing h_{dis}, in one set of scenarios (*early discharge* scenarios) we shift the baseline discharge distribution one to four hours earlier, which is a typical range for the shifting amount as discussed in Section 3.2. We also test an extreme scenario by shifting the baseline distribution seven hours earlier. We use t_{start} to differentiate among these discharge distributions, and t_{start} varies from 10 A.M. (baseline) to 3 A.M. (shifting seven hours). Figure 4(b) plots the distributions of t_h for the baseline discharge distribution and for some of the early discharge distributions tested.

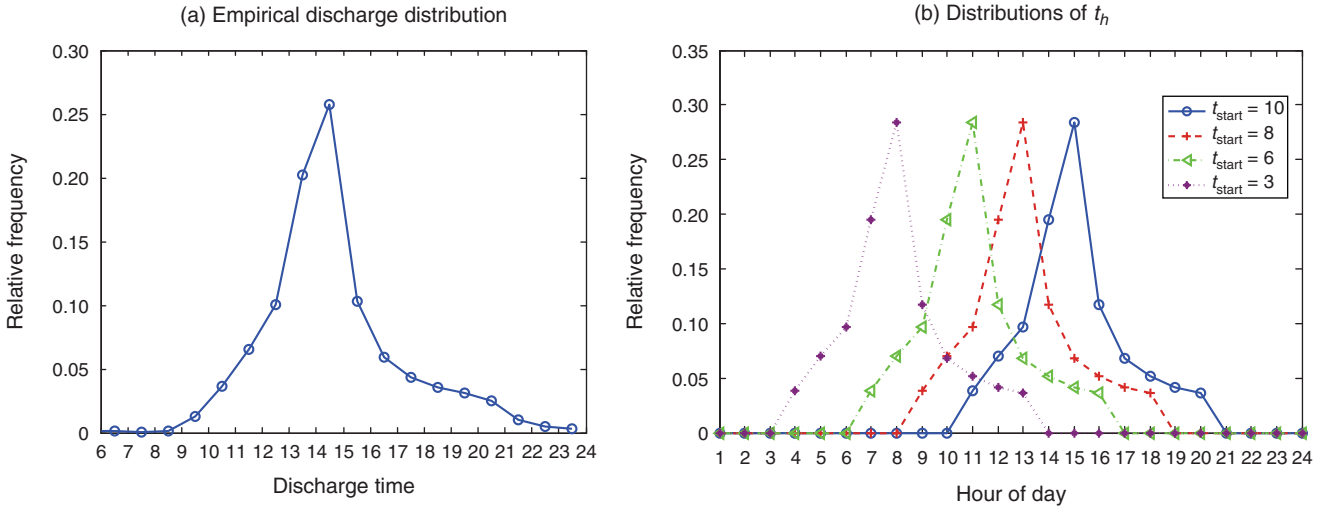
5.2. Different Impact on the Daily and Time-Dependent Performance

We focus on testing two sets of scenarios: early discharge versus capacity increase. We first summarize our observations from Figures 5 to 8 in Sections 5.2.1 to 5.2.3. Then, we explain these observations in Section 5.2.4.

Figures 5 and 6 plot the time-dependent performance curves under the early discharge and capacity increase scenarios for large systems, respectively. For Figure 5, we fix $N = 500$, but shift the baseline discharge distribution zero to three hours earlier. For Figure 6, we fix the baseline discharge distribution, but change N from 500 to 515. For each set of scenarios, we plot the time-dependent mean queue length, mean waiting time, and six-hour service level curves. Figures 7 and 8 plot similar sets of performance curves for small systems (N around 60). Table A.1 in Appendix A shows the daily performance under these tested scenarios.

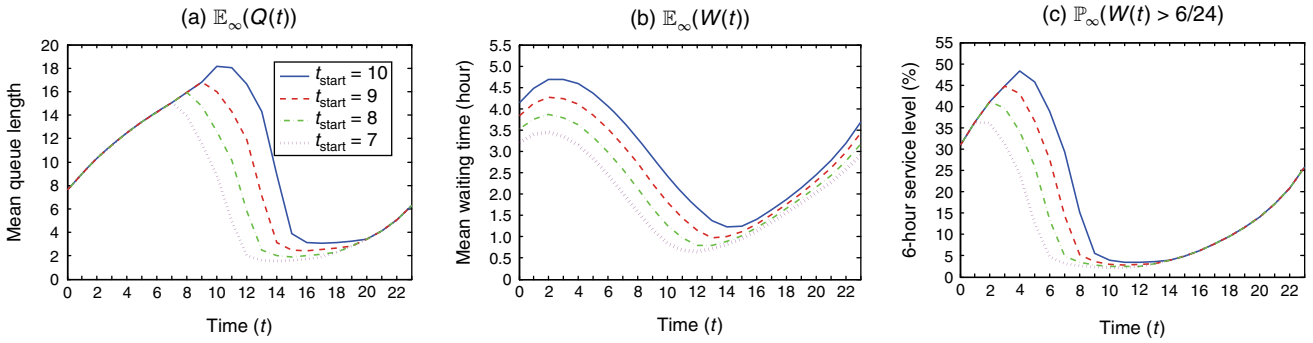
5.2.1. Early Discharge Impacts Time-Dependent Performance. Shifting h_{dis} affects the time-dependent pattern of the performance curves and mainly alleviates congestion in the morning. For example, in Figure 5, the timing of the peak performance value moves to earlier times of the day when shifting the discharge distribution, along with a reduction in the peak value. However, the daily performance shows a less significant change (see Table A.1), because the performance values in most times of the day are unaffected or only

Figure 4. (Color online) Empirical discharge hour distribution and distributions of t_h in the numerical study



Notes. In Fig. 4(a), the discharge hour distribution is obtained from empirical data between January 1, 2008 and June 30, 2009. In (4(b)), we plot the distributions of t_h under the baseline discharge distribution and a subset of early discharge distributions, with $t_{\text{start}} = 10$ representing the baseline discharge distribution.

Figure 5. (Color online) Time-dependent performance curves under the *early discharge* scenario for large systems



Notes. We fix the capacity $N = 500$ ($\rho = 0.96$), but shift the baseline discharge distribution zero to three hours earlier. We use t_{start} to denote different discharge distributions; $t_{start} = 10$ corresponds to the baseline one.

change a small amount. Still, take Figure 5 as an example. The mean queue length from midnight to 9 A.M. and from 8 P.M. to the end of the day remains the same in the four curves in Figure 5(a). The mean waiting time and six-hour service level for patients arriving after 2 P.M. remain similar in the four curves in Figures 5(b) and 5(c), respectively.

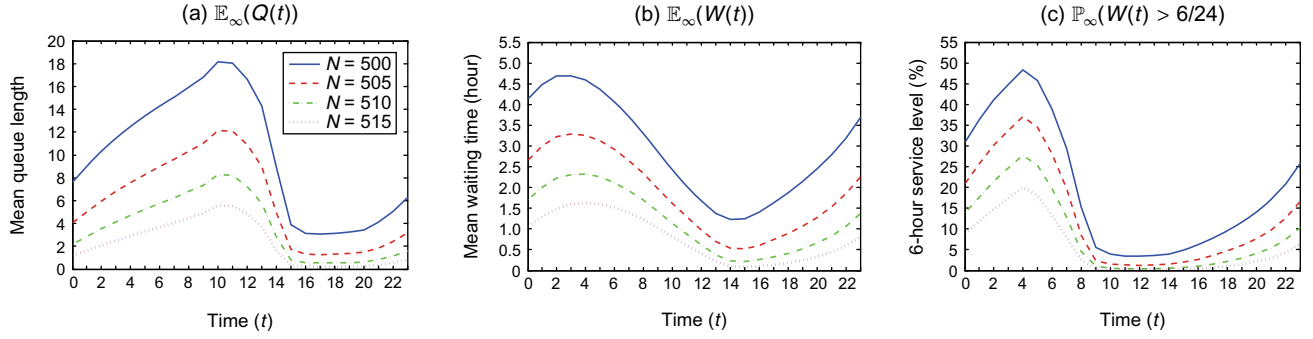
5.2.2. Capacity Increase Impacts Daily Performance. In contrast to early discharge, increasing capacity benefits customers across all different times of a day, and the entire performance curves shift down when N increases. Correspondingly, the daily performance shows more significant changes when capacity changes; see Table A.1. However, the time-dependent pattern is less affected by changing N when we keep the same discharge distribution. The performance curves maintain a similar shape when N increases, with the peak time unchanged.

5.2.3. Limited Impact of Early Discharge in Small or Heavily Utilized Systems. Comparing the curves for large systems and small systems in Figures 5 to 8, we note that the benefit of early discharge becomes much less significant in small systems, especially when compared with capacity increase. Consider patients arriving in the morning who get the largest reduction in the waiting time from early discharge in both large and small systems. We find that when $N = 500$ with $\rho = 96\%$, for patients arriving at $t = 9$ A.M., shifting the discharge distribution just three hours earlier can achieve a similar reduction in $\mathbb{E}_\infty[W(t)]$ as increasing N from 500 to 515 (a 3% capacity increase). However, when $N = 63$ with the same $\rho = 96\%$, still for $t = 9$ A.M., shifting the discharge distribution as much as seven hours now cannot even achieve a similar reduction in $\mathbb{E}_\infty[W(t)]$ as increasing N from 63 to 64 (a 1.6% capacity increase). Comparison on the mean queue length or the six-hour service level leads to similar findings.

For large systems, we test scenarios when systems are heavily utilized, e.g., $N = 485$ with $\rho = 0.99$. We observe similar phenomena, that is, the impact of early discharge becomes very limited in heavily utilized systems.

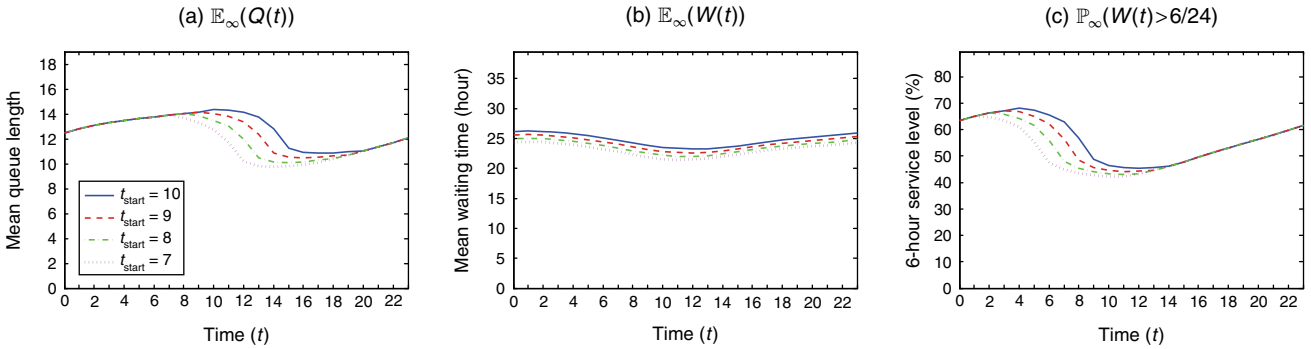
5.2.4. Explanations for the Observations. Recall from Section 2.1 that the midnight customer count distribution does not depend on the discharge time. Thus, we can see from (9) and Proposition 3 that early discharge can only reduce the customer count at certain times of a day and thus only affects the queue length at those times. This explains why in Figure 5, the mean queue

Figure 6. (Color online) Time-dependent performance curves under the *capacity increase* scenario for large systems



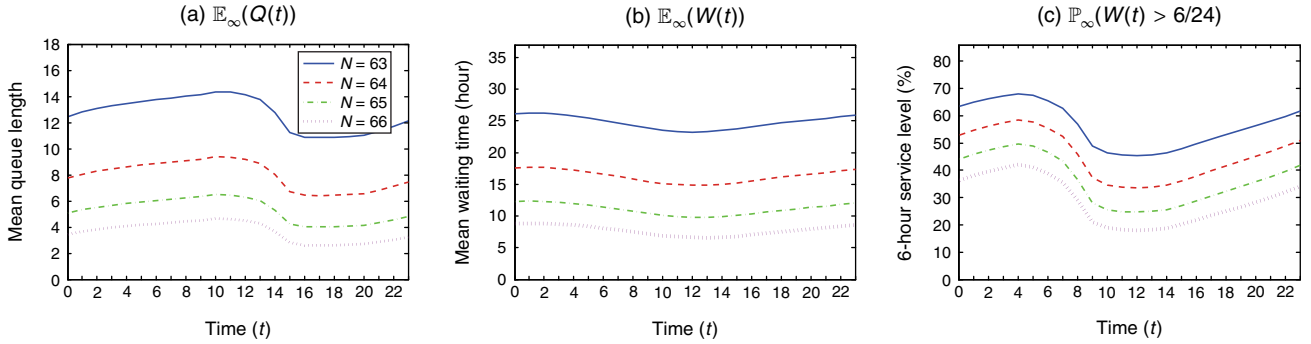
Note. We fix the baseline discharge distribution, but vary the capacity N from 500 to 515 (ρ from 0.96 to 0.94).

Figure 7. (Color online) Time-dependent performance curves under the *early discharge* scenario for small systems



Notes. We fix the capacity $N = 63$ ($\rho = 0.96$), but shift the baseline discharge distribution zero to three hours earlier. We use t_{start} to denote different discharge distributions; $t_{start} = 10$ corresponds to the baseline one.

Figure 8. (Color online) Time-dependent performance curves under the *capacity increase* scenario for small systems



Note. We fix the baseline discharge distribution, but vary the capacity N from 63 to 66 (ρ from 0.96 to 0.91).

length from midnight to 9 A.M. and from 8 P.M. to the end of the day does not change when shifting the baseline discharge distribution one to three hours earlier: the cdf of the discharge time $H(t)$ does not change for t in these time periods. In contrast, capacity increase affects the midnight count distribution, and thus it can reduce the customer count and queue length across the day.

To explain the impact on the waiting time performance, recall that in Section 3.1, we classify a customer's wait into two types: overnight wait and intraday wait. The overnight wait is in the order of days, and the intraday wait is in the order of hours. Because of the two different orders of magnitude, the overnight wait is dominating in reflecting the daily waiting time performance as long as there is a moderate number of patients waiting overnight. This explains why capacity increase can significantly impact the daily waiting time performance, whereas early discharge shows a lesser impact, since the former affects the overnight wait but the latter cannot (see Proposition 5).

This difference in the effect on the overnight wait also explains why early discharge shows a much less significant impact on heavily utilized systems compared to capacity increase. When the system load is high, many customers need to wait overnight. The waiting time is in the order of days in a heavily utilized system (the mean waiting time is longer than 20 hours across the day for $N = 485$). Early discharge can only bring a reduction in the order of hours for the waiting time, and thus it shows a small impact on heavily utilized systems. We can apply a similar argument to explain why early discharge has a lesser impact on small systems as observed in Section 5.2.3. Under the same utilization, a smaller system has less flexibility to accommodate the random fluctuations in the arrivals, which leads to a higher proportion of patients waiting overnight.

As shown in Section 3, early discharge does help to reduce the intraday wait, which mainly plays a role in the time-dependent performance. The intraday wait is caused by the nonsynchronization between the arrival and discharge time patterns, and most customers experiencing the intraday wait are those arriving in the morning before the majority of discharges occur. As a result, early discharge mainly benefit these morning arrivals. This effect is more prominent when the system is not heavily utilized and the waiting time is in the order of hours (e.g., N=500). Also, note that capacity

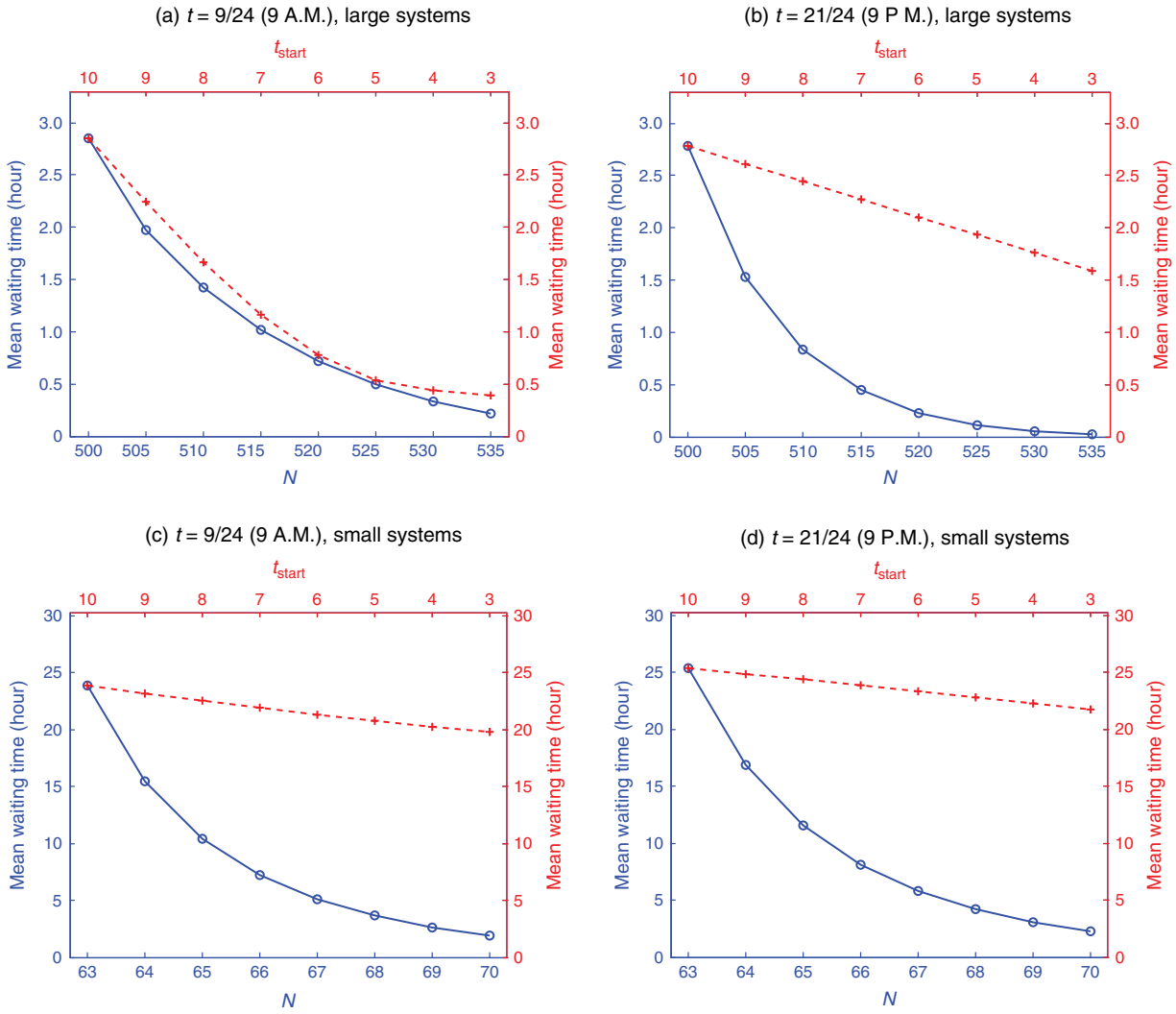
increase (in the range such that the system is still moderately utilized) cannot eliminate the nonsynchronization, which explains why we still observe the morning peak when changing N in Figure 6.

5.3. Linear Effect vs. Nonlinear Effect on $\mathbb{E}_\infty[W(t)]$

In this section we consider a fixed time t and make a more direct comparison between the impact on $\mathbb{E}_\infty[W(t)]$ from early discharge and capacity increase. We plot $\mathbb{E}_\infty[W(t)]$ for customers arriving at 9 A.M. and 9 P.M. in large systems in Figures 9(a) and 9(b), respectively. In each figure, the dashed curve represents the early discharge scenarios, and the solid curve represents the capacity increase scenarios (indicated by t_{start} and N in the upper and lower horizontal axises, respectively). Figures 9(c) and 9(d) plot a similar set of curves for small systems with N around 60.

We observe that the dashed curves in Figures 9(b), 9(c), and 9(d) are straight lines and that the first half of the dashed curve in Figure 9(a) is also straight. This linear effect in $\mathbb{E}_\infty[W(t)]$ reduction is what we have shown in (26), that is, the reduction in the mean waiting time is proportional to h, the amount of time that the discharge distribution is shifted. We can check that conditions (21) and (25) are both satisfied when we use the baseline discharge distribution and consider $t = 9$ P.M. For $t = 9$ A.M., although condition (25) is violated when t_{start} is earlier than 9 A.M., Figure 9(a) shows that there is an approximate linear effect when $1 \leqslant h \leqslant 3$ (before t_{start} shifts to 6 A.M. or earlier). This is because the majority of customers arriving at 9 A.M., if delayed, would not be admitted within three hours under the baseline discharge distribution, and (25) approximately holds. Similar explanations apply to Figure 9(c). Moreover,

Figure 9. (Color online) Comparison between early discharge and capacity increase for the mean waiting time $\mathbb{E}_\infty[W(t)]$ at 9 A.M. and 9 P.M.



Note. Solid curves represent capacity increase scenarios (use N in the lower horizontal axis), and dashed curves represent early discharge scenarios (use t_{start} in the upper horizontal axis).

the linear part of the dashed line in Figure 9(a) has a steeper slope than that in Figure 9(b) (both plots use the same scale), because the slope equals $\mathbb{P}(W(t) > 0)$, which is larger in the morning than in the night. This observation also confirms what we saw in Figure 5(b), i.e., early discharge mainly benefits morning arrivals.

Different from the dashed curves, the solid curves in Figure 9 are clearly nonlinear. The nonlinear effect associated with capacity increase is consistent with the nonlinear effect of utilization displayed in the Pollaczek-Khinchine formula for single-server queueing systems (Gross and Harris 1985).

6. Conclusion and Future Research

This paper presents a two-time-scale framework to analyze a time-varying $M_{\text{peri}}/\text{Geo}_{\text{2timeScale}}/N$ system that is motivated by modeling hospital inpatient flow. The novel feature of this model is that the service times are no longer exogenous, i.i.d. random variables but explicitly depend on LOS, admission times, and discharge times. Using the two-time-scale framework, we develop exact analysis and approximations to compute various time-dependent performance measures. Through performance analysis, we advance the understanding of the effect of this unique service time model, especially how the LOS and the discharge time h_{dis}, which are on different time scales, impact the system performance. The LOS term affects the midnight count distribution and the fraction of patients who need to wait overnight. Thus, it can change the system congestion over the entire day and significantly affect the daily performance. The h_{dis} term affects the customer count at certain times of a day and the fraction of patients who experience intraday wait because of the nonsynchronization between the arrival and discharge patterns. Consequently, it can change the system congestion at certain times of a day and mainly affects the time-dependent performance.

These findings allow us to gain insights into the trade-offs among different policies, such as early discharge and capacity increase. For example, from Section 5.2.4, we can see that when the system is heavily utilized so that the fraction of patients waiting overnight is high, implementing an early discharge policy would have very limited impact, whereas increasing bed capacity should be the first priority. These insights are consistent with those summarized in Shi et al. (2016), where the authors simulated a high-fidelity stochastic network model to evaluate the impact of early discharge. The framework and tools developed in this paper help to explain those simulation results and provide an efficient way to facilitate the full-scale simulation.

For future research, a variety of additional model features that are important in the healthcare context can be added to the current single-pool system. For example, the hospital is usually a network system. One can extend the single-pool structure to a multipool structure to better mimic the reality. Patients also demonstrate different characteristics, and one can extend the model to incorporate multiple customer classes. We believe that our two-time-scale framework will encourage the development of new tools for analyzing models with these new features.

Acknowledgments

The authors thank Costis Maglaras, the associate editor, and two anonymous referees for their critical comments to the first version of this paper. These comments motivated the authors to use Stein's method to develop approximations for the midnight count distribution in Section 4.3. The authors thank Anton Braverman and Jiekun Feng, both from Cornell University, for their stimulating discussions on Stein's method. They also thank Shuangchi He from National University of Singapore for improving the exposition of the paper.

Appendix A. Daily Performance for Early Discharge and Capacity Increase Scenarios

Table A.1 shows the daily performance for large and small systems. We can see that in the capacity increase scenarios, the daily performance changes significantly when N increases, whereas in the early discharge scenarios, the daily performance shows a less significant change when we shift the discharge distribution to earlier times of the day. This

Table A.1. Daily performance for large and small systems

(a) Large systems

Early discharge (fix $N = 500$)				Capacity increase (fix $t_{\text{start}} = 10$)			
t_{start}	$\mathbb{E}[Q]$	$\mathbb{E}[W]$	$\mathbb{P}(W > 6/24)$ (%)	N	$\mathbb{E}[Q]$	$\mathbb{E}[W]$	$\mathbb{P}(W > 6/24)$ (%)
10	9.87	2.61	16.3	500	9.87	2.61	16.3
9	8.75	2.31	15.2	505	5.91	1.56	10.3
8	7.79	2.05	14.2	510	3.66	0.97	6.56
8	6.95	1.83	13.1	515	2.29	0.60	4.14

Table A.1. (Continued)

(b) Small systems

Early discharge (fix $N = 63$)				Capacity increase (fix t_{start} =10)			
t_{start}	$\mathbb{E}[Q]$	$\mathbb{E}[W]$	$\mathbb{P}(W > 6/24)$ (%)	N	$\mathbb{E}[Q]$	$\mathbb{E}[W]$	$\mathbb{P}(W > 6/24)$ (%)
10	12.9	27.2	54.8	63	12.9	27.2	54.8
9	12.6	26.6	54.2	64	7.99	16.9	43.7
8	12.3	26.0	53.6	65	5.29	11.2	34.7
8	12.0	25.4	53.0	66	3.65	7.70	27.3

Notes. Here, the mean LOS = 5.30 days, $\Lambda = 90.95$ for large systems (the upper table), while $\Lambda = 11.37$ for small systems (the lower table). We compare the early discharge scenarios and the capacity scenarios as described in Section 5.2. We show three daily performance measures: the mean queue length $\mathbb{E}[Q]$, the mean waiting time $\mathbb{E}[W]$, and the six-hour service level $\mathbb{P}(W > 6/24)$. Note that the daily mean waiting time $\mathbb{E}[W]$ is in the unit of *hours* and is directly computed from Little's formula ($\mathbb{E}[W] = \mathbb{E}[Q]/\Lambda$).

phenomenon is particularly prominent in the small systems ($N = 63$).

Appendix B. Comparing Approximations with Exact Analysis

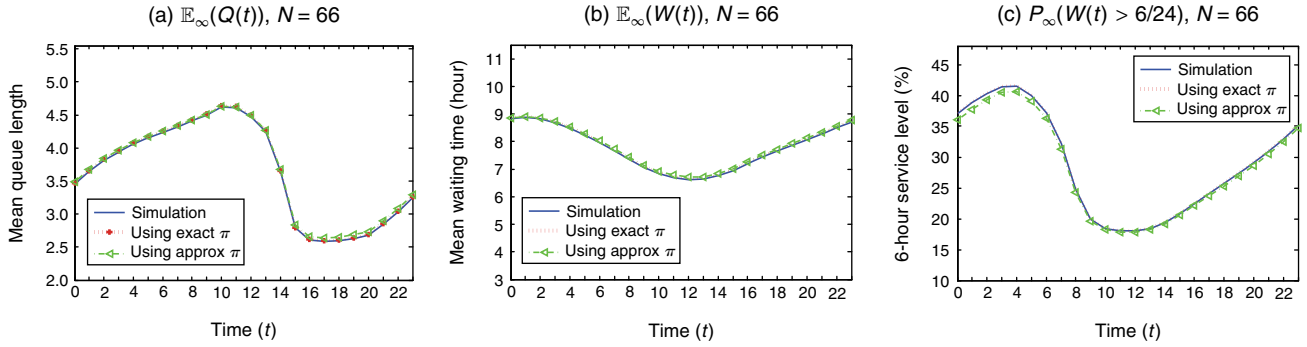
B.1. Time-Dependent Performance

Figures B.1 and B.2 show the time-dependent performance for small systems ($N = 66, 18$). The three curves in each figure are obtained from (i) simulation estimates, (ii) normal approximations using π solved from exact Markov chain analysis, and (iii) normal approximations using π approximated by (40).

B.2. Midnight Count Queue Length

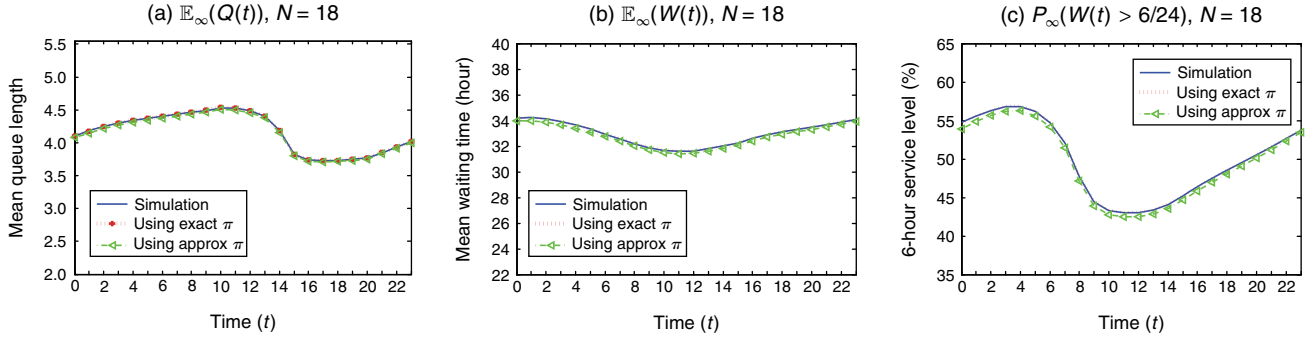
Tables B.1 and B.2 compare the midnight queue length $\mathbb{E}[(X_\infty - N)^+]$ obtained from exact Markov chain analysis and approximation in (40). Recall that X_∞ denotes the steady-state midnight customer count, whose distribution is π.

Figure B.1. (Color online) Time-dependent performance curves from simulation and approximations



Notes. Here, $\Lambda = 11.37$ for $N = 66$, the mean LOS equals 5.30 days, and we use the baseline discharge distribution. The three performance curves in each subfigure are from using (i) simulation, (ii) normal approximations and exact midnight count π, and (iii) normal approximations and π approximated by (40), respectively.

Figure B.2. (Color online) Time-dependent performance curves from simulation and approximations



Notes. Here, $\Lambda = 3.03$ for $N = 18$, the mean LOS equals 5.30 days, and we use the baseline discharge distribution. The three performance curves in each subfigure are from using (i) simulation, (ii) normal approximations and exact midnight count π, and (iii) normal approximations and π approximated by (40), respectively.

Table B.1. Midnight queue length $\mathbb{E}[(X_\infty - N)^+]$ from using exact and approximate midnight distribution

N	Λ	Stein	Exact	Relative diff (%)
504	90.95	4.78	4.59	4.17
995	181.92	6.83	6.55	4.20
1,484	272.90	8.40	8.06	4.21
1,972	363.89	9.72	9.33	4.22
2,945	545.65	11.94	11.46	4.23
3,917	727.51	13.82	13.26	4.23
7,799	1,455.22	19.61	18.81	4.24

Note. The parameter μ is fixed at 1/5.30, and $\beta = 0.977$.

Table B.2. Midnight queue length $\mathbb{E}[(X_\infty - N)^+]$ from using exact and approximate midnight distribution

N	μ	Λ	Mean LOS	Stein	Exact	Relative diff (%)
504	0.189	90.95	5.30	4.78	4.59	4.17
995	0.134	129.47	7.45	6.94	6.74	3.01
1,484	0.110	159.04	9.09	8.60	8.39	2.48
1,972	0.095	183.96	10.48	10.00	9.79	2.16
2,945	0.078	225.73	12.81	12.35	12.13	1.77
3,917	0.068	260.96	14.78	14.32	14.11	1.54
7,799	0.048	369.94	20.85	20.44	20.22	1.09

Note. The parameter $\mu = 4.23/\sqrt{N}$, and $\beta = 0.977$.

In Table B.1 we fix $\mu = 1/5.30$ and use the square-root staffing rule (30) to determine Λ with $\beta = 0.977$. In Table B.2 we change μ as N changes. Specificaly, we assume

$$\mu = 4.23/\sqrt{N}.$$

In this way, $\mu = 1/5.30$ when $N = 504$, consistent with the first row in Table B.1. Moreover, Λ is still determined from the square-root staffing rule (30) with the same staffing coefficient $\beta = 0.977$.

References

Armony M, Chan CW, Zhu B (2015a) Critical care in hospitals: When to introduce a step down unit? Working paper, New York University, New York.

Armony M, Israelit S, Mandelbaum A, Marmor YN, Tseytlin Y, Yom-Tov GB (2015b) On patient flow in hospitals: A data-based queueing-science perspective. *Stochastic Systems* 5(1): 146–194.

Bernstein SL, Aronsky D, Duseja R, Epstein S, Handel D, Hwang U, McCarthy M, et al. Society for Academic Emergency Medicine, Emergency Department Crowding Task Force (2009) The effect of emergency department crowding on clinically oriented outcomes. *Acad. Emergency Medicine* 16(1):1–10.

Bramson M (2008) *Stability of Queueing Networks*. Lecture Notes in Mathematics, Vol. 1950 (Springer, Berlin).

Braverman A, Dai JG (2017) Stein's method for steady-state diffusion approximations. *Annals Appl. Probab.* Forthcoming.

Chan CW, Dong J, Green LV (2016) Queues with time-varying arrivals and inspections with applications to hospital discharge policies. *Oper. Res.*, ePub ahead of print October 27, https://doi.org/10.1287/opre.2016.1536.

Choudhury G, Lucantoni D, Whitt W (1997) Numerical solution of piecewise-stationary $M_t/G_t/1$ queues. *Oper. Res.* 45(3):451–463.

Clark GM (1981) Use of polya distributions in approximate solutions to nonstationary $M/M/s$ queues. *Commun. ACM* 24(4): 206–217.

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.

Gao P, Wittevrongel S, Bruneel H (2004) Discrete-time multiserver queues with geometric service times. *Comput. Oper. Res.* 31(1):81–99.

Green L, Kolesar PJ (1991) The pointwise stationary approximation for queues with nonstationary arrivals. *Management Sci.* 37(1): 84–97.

Green LV, Kolesar PJ (1997) The lagged PSA for estimating peak congestion in multiserver Markovian queues with periodic arrival rates. *Management Sci.* 43(1):80–87.

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.

Griffin J, Xia S, Peng S, Keskinocak P (2012) Improving patient flow in an obstetric unit. *Health Care Management Sci.* 15(1):1–14.

Gross D, Harris CM (1985) *Fundamentals of Queueing Theory* (John Wiley & Sons, New York).

Gurvich I (2014) Diffusion models and steady-state approximations for exponentially ergodic Markovian queues. *Ann. Appl. Probab.* 24(6):2527–2559.

Halfin S, Whitt W (1981) Heavy-traffic limits for queues with many exponential servers. *Oper. Res.* 29(3):567–588.

Hoot N, Aronsky D (2008) Systematic review of emergency department crowding: Causes, effects, and solutions. *Ann. Emergency Medicine* 52(2):126–136.

Huang Q, Thind A, Dreyer J, Zaric G (2010) The impact of delays to admission from the emergency department on inpatient outcomes. *BMC Emergency Medicine* 10:16.

Ingolfsson A, Akhmetshina E, Budge S, Li Y, Wu X (2007) A survey and experimental comparison of service-level-approximation methods for nonstationary $M(t)/M/s(t)$ queueing systems with exhaustive discipline. *INFORMS J. Comput.* 19(2):201–214.

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

Liu SW, Thomas SH, Gordon JA, Hamedani AG, Weissman JS (2009) A pilot study examining undesirable events among emergency department-boarded patients awaiting inpatient beds. *Ann. Emergency Medicine* 54(3):381–385.

Liu Y, Whitt W (2011a) Large-time asymptotics for the $G_t/M_t/s_t+GI_t$ many-server fluid queue with abandonment. *Queueing Systems* 67(2):145–182.

Liu Y, Whitt W (2011b) Nearly periodic behavior in the overloaded $G/D/s + GI$ queue. *Stochastic Systems* 1(2):340–410.

Liu Y, Whitt W (2011c) A network of time-varying many-server fluid queues with customer abandonment. *Oper. Res.* 59(4):835–846.

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* 71:405–444.

Mandelbaum A, Momcilovic P, Tseytlin Y (2012) On fair routing from emergency departments to hospital wards: QED queues with heterogeneous servers. *Management Sci.* 58(7):1273–1291.

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

Pines JM, Batt RJ, Hilton JA, Terwiesch C (2011) The financial consequences of lost demand and reducing boarding in hospital emergency departments. *Ann. Emergency Medicine* 58(4):331–340.

Powell ES, Khare RK, Venkatesh AK, Van Roo BD, Adams JG, Reinhardt G (2011) The relationship between inpatient discharge timing and emergency department boarding. *J. Emergency Medicine* 42(2):186–196.

Ramakrishnan M, Sier D, Taylor P (2005) A two-time-scale model for hospital patient flow. *IMA J. Management Math.* 16(3):197–215.

Ross N (2011) Fundamentals of Stein's method. *Probab. Surv.* 8: 210–293.

Rothkopf MH, Oren SS (1979) A closure approximation for the nonstationary $M/M/s$ queue. *Management Sci.* 25(6):522–534.

Shi P (2013) Stochastic modeling and decision making in two healthcare applications: Inpatient flow management and influenza pandemics. Ph.D. thesis, Georgia Institute of Technology, Atlanta.

Shi P, Chou M, Dai JG, Ding D, Sim J (2016) Models and insights for hospital inpatient operations: Time-dependent ED boarding time. *Management Sci.* 62(1):1–28.

Shi P, Dai JG, Ding D, Ang J, Chou M, Xin J, Sim J (2014) Patient flow from emergency department to inpatient wards: Empirical observations from a Singaporean hospital. https://ssrn.com/abstract_id=2517050.

Singer AJ, Thode J Henry C, Viccellio P, Pines JM (2011) The association between length of emergency department boarding and mortality. *Acad. Emergency Medicine* 18(12):1324–1329.

U.S. Centers for Disease Control and Prevention (2010) Health, United States. http://www.cdc.gov/nchs/data/hus/hus10.pdf.

U.S. General Accounting Office (2003) *Hospital Emergency Departments: Crowded Conditions Vary Among Hospitals and Communities*. United States General Accounting Office, Washington, DC.

Whitt W (1991) The pointwise stationary approximation for $M_t/M_t/s$ queues is asymptotically correct as the rates increases. *Management Sci.* 37(3):307–314.

Yom-Tov GB, 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.

Zacharias C, Armony M (2016) Joint panel sizing and appointment scheduling in outpatient care. *Management Sci.*, ePub ahead of print September 12, https://doi.org/10.1287/mnsc.2016.2532.

J. G. Dai is the Leon C. Welch Professor of Engineering in the School of Operations Research and Information Engineering at Cornell University. His research focuses on the analysis and control of stochastic processing networks with applications to communications systems, service systems, and hospital patient flow management.

Pengyi Shi is an assistant professor of operations management at Krannert School of Management, Purdue University. Her research focuses on data-driven modeling and decision making in service systems, with applications to hospital patient flow management.