Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 0000-0000 | EISSN 0000-0000 | 00 | 0000 | 0001

INFORMS
DOI 10.1287/xxxx.0000.0000

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

J. G. Dai
School of Operations Research and Information Engineering, Cornell University, Ithaca, NY 14853, jd694@cornell.edu

Pengyi Shi
Krannert School of Management, Purdue University, West Lafayette, IN 47907, shi178@purdue.edu

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 intra-day wait caused by the non-synchronization 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.

Key words: two-time-scale, time-varying queue, time-dependent performance, approximation, Stein's method, hospital inpatient operations
History: first version: August, 2014; this version: November, 2015

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, health care delivery systems, and many other service systems [21]. 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 $t \geq 0$. 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 "Geo $_{\text{2timeScale}}$"):

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

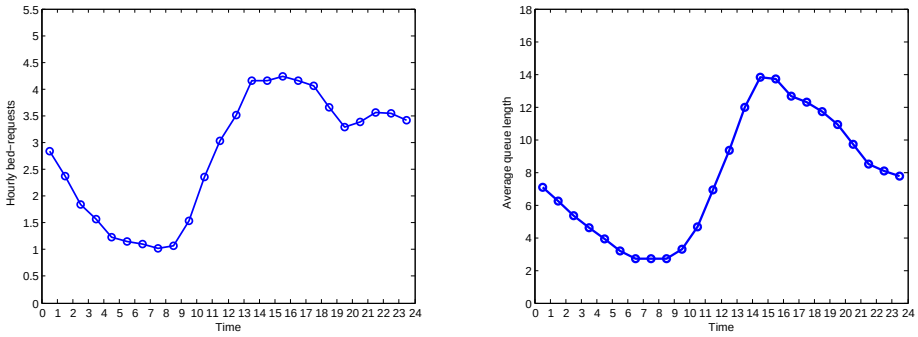
Here, LOS (Length-of-Stay) denotes the *number of days* that the customer occupies a server (which equals the departure day minus the admission day), and $h_{\text{adm}} \in (0,1)$ and $h_{\text{dis}} \in (0,1)$ represent the time-of-day when the customer is admitted and departs the system, respectively. 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 1 day according to our definition, but she spends less than 24 hours (1 day) in service. Our definition of LOS is consistent with the medical conventions on measuring inpatient stays [9]; see Section 1.1 for more details on the hospital background. Mathematically speaking,

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

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 (iid) 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 iid. This is different from the popular iid 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



(a) Hourly bed-request rate (b) Hourly average queue length

Figure 1 **Hourly bed-request rate and average queue length from empirical analysis using data from a Singaporean hospital [49, 50].**

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 [49, 50]. 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 [52]. 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 [2, 49] and can be modeled by a time-nonhomogeneous Poisson process. See Figure 1a 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 Appendix G.

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 [49] 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 [49] discusses how the conventional $M_t/GI/N$ systems with iid 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 5pm; see Figure 4a 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 [2, 10, 22] 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 [50]. 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 [48]. 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 [4, 26]. 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 [32, 51] and lead to a significant increase in the hospital operational costs [27, 43].

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 [2, 44, 49]. Figure 1b shows the time-dependent mean queue length from the empirical analysis in [50], 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 [21]. This assumption is usually not valid in the hospital inpatient setting, since a typical patient service time is 4 to 5 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 [2, 3, 39, 45].

Second, besides the exact analysis, we develop efficient approximation tools to analyze the queueing system. Specifically, for step (1) in the two-time-scale framework, we use Stein's method to identify a continuous density with an explicit formula to approximate the stationary distribution of the midnight count process. As we will see from Section 4.3.3, this method provides not only a powerful mathematical tool for us to establish the error bound for the approximation, but also a practical engineering tool that can guide one to find the appropriate continuous density for steady-state approximations. 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) *intra-day wait* occurs when the arrival and discharge patterns are non-synchronized 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 it can reduce the intra-day wait by eliminating the non-synchronization; 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, cautions 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 [49], to accurately replicate empirical performance curves such as the one in Figure 1b, 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 [49], 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 systematical 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 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 iid 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 [20]. 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 [12, 47], pointwise stationary approximation (PSA) [20, 53], lagged PSA [19],

modified offered-load approximation [40, 54], infinite-server approximation [29], and iteration algorithms [11, 15]. See [21, 28] for comprehensive surveys and comparisons on these approximation methods. Recently, Liu and Whitt [35, 36, 37] 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 well apply to analyze our $M_{\text{peri}}/\text{Geo}_{\text{2timeScale}}/N$ system.

The two most related works to our analytical framework are Ramakrishnan et al. [45] and Powell et al. [44]. The discrete-time Markov chain (DTMC) considered in [45] is similar to the one we develop for the midnight count process; see Section 2.1. However, [45] 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. [44] 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 [44] assumed that all servers are occupied at 8am 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 [25] and has been widely considered in call center research [16]. Recent studies [2, 39] 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 [1, 39, 54] 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 [24] and later systematically developed in [7]; 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 Appendix C.1, 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 Appendix C.2 using this revised system.

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 [6], i.e., when $x > N$, we have

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

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

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

Here, $f_1(k)=\frac{\Lambda^k}{k!}e^{-\Lambda}$ is the probability mass function (pmf) at point k for a Poisson distribution with mean Λ, $g_1(i,k)=\frac{i!}{k!(i-k)!}\mu^k(1-\mu)^{i-k}$ 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 $k+j-i\geq 0$.

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

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

where $P=(P_{i,j})$ 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 [17].

2.2. A two-time-scale approach: step 2, time-of-day dynamics

For $t\geq 0$, 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 $t\geq 0$, $X(t)$ 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 $X_\cdot$ 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 $X=\{X(t),t\geq 0\}$, the queue length process $Q=\{Q(t),t\geq 0\}$, and the virtual waiting time process $W=\{W(t),t\geq 0\}$. Here, for a given $t\geq 0$, the queue length and virtual waiting time are defined as

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

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

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 Appendix C.4. □

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 [33, 34]). 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 $0 \le t < 1$. For $m \in \mathbb{Z}_+$,*

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

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

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

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

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

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

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 Appendix C.2 using the revised coin-toss system.

LEMMA 1. *For any $t \ge 0$, 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 $t \ge 0$, (13) follows immediately from (12). When $0 \le t < 1$, $k_t = 0$. 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

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

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

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 1b), mean waiting time $\mathbb{E}_\infty[W(t)]$, and 6-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 [49]. 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 6-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 6-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 $0 \le t + x < 1$.

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

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

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 Appendix C.3, where a more general version of the proposition will be stated and proved. The general version covers all cases of $t + x \ge 0$.

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

$$\mathbb{P}_\infty(W(t) > x) = \mathbb{P}_\infty(X(0) + A_{(0,t]} - N \ge D_{(0,t+x]}) \tag{18}$$

$$= \sum_{n=0}^{\infty} \mathbb{P}_\infty\big(D_{(0,t+x]} - A_{(0,t]} \le n - N | X(0) = n\big)\pi(n). \tag{19}$$

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

Note that (17) does *not* apply to the scenarios when $t + x \ge 1$. 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 $t + x \ge 1$; see Appendix C.3 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: intra-day wait and overnight wait, and show that changing h_{dis} only affects the intra-day 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 Sections 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

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

The above says that if $X(0) + A_{(0,t]} - N \geq D_{(0,1]}$, 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 $0 < w \leq 1-t$, where w is the first time such that $X(0) + A_{(0,t]} - D_{(0,t+w]} < N$. We call the customer experiences an *intra-day 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 intra-day 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 intra-day 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):

$$H(h/24) = 0. \tag{21}$$

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 0 and 4 according to our communications with hospital managers), and rarely does a patient discharge between midnight and 4am. As shown in [50], after an "early discharge campaign" in a Singaporean hospital, a new discharge peak emerged around 11am to noon, 3 hours earlier than the old 2-3pm peak.

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

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

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 $x \geq h/24$,*

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

This proposition is intuitive, and we leave its proof to Appendix C.5. 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.,*

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

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 6-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 6 hours or longer than 7 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*

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

Then,

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

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 Appendix C.6.

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 $t + x$. With these many summations, 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, which poses additional challenges on getting the time-dependent performance. In contrast, using the approximations developed in this section, we need less than 3 minutes to produce the entire mean waiting time curve for $N = 500$ (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 affects the system performance. Also see Appendix F 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 $(0 \le t < 1)$ 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

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

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 $0 \le m < \infty$,*

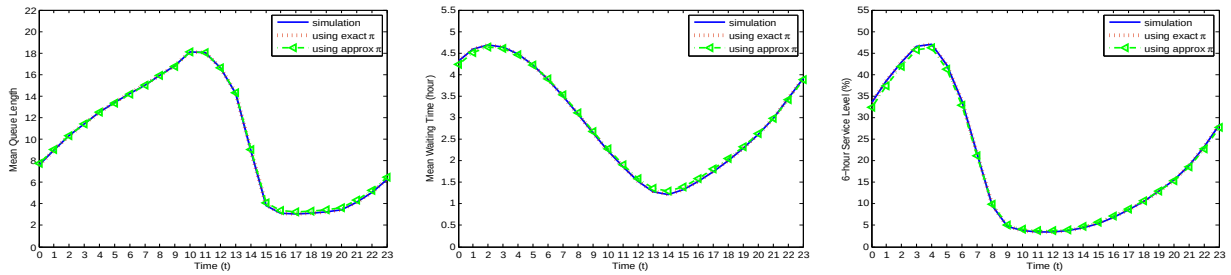
$$\mathbb{P}_\infty(X(t) \le m) \approx \sum_{n=0}^{\infty} \Phi\big(M(t,m,n)\big)\pi(n), \tag{28}$$

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 $n \le m$ and $-\infty$ when $n > m$. We adopt the convention that $\Phi(\infty) = 1$ and $\Phi(-\infty) = 0$.

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 benchmark ones 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.



(a) $\mathbb{E}_\infty(Q(t)),\ N=500$ (b) $\mathbb{E}_\infty(W(t)),\ N=500$ (c) $P_\infty(W(t)>6/24),\ N=500$

Figure 2 **Time-dependent performance curves from simulation and approximations. 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 (41), respectively.**

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*

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

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 $\sqrt{\mu}/\sqrt{\Lambda}=\Big(\frac{1}{\sqrt{1-\beta/\sqrt{N}}}\Big)/\sqrt{N}$ 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 [16]; see discussions on the QED regime in Section 1.4.

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

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

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 $n\geq 1$ (the $n=0$ case can be treated in the same way as $H(t)=0$). Then, Theorem 1 follows from

$$\mathbb{P}\Big(n+A_{(0,t]}-D_{(0,t]}\leq m+0.5|X(0)=n\Big) \tag{31}$$

$$=\mathbb{P}\left(\frac{\big(A_{(0,t]}-\Lambda G(t)\big)-\big(D_{(0,t]}-z(n)\mu H(t)\big)}{\sqrt{\Lambda G(t)+z(n)\mu H(t)(1-\mu H(t))}}\leq M(t,m,n)|X(0)=n\right) \tag{32}$$

and the following lemma.

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

$$\sup_{x\in\mathbb{R}}\left|\mathbb{P}\left(\frac{\big(A_{(0,t]}-\Lambda G(t)\big)-\big(D_{(0,t]}-z(n)\mu H(t)\big)}{\sqrt{\Lambda G(t)+z(n)\mu H(t)(1-\mu H(t))}}\leq x|X(0)=n\right)-\Phi(x)\right|$$
$$\leq 0.4785\left(\frac{1}{\sqrt{\Lambda G(t)}}\mathbb{1}_{\{G(t)>0\}}+\frac{1}{\sqrt{z(n)}}\frac{\big(\mu H(t)\big)^2+\big(1-\mu H(t)\big)^2}{\sqrt{\mu H(t)(1-\mu H(t))}}\mathbb{1}_{\{H(t)>0\}}\right). \tag{33}$$

(b) Assume condition (4). Then

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

Appendix D details the proof of Lemma 2, and the constant 0.4785 comes from Theorem 1.1 of [46].

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 6-hour service level, the key step is to approximate the probability $\mathbb{P}_\infty(W(t)>x)$, which we state below.

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

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

When $t+x\geq 1$,

$$\mathbb{P}_\infty\left(W(t)>x\right)\approx\sum_{n=0}^{\infty}\Phi\left(\frac{0.5+n+\Lambda G(t)-N-\big(z(n)\mu+(k_{t,x}-1)N\mu+N\mu H(t+x-k_{t,x})\big)}{\sqrt{\Lambda G(t)+\big(z(n)+(k_{t,x}-1)N\big)\mu(1-\mu)+Nh(t,x)}}\right)\pi(n). \tag{35}$$

Here,

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

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 (c) show time-dependent mean waiting time and 6-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 $0\le t+x<1$; the scenarios when $t+x\ge 1$ can be adapted accordingly.

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

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

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 $0\le t+x<1$ (see Lemma 1). Using (18) and a similar transformation as in (32), we have

$$\begin{aligned}\mathbb{P}_\infty(W(t)>x|X(0)=n)&=\mathbb{P}_\infty\big(0.5+n+A_{(0,t]}-N\ge D_{(0,t+x]}|X(0)=n\big)\\&=\mathbb{P}_\infty\Bigg(\frac{0.5+n+\Lambda G(t)-N-z(n)\mu H(t+x)}{\sqrt{\Lambda G(t)+z(n)h(t,x)}}\\&\qquad\ge\frac{\big(D_{(0,t+x]}-z(n)\mu H(t+x)\big)-\big(A_{(0,t]}-\Lambda G(t)\big)}{\sqrt{\Lambda G(t)+z(n)h(t,x)}}|X(0)=n\Bigg).\quad(36)\end{aligned}$$

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 10 and 11 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 *Stein's method* to identify a continuous density to approximate the midnight distribution π and establish an error bound of the approximation; see systematic descriptions of Stein's method for steady-state approximations in [7]. 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.

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

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

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 $x \in \mathbb{R}$. Note that

$$\sigma^2(x) \geq \frac{1}{4}\delta^2\Big(4(2-\mu)\Lambda - (1-\mu)^2\Big), \ \forall x \in \mathbb{R}.$$

Thus, to ensure $\sigma^2(x) > 0$ for all $x \in \mathbb{R}$, Λ and μ need to satisfy

$$4(2-\mu)\Lambda - (1-\mu)^2 > 0. \tag{39}$$

For example, this condition is met when $\Lambda \geq 1/4$ and $\mu \in (0, 1)$.

We define the following function $p : \mathbb{R} \to \mathbb{R}_+$

$$p(x) = C_1 \frac{1}{\sigma^2(x)} \exp\Big(\int_0^x \frac{2b(y)}{\sigma^2(y)} dy\Big), \tag{40}$$

where $C_1 = C_1(\Lambda, N, \mu) > 0$ is the normalizing constant such that $\int_{\mathbb{R}} p(x)dx = 1$. At the end of this section, we spell out the explicit form for $p(\cdot)$, and show $p(\cdot)$ is indeed integrable over $\mathbb{R}$. Note that $p(\cdot)$ can be seen as the probability density function of a continuous random variable. Theorem 3 in Section 4.3.3 suggests that this random variable 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) *and* (39) *are satisfied, we approximate the stationary distribution of the midnight customer count process $\pi(\cdot)$ by*

$$\begin{aligned}\pi(n) &= \mathbb{P}(X_\infty = n)\\ &= \mathbb{P}\left(\delta(n-N-0.5) < \tilde{X}_\infty < \delta(n-N+0.5)\right)\\ &\approx \int_{\delta(n-N-0.5)}^{\delta(n-N+0.5)} p\,(s)\,ds. \end{aligned} \tag{41}$$

The 0.5 term in (41) is added as a continuity correction factor, similar to those in (27) and (34).

We now specify the explicit formulas for $p(\cdot)$. Plugging (37) and (38) to (40), we can eventually get

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

where

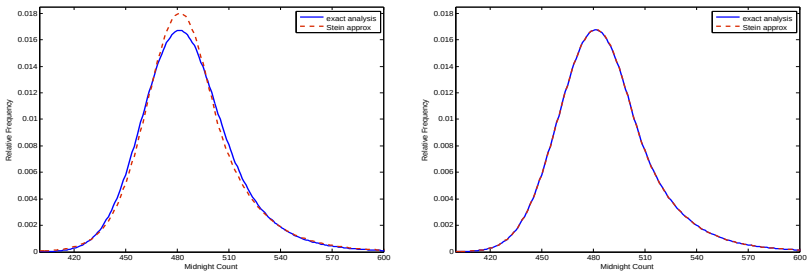
$$\nu = \frac{2\delta(1-\mu)}{\mu\eta}, \quad \eta = \delta\sqrt{4(2-\mu)\Lambda - (1-\mu)^2},$$
$$\zeta(x) = \frac{2(\mu x - b(0)) + \delta(1-\mu)}{\eta}.$$

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. Thus, $p(\cdot)$ is integrable over $\mathbb{R}$. Moreover, one can check that δ does not affect the approximation values in the right-hand side of (41) since $x = \delta(n-N)$ for $n \in \mathbb{Z}_+$. Thus, one can choose any arbitrary δ when using the approximation in (41), 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 (41). Figures 10 and 11 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 (41) does *not* decrease when N increases (with μ fixed); for example, see Table 2 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



(a) $N = 504$, $\Lambda = 90.95$, $\mu = .1887$ (b) $N = 504$, $\Lambda = 2.27$, $\mu = .0047$

Figure 3 Stationary distribution of the midnight customer count from exact Markov chain analysis and approximation in (41) **from Stein's method.**

decreases. Figure 3(a) and (b) compare the exact and approximate midnight count distributions when $\mu = 1/5.3 = .1887$ and $\mu = .0047$, respectively (with $N = 504$ in both figures). We can see that the approximation quality improves significantly when μ decreases. Also see Table 3 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:

$$\mu = \delta = 1/\sqrt{N}, \tag{43}$$

$$\Lambda = \sqrt{N} - \beta \text{ for some } \beta > 0. \tag{44}$$

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.

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

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

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

$$\mathrm{Lip}(1) = \{h : \mathbb{R} \to \mathbb{R}, |h(x) - h(y)| \le |x - y|\}.$$

The continuous random variable Y_∞ has a density in the form of (40) 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 $N^{-\frac{1}{4}}$. 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,

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

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 sketch of proof. For a twice (continuous) differentiable function f, define

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

Although the following two facts are *not* needed in our paper, we note that (i) G_Y is the generator of a diffusion process with infinitesimal variance σ_0^2 and drift $b(x)$; and (ii) this diffusion process has a stationary distribution whose density is given by the form in (40) with σ_0^2 replacing $\sigma^2(x)$.

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

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

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 $\big(\min(n, N), \mu\big)$ and is independent of A_0. Using the basic adjoint relation (which is verified in Appendix E.4.1), 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

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

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 – this is why Equation (51) is referred as generator coupling; see [7]. 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

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

where

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

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,

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

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 Appendix G. Then, to prove Theorem 3, we just need to bound the error terms in (52). See the complete proof in Appendix E.

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 $N^{-\frac{1}{4}}$ to $N^{-\frac{\epsilon}{2}}$. 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 5 to 6 days and the resulting μ is relatively small.

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 [50].

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 [50] 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.

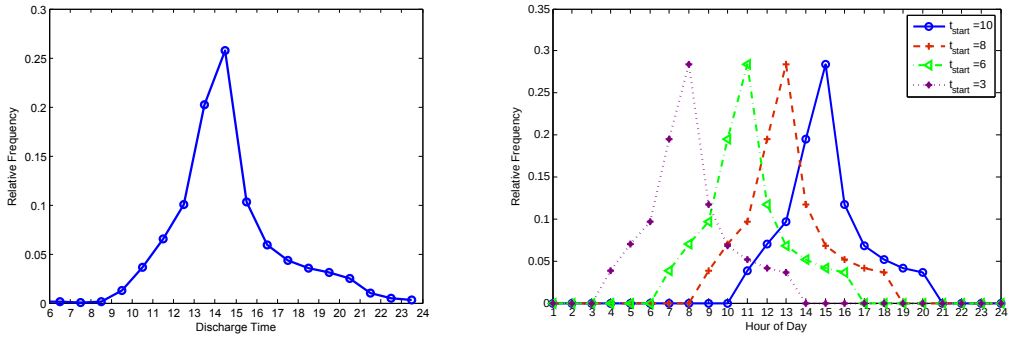
We assume that the arrival rate function $\lambda(t)$ is constant in each hour. This function has a shape similar to the curve in Figure 1a, while we proportionally adjust the actual value of $\lambda(t)$ in each hour to be consistent with the value of Λ.

5.1.2. Discharge distributions

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

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

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 $i/24$,



(a) Empirical discharge distribution (b) Distributions of t_h

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

following a uniform distribution. In this paper, we always use *discharge distribution* to refer to the distribution of h_{dis}, not t_h.

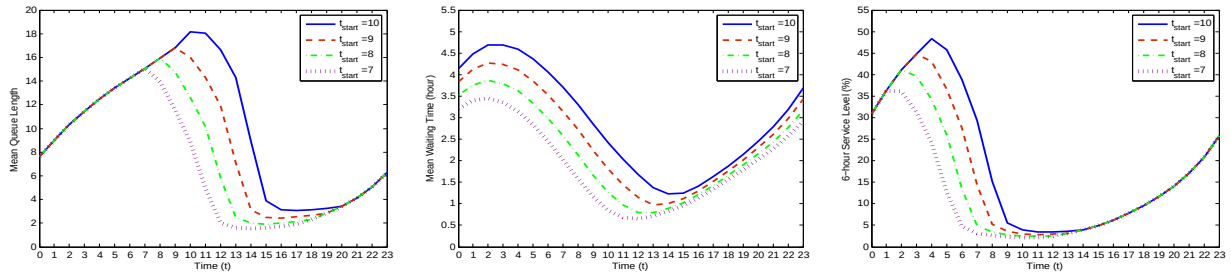
We first construct a *baseline* discharge distribution. We use the empirical distribution of patients' discharge hours to estimate the distribution of t_h in the baseline. Figure 4a shows a plot of the empirical distribution. To simplify the analysis, we further assume that $p_{t_i} = 0$ if less than 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 10am each day. We use t_{start} to denote the first time that patients can discharge each day, and set t_{start} to be 10am 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 1 to 4 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 7 hours earlier. We use t_{start} to differentiate among these discharge distributions, and t_{start} varies from 10am (baseline) to 3am (shifting 7 hours). Figure 4b 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

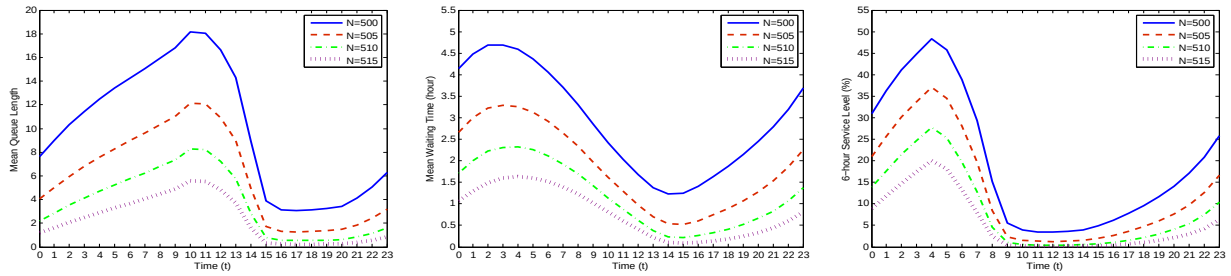
We focus on testing two sets of scenarios: early discharge versus capacity increase. We first summarize our observations from Figures 5 to 8 in Sections 5.2.1 to 5.2.3. Then, we explain these observations in Section 5.2.4.

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 0 to 3 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 6-hour service level curves. Figures 7 and 8 plot a



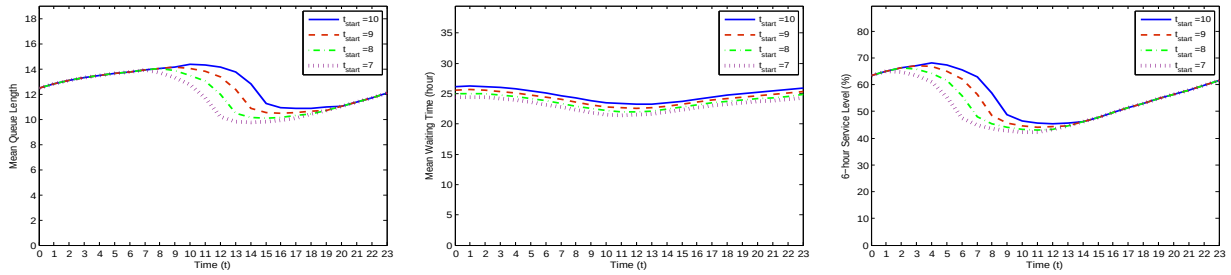
(a) $\mathbb{E}_\infty(Q(t))$ (b) $\mathbb{E}_\infty(W(t))$ (c) $\mathbb{P}_\infty(W(t) > 6/24)$

Figure 5 **Time-dependent performance curves under the early discharge scenario for large systems. We fix the capacity $N = 500$ ($\rho = 0.96$), but shift the baseline discharge distribution 0 to 3 hours earlier. We use $t_{\mathbf{start}}$ to denote different discharge distributions; $t_{\mathbf{start}} = 10$ corresponds to the baseline one.**



(a) $\mathbb{E}_\infty(Q(t))$ (b) $\mathbb{E}_\infty(W(t))$ (c) $\mathbb{P}_\infty(W(t) > 6/24)$

Figure 6 **Time-dependent performance curves under the capacity increase scenario for large systems. We fix the baseline discharge distribution, but vary the capacity N from 500 to 515 (ρ from 0.96 to 0.94).**



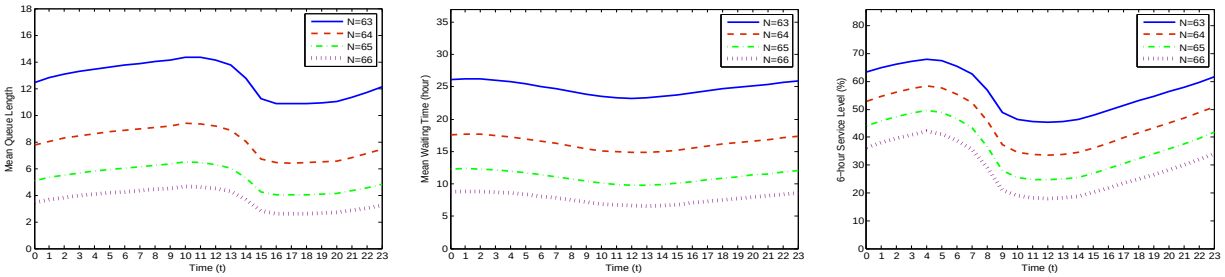
(a) $\mathbb{E}_\infty(Q(t))$ (b) $\mathbb{E}_\infty(W(t))$ (c) $\mathbb{P}_\infty(W(t) > 6/24)$

Figure 7 **Time-dependent performance curves under the early discharge scenario for small systems. We fix the capacity $N = 63$ ($\rho = 0.96$), but shift the baseline discharge distribution 0 to 3 hours earlier. We use $t_{\mathbf{start}}$ to denote different discharge distributions; $t_{\mathbf{start}} = 10$ corresponds to the baseline one.**

similar sets of performance curves for small systems (N around 60). Table 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 1),



(a) $\mathbb{E}_\infty(Q(t))$ (b) $\mathbb{E}_\infty(W(t))$ (c) $\mathbb{P}_\infty(W(t) > 6/24)$

Figure 8 **Time-dependent performance curves under the capacity increase scenario for small systems. We fix using the baseline discharge distribution, but vary the capacity N from 63 to 66 (ρ from 0.96 to 0.91).**

because the performance values in most times of the day are unaffected or only change a small amount. Still take Figure 5 as an example. The mean queue length from midnight to 9am and from 8pm to the end of the day remains the same in the four curves in Figure 5a. The mean waiting time and 6-hour service level for patients arriving after 2pm remain similar in the four curves in Figures 5b and 5c, 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 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 $t = 9$am, shifting the discharge distribution just 3 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 $t = 9$am, shifting the discharge distribution as much as 7 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 6-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 length from midnight to 9am and from 8pm to the end of the day does not change when shifting the baseline discharge distribution 1 to 3 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.

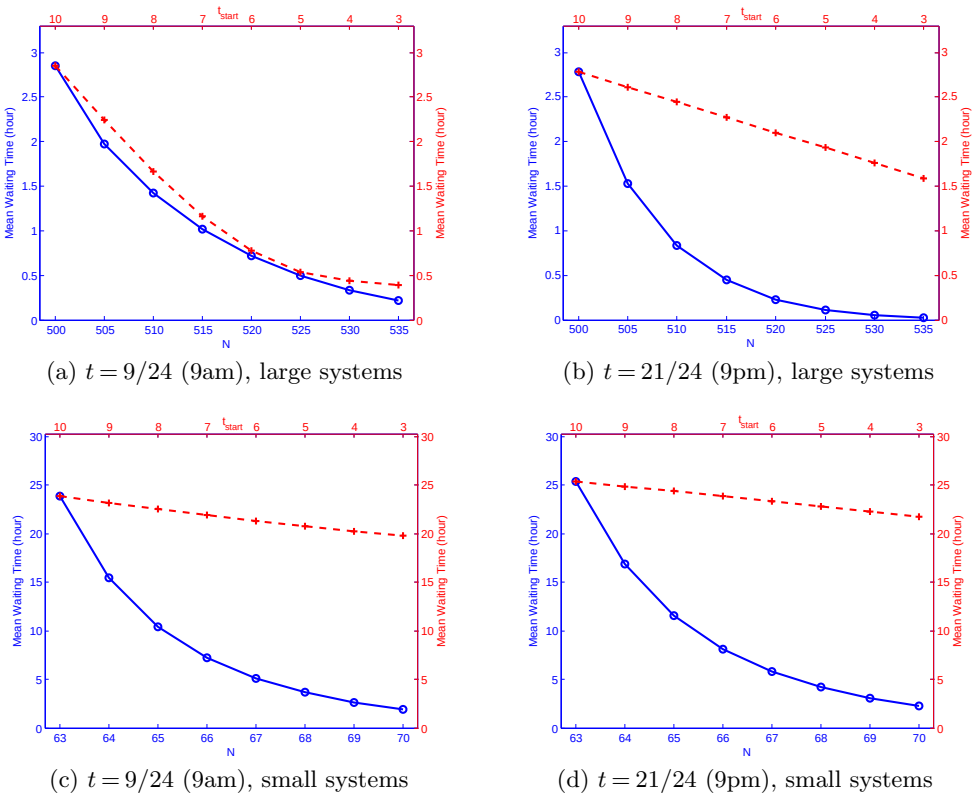
To explain the impact on the waiting time performance, recall that in Section 3.1, we classify a customer's wait into two types: overnight wait and intra-day wait. The overnight wait is in the order of days, and the intra-day wait is in the order of hours. Because of the two different orders of magnitude, 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 intra-day wait, which mainly plays a role in the time-dependent performance. The intra-day wait is caused by the non-synchronization between the arrival and discharge time patterns, and most customers experiencing the intra-day 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 non-synchronization, which explains why we still observe the morning peak when changing N in Figure 6.

5.3. Linear effect versus 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 9am and 9pm in large systems in Figures 9a and 9b, respectively. In each figure, the dashed curve represents the early discharge scenarios, and the solid curve represents the capacity



(a) $t = 9/24$ (9am), large systems

(b) $t = 21/24$ (9pm), large systems

(c) $t = 9/24$ (9am), small systems

(d) $t = 21/24$ (9pm), small systems

Figure 9 **Comparison between early discharge and capacity increase for the mean waiting time $\mathbb{E}_\infty[W(t)]$ at 9am and 9pm. Solid curves represent capacity increase scenarios (use N in the lower horizontal axis), and dashed curves represent early discharge scenarios (use $t_{\mathbf{start}}$ in the upper horizontal axis).**

increase scenarios (indicated by t_{start} and N in the upper and lower horizontal axises, respectively). Figures 9c and 9d plot a similar set of curves for small systems with N around 60.

We observe that the dashed curves in Figures 9b, 9c, and 9d are straight lines, and that the first half of the dashed curve in Figure 9a 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 $t = 9$pm. For $t = 9$am, although condition (25) is violated when t_{start} is earlier than 9am, Figure 9a shows that there is an approximate linear effect when $1 \le h \le 3$ (before t_{start} shifts to 6am or earlier). This is because the majority of customers arriving at 9am, if delayed, would not be admitted within 3 hours under the baseline discharge distribution, and (25) approximately holds. Similar explanations apply to Figure 9c. Moreover, the linear part of the dashed line in Figure 9a has a steeper slope than that in Figure 9b (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 5b, i.e., early discharge mainly benefits morning arrivals.

Different from the dashed curves, the solid curves in Figure 9 are clearly non-linear. The non-linear effect associated with capacity increase is consistent with the non-linear effect of utilization displayed in the Pollaczek-Khinchine formula for single-server queueing systems [23].

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, iid random variables but explicitly depend on LOS, admission 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 intra-day wait because of the non-synchronization 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 tradeoffs 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 [49], 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 multi-pool 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

We thank Shuangchi He from National University of Singapore for his useful comments to this paper. This research is supported in part by NSF Grants CMMI-1030589, CNS-1248117 and CMMI-1335724.

References

[1] M. Armony, C. W. Chan, and B. Zhu, "Critical care in hospitals: When to introduce a step down unit?" working paper, 2014.

[2] M. Armony, S. Israelit, A. Mandelbaum, Y. Marmor, Y. Tseytlin, and G. Yom-Tov, "Patient flow in hospitals: A data-based queueing perspective," 2011, working paper. [Online]. Available: http://www.stern.nyu.edu/om/faculty/armony/Patient%20flow%20main.pdf

[3] M. Armony and C. Zacharias, "Panel sizing and appointment scheduling in outpatient medical care," 2013, working paper.

[4] S. L. Bernstein, D. Aronsky, R. Duseja, S. Epstein, D. Handel, U. Hwang, M. McCarthy, K. John McConnell, J. M. Pines, N. Rathlev, R. Schafermeyer, F. Zwemer, M. Schull, B. R. Asplin, and Society for Academic Emergency Medicine, Emergency Department Crowding Task Force, "The effect of emergency department crowding on clinically oriented outcomes," *Academic Emergency Medicine*, vol. 16, no. 1, pp. 1–10, 2009.

[5] P. Billingsley, *Convergence of probability measures*, 2nd ed. New York: Wiley, 1999.

[6] M. Bramson, *Stability of Queueing Networks:*, ser. Lecture Notes in Mathematics / École d'Été de Probabilités de Saint-Flour. Springer, 2008.

[7] A. Braverman and J. G. Dai, "Stein's method for steady-state diffusion approximations," 2015, preprint. [Online]. Available: http://people.orie.cornell.edu/jdai/publications/preprints.html

[8] A. Braverman, J. G. Dai, and J. Feng, "Steins method for steady-state diffusion approximations: an introduction," 2015, working paper.

[9] Centers for Disease Control and Prevention, USA, "Health, United States," 2010. [Online]. Available: http://www.cdc.gov/nchs/data/hus/hus10.pdf

[10] C. W. Chan, J. Dong, and L. V. Green, "Queues with time-varying arrivals and inspections with applications to hospital discharge policies," 2015, working paper. [Online]. Available: http://www.columbia.edu/~cc3179/inspections_2015.pdf

[11] G. Choudhury, D. Lucantoni, and W. Whitt, "Numerical solution of piecewise-stationary $M_t/G_t/1$ queues," *Operations Research*, vol. 45, no. 3, pp. 451–463, MAY-JUN 1997.

[12] G. M. Clark, "Use of polya distributions in approximate solutions to nonstationary $M/M/s$ queues," *Commun. ACM*, vol. 24, no. 4, pp. 206–217, Apr. 1981.

[13] D. Cox and P. Lewis, *The statistical analysis of series of events*, ser. Methuen's monographs on applied probability and statistics. Methuen, 1966.

[14] S. N. Ethier and T. G. Kurtz, *Markov Processes: Characterization and Convergence*. New York: Wiley, 1986.

[15] Z. Feldman, A. Mandelbaum, W. A. Massey, and W. Whitt, "Staffing of time-varying queues to achieve time-stable performance," *Management Science*, vol. 54, no. 2, pp. 324–338, 2008.

[16] N. Gans, G. Koole, and A. Mandelbaum, "Telephone call centers: Tutorial, review, and research prospects," *Manufacturing & Service Operations Management*, vol. 5, no. 2, pp. 79–141, 2003.

[17] P. Gao, S. Wittevrongel, and H. Bruneel, "Discrete-time multiserver queues with geometric service times," *Computers & Operations Research*, vol. 31, no. 1, pp. 81–99, 2004.

[18] P. W. Glynn and A. Zeevi, "Bounding stationary expectations of Markov processes," in *Markov processes and related topics: a Festschrift for Thomas G. Kurtz*, ser. Inst. Math. Stat. Collect. Inst. Math. Statist., Beachwood, OH, 2008, vol. 4, pp. 195–214. [Online]. Available: http://dx.doi.org/10.1214/074921708000000381

[19] L. V. Green and P. J. Kolesar, "The lagged PSA for estimating peak congestion in multiserver Markovian queues with periodic arrival rates," *Management Science*, vol. 43, no. 1, pp. 80–87, JAN 1997.

[20] L. Green and P. J. Kolesar, "The pointwise stationary approximation for queues with non-stationary arrivals," *Management Science*, vol. 37, pp. 84–97, 1991.

[21] L. V. Green, P. J. Kolesar, and W. Whitt, "Coping with time-varying demand when setting staffing requirements for a service system," *Production and Operations Management*, vol. 16, no. 1, pp. 13–39, JAN-FEB 2007.

[22] J. Griffin, S. Xia, S. Peng, and P. Keskinocak, "Improving patient flow in an obstetric unit," *Health Care Manag Sci*, 2011.

[23] D. Gross and C. M. Harris, *Fundamentals of Queueing Theory*. New York: Wiley, 1985.

[24] I. Gurvich, "Diffusion models and steady-state approximations for exponentially ergodic Markovian queues," *The Annals of Applied Probability*, vol. 24, no. 6, pp. 2527–2559, 12 2014. [Online]. Available: http://dx.doi.org/10.1214/13-AAP984

[25] S. Halfin and W. Whitt, "Heavy-traffic limits for queues with many exponential servers," *Oper. Res.*, vol. 29, no. 3, pp. 567–588, 1981.

[26] N. Hoot and D. Aronsky, "Systematic review of emergency department crowding: Causes, effects, and solutions." *Ann Emerg Med*, vol. 52, pp. 126–36, 2008.

[27] Q. Huang, A. Thind, J. Dreyer, and G. Zaric, "The impact of delays to admission from the emergency department on inpatient outcomes," *BMC Emergency Medicine*, vol. 10, no. 1, p. 16, 2010.

[28] A. Ingolfsson, E. Akhmetshina, S. Budge, Y. Li, and X. Wu, "A survey and experimental comparison of service-level-approximation methods for nonstationary $M(t)/M/s(t)$ queueing systems with exhaustive discipline," *INFORMS Journal on Computing*, vol. 19, no. 2, pp. 201–214, 2007.

[29] O. B. Jennings, A. Mandelbaum, W. A. Massey, and W. Whitt, "Server staffing to meet time-varying demand," *Management Science*, vol. 42, pp. 1383–1394, 1996.

[30] J. F. C. Kingman, "Two similar queues in parallel," *The Annals of Mathematical Statistics*, vol. 32, no. 4, pp. 1314–1323, 1961.

[31] P. A. Lewis and G. S. Shedler, "Simulation methods for Poisson processes in nonstationary systems," in *Proceedings of the 10th conference on Winter simulation - Volume 1*. IEEE Press, 1978, pp. 155–163.

[32] S. W. Liu, S. H. Thomas, J. A. Gordon, A. G. Hamedani, and J. S. Weissman, "A pilot study examining undesirable events among emergency department-boarded patients awaiting inpatient beds," *Annals of Emergency Medicine*, vol. 54, no. 3, pp. 381–385, 2009.

[33] Y. Liu and W. Whitt, "Large-time asymptotics for the $G_t/M_t/s_t + GI_t$ many-server fluid queue with abandonment," *Queueing Systems*, vol. 67, no. 2, pp. 145–182, Feb. 2011.

[34] ——, "Nearly periodic behavior in the overloaded $G/D/s + GI$ queue," *Stochastic Systems*, vol. 1, no. 2, pp. 340–410, 2011.

[35] ——, "A network of time-varying many-server fluid queues with customer abandonment," *Operations Research*, vol. 59, no. 4, pp. 835–846, 2011.

[36] ——, "Stabilizing customer abandonment in many-server queues with time-varying arrivals," *Operations Research*, vol. 60, no. 6, pp. 1551–1564, 2012.

[37] ——, "The $G_t/GI/s_t + GI$ many-server fluid queue," *Queueing Systems*, vol. 71, pp. 405–444, 2012.

[38] S. Maman, "Uncertainty in the demand for service: The case of call centers and emergency departments," July 2009. [Online]. Available: http://ie.technion.ac.il/serveng/References/Thesis_Shimrit.pdf

[39] A. Mandelbaum, P. Momcilovic, and Y. Tseytlin, "On fair routing from emergency departments to hospital wards: QED queues with heterogeneous servers," *Management Science*, 2012.

[40] W. A. Massey and W. Whitt, "An analysis of the modified offered-load approximation for the nonstationary Erlang loss model," *Annals of Applied Probability*, vol. 4, no. 4, pp. 1145–1160, 1994.

[41] S. Meyn and R. L. Tweedie, *Markov chains and stochastic stability*, 2nd ed. Cambridge: Cambridge University Press, 2009.

[42] A. Mood, F. Graybill, and D. Boes, *Introduction to the Theory of Statistics*, 3rd ed. McGraw-Hill, 1974.

[43] J. M. Pines, R. J. Batt, J. A. Hilton, and C. Terwiesch, "The financial consequences of lost demand and reducing boarding in hospital emergency departments," *Annals of Emergency Medicine*, vol. 58, no. 4, pp. 331–340, 2011.

[44] E. S. Powell, R. K. Khare, A. K. Venkatesh, B. D. Van Roo, J. G. Adams, and G. Reinhardt, "The relationship between inpatient discharge timing and emergency department boarding," *The Journal of Emergency Medicine*, 2011.

[45] M. Ramakrishnan, D. Sier, and P. Taylor, "A two-time-scale model for hospital patient flow," *IMA Journal of Management Mathematics*, vol. 16, no. 3, pp. 197–215, 2005.

[46] N. Ross, "Fundamentals of Stein's method," *Probab. Surv.*, vol. 8, pp. 210–293, 2011. [Online]. Available: http://dx.doi.org/10.1214/11-PS182

[47] M. H. Rothkopf and S. S. Oren, "A closure approximation for the nonstationary $M/M/s$ queue," *Management Science*, vol. 25, no. 6, pp. 522–534, 1979.

[48] P. Shi, "Stochastic Modeling and Decision Making in Two Healthcare Applications: Inpatient Flow Management and Influenza Pandemics," Ph.D. dissertation, Georgia Institute of Technology, December 2013. [Online]. Available: http://web.ics.purdue.edu/~shi178/Pengyi_Shi_phd_thesis.pdf

[49] P. Shi, M. Chou, J. G. Dai, D. Ding, and J. Sim, "Models and Insights for Hospital Inpatient Operations: Time-Dependent ED Boarding Time," *Management Science*, 2014, forthcoming. [Online]. Available: http://www2.isye.gatech.edu/~dai/publications/NUH12252013-final-namedCopy.pdf

[50] P. Shi, J. G. Dai, D. Ding, J. Ang, M. Chou, J. Xin, and J. Sim, "Patient Flow from Emergency Department to Inpatient Wards: Empirical Observations from a Singaporean Hospital," 2013. [Online]. Available: http://www2.isye.gatech.edu/people/faculty/dai/publications/CompanionPaper-named.pdf

[51] A. J. Singer, J. Thode, Henry C., P. Viccellio, and J. M. Pines, "The association between length of emergency department boarding and mortality," *Academic Emergency Medicine*, vol. 18, no. 12, pp. 1324–1329, 2011.

[52] United States General Accounting Office, *Hospital emergency departments: crowded conditions vary among hospitals and communities*. Washington, D.C.: United States General Accounting Office, 2003.

[53] W. Whitt, "The pointwise stationary approximation for $M_t/M_t/s$ queues is asymptotically correct as the rates increases," *Management Science*, vol. 37, no. 3, pp. 307–314, MAR 1991.

[54] G. B. Yom-Tov and A. Mandelbaum, "Erlang-R: A time-varying queue with reentrant customers, in support of healthcare staffing," *Manufacturing & Service Operations Management*, vol. 16, no. 2, pp. 283–299, 2014.

Appendix A: Daily performance for early discharge and capacity increase scenarios

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

Appendix B: Comparing approximations with exact analysis

B.1. Time-dependent performance

Figures 10 and 11 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 (41).

B.2. Midnight count queue length

Tables 2 and 3 compare the midnight queue length $\mathbb{E}[(X_\infty - N)^+]$ obtained from exact Markov chain analysis and approximation in (41). Recall that X_∞ denotes the steady-state midnight customer count, whose distribution is π.

In Table 2, we fix $\mu = 1/5.30$, and use the square-root staffing rule (30) to determine Λ with $\beta = 0.977$. In Table 3, 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 2. Moreover, Λ is still determined from the square-root staffing rule (30) with the same staffing coefficient $\beta = 0.977$.

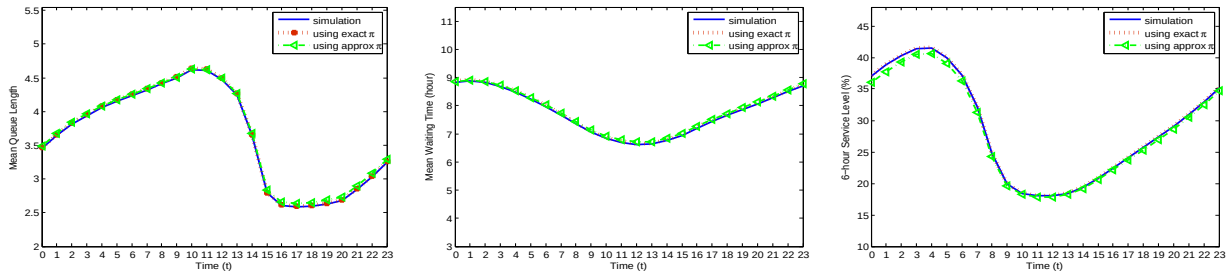
Early discharge (fix $N = 500$)				Capacity increase (fix t_{start} =10)			
t_{start}	$\mathbb{E}[Q]$	$\mathbb{E}[W]$	$\mathbb{P}(W > 6/24)$	N	$\mathbb{E}[Q]$	$\mathbb{E}[W]$	$\mathbb{P}(W > 6/24)$
10	9.87	2.61	16.3%	500	9.87	2.61	16.3%
9	8.75	2.31	15.2%	505	5.91	1.56	10.3%
8	7.79	2.05	14.2%	510	3.66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

(a) Large 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%

(b) Small systems

Table 1 Daily performance for large and small systems. 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 6-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 the Little's formula ($\mathbb{E}[W] = \mathbb{E}[Q]/\Lambda$).

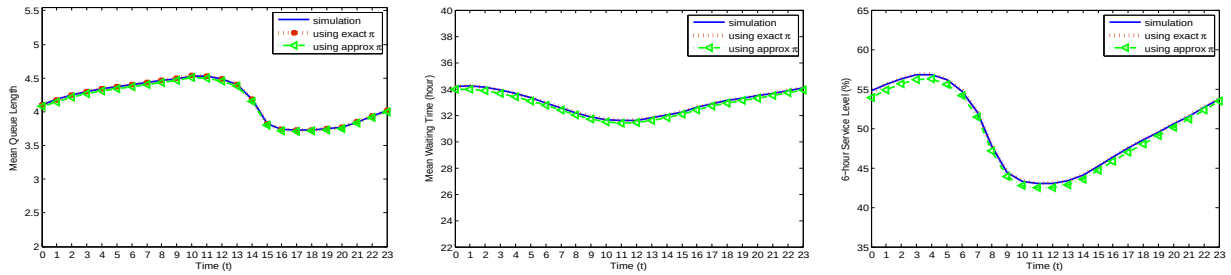


(a) $\mathbb{E}_\infty(Q(t))$, $N = 66$ (b) $\mathbb{E}_\infty(W(t))$, $N = 66$ (c) $P_\infty(W(t) > 6/24)$, $N = 66$

Figure 10 **Time-dependent performance curves from simulation and approximations. 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 (41), respectively.**

Appendix C: Proofs for Sections 2 and 3

To prove propositions and lemmas related to the exact analysis in Sections 2 and 3 of the main paper, we introduce a revised system. Recall that in the original system, we directly generate each customer's LOS according to a geometric distribution. In the revised system, we instead toss a coin for each server which is serving a customer at the beginning of each day. Only when getting a "successful" coin toss, the customer being served by that server will be discharged on that day, and her LOS is then determined by counting the number of midnights she has spent in the system. In Section C.1, we give a detailed description of this revised system, which is shown to be equivalent



(a) $\mathbb{E}_\infty(Q(t))$, $N=18$ (b) $\mathbb{E}_\infty(W(t))$, $N=18$ (c) $P_\infty(W(t)>6/24)$, $N=18$

Figure 11 **Time-dependent performance curves from simulation and approximations. 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 (41), respectively.**

N	Λ	Stein	Exact	Relative Diff
504	90.95	4.78	4.59	4.17%
995	181.92	6.83	6.55	4.20%
1484	272.90	8.40	8.06	4.21%
1972	363.89	9.72	9.33	4.22%
2945	545.65	11.94	11.46	4.23%
3917	727.51	13.82	13.26	4.23%
7799	1455.22	19.61	18.81	4.24%

Table 2 Midnight queue length $\mathbb{E}[(X_\infty-N)^+]$ from using exact and approximate midnight distribution. The parameter μ is fixed at $1/5.30$, and $\beta=0.977$.

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%
1484	0.110	159.04	9.09	8.60	8.39	2.48%
1972	0.095	183.96	10.48	10.00	9.79	2.16%
2945	0.078	225.73	12.81	12.35	12.13	1.77%
3917	0.068	260.96	14.78	14.32	14.11	1.54%
7799	0.048	369.94	20.85	20.44	20.22	1.09%

Table 3 Midnight queue length $\mathbb{E}[(X_\infty-N)^+]$ from using exact and approximate midnight distribution. The parameter $\mu=4.23/\sqrt{N}$, and $\beta=0.977$.

in distribution to the original system. In Sections C.2 to C.6, we use this revised system to prove the major propositions and lemmas in Sections 2 and 3 of the main paper.

C.1. A revised system

In the revised system, we assume that the arrival process remains the same as in the original system. However, instead of tracking each customer's service time as in the original system, we use the following scheme to generate departures.

We index the N servers by $1,2,\ldots,N$. We use a index set $\mathcal{B}(t)$ to keep track of the indices of the servers who are busy serving a customer at time t. At the midnight of day k, we generate a pair of random variables $(\xi_{j,k},h_{j,k})$ for each server $j\in\mathcal{B}(k)$. In other words, we generate such a pair of

random variables for each server who is busy at the midnight of day k. Here, $\xi_{j,k} \in \{0,1\}$ is the outcome of a coin toss with success probability μ, i.e.,

$$\xi_{j,k} = \begin{cases} 1 & \text{with probability } \mu, \\ 0 & \text{with probability } 1-\mu, \end{cases}$$

while $h_{j,k} \in (0,1)$ is independent of $\xi_{j,k}$ and follows a common distribution with cdf $H(\cdot)$. We assume that for each server j, the pair of random variables generated at each midnight forms an iid sequence $\{(\xi_{j,k}, h_{j,k}), k = 0,1,\dots\}$. We further assume that sequences associated with different servers are independent.

At the midnight of day k, if $\xi_{j,k} = 1$, the customer being served by server j will be discharged on day k, at time $h_{j,k}$; otherwise, the customer continues to stay in the system for at least one more day. If a discharge occurs from server j at time t, we check the buffer. If the buffer is not empty, we admit a waiting customer to server j following the first-come, first-served (FCFS) rule; otherwise, server j becomes idle and we delete j from the index set $\mathcal{B}(t)$. If a customer arrives at time t and is admitted to an idle server j following a predetermined server assignment rule, we add j to the index set $\mathcal{B}(t)$. Thus, we update this index set only when a new arrival occurs or when a server becomes idle.

The original and revised systems are equal in distribution as stated in the following lemma.

LEMMA 3.

$$\big(T^{(1)}_{adm}, T^{(2)}_{adm}, \dots, T^{(1)}_{dis}, T^{(2)}_{dis}, \dots\big) =^d \big(\tilde{T}^{(1)}_{adm}, \tilde{T}^{(2)}_{adm}, \dots, \tilde{T}^{(1)}_{dis}, \tilde{T}^{(2)}_{dis}, \dots\big),$$

where $=^d$ *denotes equal in distribution,* $T^{(i)}_{adm}$ *and* $T^{(i)}_{dis}$ *are the admission and discharge times of the* i*th customer in the original system, respectively, and* $\tilde{T}^{(i)}_{adm}$ *and* $\tilde{T}^{(i)}_{dis}$ *are the admission and discharge times of the* i*th customer in the revised system, respectively.*

Sketch of Proof for Lemma 3. Because the revised system is in a probability space different from the original system, we use a coupling argument to prove Lemma 3. In the interest of space, we only give a sketch of the proof. We focus on the case when $X(0) = 0$ for each system and highlight the key step of the coupling – constructing the original and revised systems from two sets of *common* variables, $\{\tilde{\xi}_{i,l}\}$ and $\{\tilde{h}_{i,l}\}$. Specifically, for each customer $i = 1,2,\dots$, we define two independent sequences of iid random variables $\{\tilde{\xi}_{i,1}, \tilde{\xi}_{i,2}, \dots\}$ and $\{\tilde{h}_{i,1}, \tilde{h}_{i,2}, \dots\}$, which follow the same distributions as $\xi_{1,1}$ and $h_{1,1}$, respectively. We further assume that sequences associated with different customers are independent. Then, for the ith customer arriving to the **original** system, we set her LOS and discharge hour to be

$$\text{LOS}^{(i)} = \inf\{\ell \geq 1 : \tilde{\xi}_{i,\ell} = 1\}, \quad h^{(i)}_{\text{dis}} = \tilde{h}_{i,\text{LOS}^{(i)}}. \tag{54}$$

It is easy to verify that $\{\text{LOS}^{(i)} : i \geq 1\}$ and $\{h^{(i)}_{\text{dis}} : i \geq 1\}$ are two independent iid sequences, following a geometric distribution with mean $1/\mu$ and a random distribution with CDF $H(\cdot)$, respectively.

This means the sample path of $\{(\text{LOS}, h_{\text{dis}})\}$ constructed via (54) satisfies all the assumptions in the original system. For the **revised** system, at the midnight of day k, for server j, we set

$$\xi_{j,k} = \tilde{\xi}_{C^j(k),l(k)}, \quad h_{j,k} = \tilde{h}_{C^j(k),l(k)}. \tag{55}$$

Here, $C^j(k)$ is the index of the customer being served by server j at time k (and equals 0 if the server is idle), and $l(k) = k - \lfloor T_{\text{adm}}^{(C^j(k))} \rfloor$ is the number of midnights that customer $C^j(k)$ has spent in the system till the midnight of day k, after her admission at time $T_{\text{adm}}^{(C^j(k))}$. Again, one can verify the sample path of $\{(\xi_{j,k}, h_{j,k})\}$ constructed via (55) meets all the requirements for the revised system. Once we construct both systems from the common variables, we then apply a standard pathwise comparison to prove that on any given sample path of the arrival times, $\{\tilde{\xi}_{i,l}\}$'s, and $\{\tilde{h}_{i,l}\}$'s, each customer i is served by the server with the same index and her admission and discharge times are the same in both systems, finishing the proof of Lemma 3. For the case when $X(0) > 0$, we just need to slightly modify the proof to consider the "residual" customers that are initially in service, whose LOS in fact represents the *remaining* number of days to-be-spent in the system (including the day of discharge). □

C.2. Proof of Lemma 1

Proof. To prove this lemma, we use the revised system introduced in Section C.1. From the dynamics of this revised system, we can express the total number of discharges between k_t and t as

$$D_{(k_t,t]} = \sum_{j \in \mathcal{B}(k_t)} \mathbb{1}_{\{\xi_{j,k_t}=1\}} \mathbb{1}_{\{h_{j,k_t} \le t-k_t\}}, \tag{56}$$

where $k_t = \lfloor t \rfloor$ for any given $t \ge 0$. Note that the multiplication term $\mathbb{1}_{\{\xi_{j,k_t}=1\}} \mathbb{1}_{\{h_{j,k_t} \le t-k_t\}}$ takes value 1 with probability $\mu H(t-k_t)$. The size of the index set $\mathcal{B}(k_t)$ equals $Z(k_t)$, the number of busy servers at the midnight k_t. Because of the independence assumption we made in the revised system, $D_{(k_t,t]}$ is in fact the summation of $Z(k_t)$ independent Bernoulli random variables with success probability $\mu H(t-k_t)$. As a result, conditioning on $X(k_t) = n$, $D_{(k_t,t]}$ follows a binomial distribution with parameter $\big(z(n), \mu H(t-k_t)\big)$ where $z(n)$ is defined in (6). □

Note that for $k = 0, 1, 2, \ldots,$

$$D_{(k,k+1]} = \sum_{j \in \mathcal{B}(k)} \mathbb{1}_{\{\xi_{j,k}=1\}}.$$

Therefore, we have the following corollary.

COROLLARY 2. *For any $t \ge 0$, conditioning on $X(k) = n$, $D_k = D_{(k,k+1]}$ follows a binomial distribution with parameters $(z(n), \mu)$.*

C.3. Proof for the distribution of waiting time

We state and prove the following proposition on the distribution of the virtual waiting time $W(t)$. This is a general version of Proposition 4 in the main paper, since (i) the proposition here covers all scenarios of $t+x$, and (ii) the distribution we evaluate here is *not* necessarily a stationary distribution.

PROPOSITION 8. *For a given $t \geq 0$ and $x \in \mathbb{R}_+$, let $k_t = \lfloor t \rfloor$. Assume that the distribution of $X(k_t)$ is $p(\cdot)$, i.e., $\mathbb{P}\big(X(k_t) = n\big) = p(n)$. If $k_t \leq t+x < k_t+1$,*

$$\mathbb{P}(W(t) > x) = \sum_{n=0}^{\infty} \left(\sum_{a=(N-n)^+}^{\infty} J_{t+x}\big(z(n), a+n-N\big) f_t(a) \right) p(n).$$

If $k_t+1 \leq t+x < k_t+2$,

$$\mathbb{P}(W(t) > x) = \sum_{n=0}^{\infty} \left(\sum_{d=0}^{z(n)} \sum_{a=(N+d-n)^+}^{\infty} J_{t+x}\left(N, a+n-d-N\right) g\big(z(n), d\big) f_t(a) \right) p(n).$$

If $t+x \geq k_t+2$,

$$\mathbb{P}(W(t) > x) = \sum_{n=0}^{\infty} \left(\sum_{d=0}^{z(n)} \sum_{l=0}^{N(k_{t,x}-k_t-1)} \sum_{a=(N+d+l-n)^+}^{\infty} J_{t+x}\left(N, a+n-d-l-N\right) g\big(N(k_{t,x}-k_t-1), l\big) g\big(z(n), d\big) f_t(a) \right) p(n).$$

Here, $z(n)$ is defined in (6), $k_{t,x} = \lfloor t+x \rfloor$, $f_t(a)$ is the pmf of a Poisson distribution with mean $\Lambda G(t-k_t)$ evaluated at point a, $J_{t+x}(i,a)$ is the cdf of a binomial distribution with parameters $(i, \mu H(t+x-k_{t,x}))$ evaluated at point a, and $g(i,a)$ is the pmf of a binomial distribution with parameters (i,μ) evaluated at point a.

Proof of Proposition 8.

We extend (18) and obtain an equivalent expression for $\mathbb{P}(W(t) > x)$ for any $t \geq 0$, that is,

$$\mathbb{P}(W(t) > x) = \mathbb{P}(X(k_t) + A_{(k_t,t]} - N \geq D_{(k_t,t+x]}). \tag{57}$$

Since the distribution of $X(k_t)$ is $p(\cdot)$, it is sufficient for us to focus on evaluating the following conditional probability

$$r(t,x,n) = \mathbb{P}(D_{(k_t,t+x]} - A_{(k_t,t]} \leq n - N | X(k_t) = n). \tag{58}$$

Now, we consider three different scenarios to obtain $r(t,x,n)$.

Scenario 1: $k_t \leq t+x < k_t+1$. Under this scenario, we know that $k_t = k_{t,x} = \lfloor t+x \rfloor$. Lemma 1 tells us that conditioning on $X(k_t) = n$, $D_{(k_t,t+x]}$ follows a binomial distribution with parameters $\big(z(n), \mu H(t+x-k_{t,x})\big)$. Recall that $A_{(k_t,t]}$ follows a Poisson distribution with mean $\Lambda G(t-k_t)$ and is independent of $(X(k_t), D_{(k_t,t+x]})$. Therefore, to evaluate the conditional probability $r(t,x,n)$, we

need only *one* convolution between a binomial random variable and a Poisson random variable, that is,

$$\begin{aligned} r(t,x,n) &= \sum_{a=0}^{\infty} \mathbb{P}\left(D_{(k_t,t+x]} \le a+n-N | A_{(k_t,t]} = a, X(k_t) = n\right) \mathbb{P}\left(A_{(k_t,t]} = a | X(k_t) = n\right) \\ &= \sum_{a=a^*}^{\infty} J_{t+x}\big(z(n), a+n-N\big) f_t(a). \end{aligned} \tag{59}$$

Note that in (59), a starts from $a^* = (N-n)^+$ to ensure $a+n-N \ge 0$; otherwise, $J_{t+x}\big(z(n), a+n-N\big) = 0$ if $a < (N-n)^+$. Un-conditioning (59) with respect to $p(\cdot)$ gives the first equation in Proposition 8.

Scenario 2: $k_t + 1 \le t + x < k_t + 2$**.** Under this scenario, $D_{(k_t,t+x]} = D_{(k_t,k_t+1]} + D_{(k_t+1,t+x]}$. Conditioning on $X(k_t) = n$, Lemma 1 tells us that $D_{(k_t,k_t+1]}$ follows a binomial distribution with parameters $(z(n), \mu)$, while $A_{(k_t,t]}$ still follows a Poisson distribution with mean $\Lambda G(t-k_t)$ and is independent of $(X(k_t), D_{(k_t,k_t+1]})$. Therefore, we calculate the conditional probability $r(t,x,n)$ as

$$\begin{aligned} r(t,x,n) &= \sum_{d=0}^{z(n)} \sum_{a=0}^{\infty} \mathbb{P}\left(D_{(k_t+1,t+x]} \le a+n-d-N | A_{(k_t,t]} = a, D_{(k_t,k_t+1]} = d, X(k_t) = n\right) \\ &\qquad \cdot \mathbb{P}\left(D_{(k_t,k_t+1]} = d | A_{(k_t,t]} = a, X(k_t) = n\right) \mathbb{P}\left(A_{(k_t,t]} = a | X(k_t) = n\right) \\ &= \sum_{d=0}^{z(n)} \sum_{a=a^*}^{\infty} \mathbb{P}\left(D_{(k_t+1,t+x]} \le a+n-d-N | A_{(k_t,t]} = a, D_{(k_t,k_t+1]} = d, X(k_t) = n\right) \\ &\qquad \cdot g\big(z(n), d\big) f_t(a). \end{aligned} \tag{60}$$

Note that in (60), a starts from $a^* = (N+d-n)^+$ for the same reason discussed for (59). Thus, we have $a+n-d-N \ge 0$, which implies that

$$X(k_t+1) = X(k_t) + A_{(k_t,k_t+1]} - D_{(k_t,k_t+1]} \ge X(k_t) + A_{(k_t,t]} - D_{(k_t,k_t+1]} = n + a - d \ge N.$$

As a result, $Z(k_t+1) = N$ and every server is busy at k_t+1. Considering the revised system introduced in Section C.1, we know that the index set $\mathcal{B}(k_t+1) = \{1, \ldots, N\}$. Adapting (56), we can represent $D_{(k_t+1,t+x]}$ as

$$D_{(k_t+1,t+x]} = \sum_{j=1}^{N} \mathbb{1}_{\{\xi_{j,k_t+1}=1\}} \mathbb{1}_{\{h_{j,k_t+1} \le t+x-k_t-1\}}.$$

Therefore, conditioning on $A_{(k_t,t]} = a, D_{(k_t,k_t+1]} = d, X(k_t) = n$ such that $a+n-d-N \ge 0$, $D_{(k_t+1,t+x]}$ simply follows a binomial distribution with parameters $(N, \mu H(t+x-k_{t,x}))$, where $k_{t,x} = \lfloor t+x \rfloor = k_t + 1$. Consequently, we can further simplify (60) as

$$r(t,x,n) = \sum_{d=0}^{z(n)} \sum_{a=a^*}^{\infty} J_{t+x}\left(N, a+n-d-N\right) g\big(z(n), d\big) f_t(a). \tag{61}$$

Un-conditioning (61) with respect to $p(\cdot)$ gives the second equation in Proposition 8.

Note that (61) is in fact equivalent to calculating *two* convolutions among one Poisson random variable and two binomial random variables which are independent of each other. When we use normal approximations to approximate $r(t,x,n)$ in Section 4.2, we use this fact and approximate $\big(D_{(k_t,k_t+1]}+D_{(k_t+1,t+x]}\big)-A_{(k_t,t]}$ by the summation of three independent normal random variables.

Scenario 3: $t+x\geq k_t+2$**.** Under this scenario, $D_{(k_t,t+x]}=D_{(k_t,k_t+1]}+D_{(k_t+1,k_t+2]}+\cdots+D_{(k_t+k^*+1,t+x]}$ with $k^*=k_{t,x}-k_t-1$. Similar to (60) and (61), we compute the conditional probability $r(t,x,n)$ as

$$\begin{aligned}
r(t,x,n)=\sum_{d=0}^{z(n)}\sum_{d_1=0}^{N}\cdots\sum_{d_{k^*}=0}^{N}\sum_{a=0}^{\infty}&\mathbb{P}\big(D_{(k_t+k^*+1,t+x]}\leq a+n-d-d_1-\cdots-d_{k^*}-N\\
&\qquad|A_{(k_t,t]}=a,D_{(k_t,k_t+1]}=d,D_{(k_t+1,k_t+2]}=d_1,\ldots,D_{(k_t+k^*,k_t+k^*+1]}=d_{k^*},X(k_t)=n\big)\\
&\cdot\mathbb{P}\big(D_{(k_t+k^*,k_t+k^*+1]}=d_{k^*}|A_{(k_t,t]}=a,D_{(k_t,k_t+1]}=d,\ldots,D_{(k_t+k^*-1,k_t+k^*]}=d_{k^*-1},X(k_t)=n\big)\\
&\cdots\\
&\cdot\mathbb{P}\big(D_{(k_t+1,k_t+2]}=d_1|A_{(k_t,t]}=a,D_{(k_t,k_t+1]}=d,X(k_t)=n\big)\\
&\cdot\mathbb{P}\big(D_{(k_t,k_t+1]}=d|A_{(k_t,t]}=a,X(k_t)=n\big)\,\mathbb{P}\big(A_{(k_t,t]}=a|X(k_t)=n\big) \qquad (62)\\
=\sum_{d=0}^{z(n)}\sum_{d_1=0}^{N}\cdots\sum_{d_{k^*}=0}^{N}\sum_{a=a^*}^{\infty}&J_{t+x}\left(N,a+n-d-d_1-\cdots-d_{k^*}-N\right)g\big(N,d_1\big)\cdots g\big(N,d_{k^*}\big)g\big(z(n),d\big)f_t(a). \qquad (63)
\end{aligned}$$

To get (63), note that a starts from $a^*=(N+d+d_1+\cdots+d_{k^*}-n)^+$, which ensures $a+n-d-d_1-\cdots-d_{k^*}\geq N$. Consequently,

$$\begin{aligned}
X(k_t+1)&=X(k_t)+A_{(k_t,k_t+1]}-D_{(k_t,k_t+1]}\geq n+a-d\geq N.\\
&\cdots\\
X(k_t+k^*+1)&=X(k_t)+A_{(k_t,k_t+k^*+1]}-D_{(k_t,k_t+k^*+1]}\geq n+a-d-d_1-\cdots-d_{k^*}\geq N.
\end{aligned}$$

As a result, $Z(k_t+1)=\cdots=Z(k_{t,x})=N$. Extending the above argument for Scenario 2, we know that when conditioning on the corresponding events specified in (62), each of $D_{(k_t+1,k_t+2]},\ldots,D_{(k_t+k^*,k_t+k^*+1]}$ follows a binomial distribution with parameters (N,μ), and $D_{(k_t,x,t+x]}=D_{(k_t+k^*+1,t+x]}$ follows a binomial distribution with parameters $(N,\mu H(t+x-k_{t,x}))$. Then, (63) follows.

Similar to (61), note that (63) is in fact equivalent to calculating convolutions among one Poisson random variable and k^*+2 binomial random variables which are independent of each other. Since $D_{(k_t+1,k_t+2]},\ldots,D_{(k_t+k^*,k_t+k^*+1]}$ follows the same binomial distribution (when conditioning on appropriate events), their sum follows a binomial distribution with parameters (Nk^*,μ). Thus, we further simplify (63) as

$$r(t,x,n)=\sum_{d=0}^{z(n)}\sum_{l=0}^{Nk^*}\sum_{a=a^*}^{\infty}J_{t+x}\left(N,a+n-d-l-N\right)g\big(Nk^*,l\big)g\big(z(n),d\big)f_t(a), \qquad (64)$$

where $a^*=(N+d+l-n)^+$. Finally, un-conditioning (64) with respect to $p(\cdot)$ gives the last equation in Proposition 8. □

C.4. Proof for Proposition 2

Proof. Assume that the distribution of $X(0)$ is π. We prove below the periodicity of the customer count processes $X = \{X(t), t \geq 0\}$, the queue length process $\{Q(t), t \geq 0\}$, and the waiting time process $\{W(t), t \geq 0\}$, respectively.

Customer count processes. We first show the periodicity of the process $X = \{X(t), t \geq 0\}$. We start from showing that $X(t) =^d X(t+1)$ for $t \geq 0$ with $=^d$ denoting equal in distribution. Since $X(0)$ follows the stationary distribution π, we know the customer count at each midnight, $X(k)$, $k = 1, 2, \ldots$, also follows the stationary distribution π. Meanwhile, for any $t \geq 0$, we can represent $X(t)$ and $X(t+1)$ as

$$X(t) = X(k_t) + A_{(k_t,t]} - D_{(k_t,t]}, \tag{65}$$

$$X(t+1) = X(k_t+1) + A_{(k_t+1,t+1]} - D_{(k_t+1,t+1]}, \tag{66}$$

where $k_t = \lfloor t \rfloor$ is the most recent midnight before time t. From Lemma 1, we know that $D_{(k_t,t]}$ only depends on $X(k_t)$, and when conditioning on $X(k_t) = n$, $D_{(k_t,t]}$ follows a binomial distribution with parameters $\big(z(n), \mu H(t-k_t)\big)$. Similarly, $D_{(k_t+1,t+1]}$ follows the same binomial distribution when conditioning on $X(k_t+1) = n$. Since $X(k_t)$ and $X(k_t+1)$ have the same distribution π, we can check that $(X(k_t), D_{(k_t,t]})$ and $(X(k_t+1), D_{(k_t+1,t+1]})$ have the same distributions. The arrival process has the period of one day. Thus, both $A_{(k_t,t]}$ and $A_{(k_t+1,t+1]}$ follow a Poisson distribution with the same mean $\int_0^{t-k_t} \lambda(s)ds$. Moreover, $A_{(k_t,t]}$ is independent of $(X(k_t), D_{(k_t,t]})$ and $A_{(k_t+1,t+1]}$ is independent of $(X(k_t+1), D_{(k_t+1,t+1]})$. Consequently, from (65) and (66) we see that $X(t)$ and $X(t+1)$ have the same distributions. In fact, it is easy to check that their distributions take the same form as displayed in (12).

We can generalize the argument above to prove $(X(t_1), X(t_2)) =^d (X(t_1+1), X(t_2+1))$ for any $k \in \mathbb{Z}_+$ and $t_1, t_2 \in [k, k+1)$. To do so, we check that $(X(k), D_{(k,t_1]}, D_{(k,t_2]})$ and $(X(k+1), D_{(k+1,t_1+1]}, D_{(k+1,t_2+1]})$ have the same distribution, and then check their independence with the arrival quantities $(A_{(k,t_1]}, A_{(k,t_2]})$ and $(A_{(k+1,t_1+1]}, A_{(k+1,t_2+1]})$, respectively. Then, for any finite K-dimensional joint distribution, we can show $(X(t_1), \ldots, X(t_K)) =^d (X(t_1+1), \ldots, X(t_K+1))$ for $t_1, \ldots, t_K \in [k, k+1)$ and $k \in \mathbb{Z}_+$. Eventually, because the sample paths of $X(\cdot)$ are right-continuous, by a standard argument we can show that for any $k \in \mathbb{Z}_+$, $\{X(t), k \leq t < k+1\} =^d \{X(t), k+1 \leq t < k+2\}$; see, for example, [5, 14].

Queue length processes. Since $X(\cdot)$ is periodic in distribution, it is obvious that $Q(\cdot) = (X(\cdot) - N)^+$ is also periodic in distribution with one day as a period.

Waiting time processes. To show the periodicity of the waiting time process $\{W(t), t \geq 0\}$, we start by showing the tail distributions of $W(t)$ and $W(t+1)$ are the same for a given $t \geq 0$, that is,

$$\mathbb{P}(W(t) > x) = \mathbb{P}(W(t+1) > x) \text{ for any } x \geq 0.$$

This is easy to check by applying Proposition 8 and using the fact that $X(k_t)$ and $X(k_t+1)$ have the same distribution π.

The above implies that $W(t) =^d W(t+1)$. We can generalize the proof for Proposition 8 to show that $(W(t_1),\ldots,W(t_K)) =^d (W(t_1+1),\ldots,W(t_K+1))$ with $t_1,\ldots,t_K \in [k,k+1)$ for any given $k \in \mathbb{Z}_+$ and for any K-dimensional joint distribution. Then, similar to the proof for $\{X(t), t \geq 0\}$, we can show $\{W(t), k \leq t < k+1\} =^d \{W(t), k+1 \leq t < k+2\}$ for any $k \in \mathbb{Z}_+$, because the sample paths of $W(\cdot)$ are right-continuous. To check this right-continuity, recall that for a given $t \geq 0$,

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

is defined in (11). We can check that $W(\cdot)$ may only have positive jumps (from 0 to a positive number) at instances when an arrival occurs. The discharge events do not cause jumps in $W(\cdot)$. To see this, assume *one* customer is discharged at time t, and there is no arrival or another discharge occurring during $(t-\epsilon_0, t+\epsilon_0)$ for some $\epsilon_0 > 0$ sufficiently small. We then check that $W(\cdot)$ is continuous at this discharge instance t. Specifically, for any given $\epsilon \in (0, \epsilon_0)$, we consider three scenarios:

(i) If $X(t) - N \geq 0$, then $W(t) = \inf_{x\geq 0}\{D_{(t,t+x]} > X(t) - N\} \geq \epsilon_0$ because no discharge occurs during $(t, t+\epsilon_0)$. Fix a $t^* \in (t, t+\epsilon)$. For any $x \geq t^* - t$, we have $D_{(t^*,t+x]} = D_{(t,t+x]}$ and $X(t^*) = X(t)$ since there is no arrival or discharge during $(t, t+\epsilon)$. As a result,

$$\begin{aligned} W(t^*) &= \inf_{y\geq 0}\{D_{(t^*,t^*+y]} > X(t^*) - N\} \\ &= \inf_{y\geq 0}\{D_{(t^*,t+(t^*-t+y)]} > X(t^*) - N\} \\ &= \inf_{y\geq 0}\{D_{(t,t+(t^*-t+y)]} > X(t) - N\} \\ &= W(t) + t - t^* \in (W(t) - \epsilon, W(t)), \end{aligned}$$

where $W(t) - \epsilon > 0$. Now, fix a $t^* \in (t-\epsilon, t)$. Similarly, using the fact that $D_{(t^*,t+x]} = D_{(t,t+x]} + 1$ and $X(t^*) = X(t) + 1$, we can show that $W(t^*) = W(t) + t - t^* \in (W(t), W(t) + \epsilon)$. As a result, we have $|W(t^*) - W(t)| < \epsilon$ for any $t^* \in (t-\epsilon, t+\epsilon)$.

(ii) If $X(t) - N = -1$, then from the definition of $W(\cdot)$ in (11), we have $W(t) = 0$. Because there is no arrival or discharge during $(t, t+\epsilon)$, $X(t^*) = X(t)$ for $t^* \in (t, t+\epsilon)$. From (11), we get $W(t^*) = 0$ for $t^* \in (t, t+\epsilon)$. For $t^* \in (t-\epsilon, t)$, $X(t^*) - N = 0$ and $D_{(t^*,t]} = 1$. Also, for any arbitrarily small $\delta > 0$, we have $D_{(t^*,t-\delta]} = 0$. Again, using the definition of $W(\cdot)$ in (11), we get $W(t^*) = t - t^* \in (0, \epsilon)$. Thus, we have $|W(t^*) - W(t)| < \epsilon$ for any $t^* \in (t-\epsilon, t+\epsilon)$.

(iii) If $X(t) - N < -1$, using a similar argument as above, we can show that $W(t^*) = 0$ for any $t^* \in (t-\epsilon, t+\epsilon)$.

If there are k batch discharges occurring at t simultaneously, the same proof can be carried – we just need to change to use $D_{(t^*,t+x]} = D_{(t,t+x]} + k$ and $X(t^*) = X(t) + k$ for $t^* \in (t-\epsilon, t)$. At each arrival instance, we use the convention that the virtual customer arrives instantaneously *after* the real arrival event occurs. Under this convention, we can verify that the sample paths of $W(\cdot)$ are right-continuous. □

C.5. Proof for Proposition 6

Proof. Let $D^{(-h)}_{(0,t]}$ denote the number of discharges between 0 and t under the shifted discharge distribution $H^{(-h)}(\cdot)$. For a given $t \in [0,1)$ and $x \geq h/24$, we first prove (23) when $h/24 \leq t+x < 1$. We claim that

$$\mathbb{P}_\infty(W^{(-h)}(t) > x - h/24) = \mathbb{P}_\infty(X(0) + A_{(0,t]} - D^{(-h)}_{(0,t+x-h/24]} \geq N) \tag{67}$$

$$= \mathbb{P}_\infty(W(t) > x). \tag{68}$$

Equation (67) follows from (18). Now we argue that (68) holds. When $h/24 \leq t+x < 1$, conditioning on $X(0) = n$, Lemma 1 says that $D^{(-h)}_{(0,t+x-h/24]}$ follows a binomial distribution with parameters $\big(z(n), \mu H^{(-h)}(t+x-h/24)\big)$. Since $H^{(-h)}(t+x-h/24) = H(t+x)$ from (22), we can adapt the proof of Proposition 4 and show that the probability in (67) takes the same form as the right side of (17). Thus, (68) follows. For the case of $t+x \geq 1$, we can adapt the proof in Appendix C.3 and check that (68) still holds. □

C.6. Proof for Proposition 7

Proof. We will prove that under condition (21),

$$\mathbb{E}_\infty[W(t)] - \mathbb{E}_\infty[W^{(-h)}(t)] = \int_0^{h/24} \mathbb{P}_\infty(W(t) > x)dx. \tag{69}$$

Then, (26) follows from (69) and assumption (25) since

$$\begin{aligned}\int_0^{h/24} \mathbb{P}_\infty(W(t) > x)dx &= \int_0^{h/24} \mathbb{P}_\infty\left(x < W(t) \leq \frac{h}{24}\right)dx + \int_0^{h/24} \mathbb{P}_\infty\left(W(t) > \frac{h}{24}\right)dx \\ &= 0 + \frac{h}{24}\mathbb{P}_\infty\left(W(t) > \frac{h}{24}\right) \\ &= \frac{h}{24}\mathbb{P}_\infty\left(W(t) > 0\right). \end{aligned} \tag{70}$$

Note that the last equation in (70) holds because under condition (25), a delayed customer always waits more than h hours, and thus, $\mathbb{P}_\infty\left(W(t) > \frac{h}{24}\right)$ simply equals the delay probability $\mathbb{P}_\infty\left(W(t) > 0\right)$.

Finally, to prove (69), we have that

$$\begin{aligned}\mathbb{E}_\infty[W(t)] - \mathbb{E}_\infty[W^{(-h)}(t)] &= \int_0^\infty \mathbb{P}_\infty(W(t) > x)dx - \int_{h/24}^\infty \mathbb{P}_\infty(W(t) > x)dx \\ &= \int_0^{h/24} \mathbb{P}_\infty(W(t) > x)dx, \end{aligned} \tag{71}$$

where the first equation in (71) holds because

$$\begin{aligned}\mathbb{E}[W^{(-h)}(t)] &= \int_0^\infty \mathbb{P}(W^{(-h)}(t) > y)dy \\ &= \int_{h/24}^\infty \mathbb{P}(W^{(-h)}(t) > x - h/24)dx \quad (\text{let } x = y + h/24) \\ &= \int_{h/24}^\infty \mathbb{P}(W(t) > x)dx \quad (\text{applying Proposition 6}). \quad \square \end{aligned}$$

Appendix D: Proof for Lemma 2

Proof of Part (a) of Lemma 2. For two random variables U and V, we define their Kolmogorov distance $d_K(U,V)$ to be

$$d_K(U,V) = \sup_x |\mathbb{P}(U \le x) - \mathbb{P}(V \le x)|.$$

One can check that

$$d_K(aU, aV) = d_K(U,V), \tag{72}$$

$$d_K(U+B, U+B) \le d_K(U,V) \tag{73}$$

for any constant $a \neq 0$ and any random variable B that is independent of U and V.

Let U be a Poission random variable with mean λ. Then by Berry-Esseen theorem (see for example, Theorem 1.1 of [46]), one has

$$d_K(\hat{U}, Z_1) \le \frac{0.4785}{\sqrt{\lambda}}, \tag{74}$$

where $\hat{U} = (U-\lambda)/\sqrt{\lambda}$, Z_1 is a standard normal random variable. Similarly, let V be a binomial random variable with parameters (l,p), and one has

$$d_K(\hat{V}, Z_2) \le 0.4785\Big(\frac{p^2+q^2}{\sqrt{lpq}}\Big), \tag{75}$$

where $q = 1-p$, $\hat{V} = (V-lp)/\sqrt{lpq}$, and Z_2 is a standard normal variable independent of Z_1. Now, define

$$W = \frac{U-\lambda-(V-lp)}{\sqrt{\lambda+lpq}} = a\hat{U} - b\hat{V},$$
$$Z = aZ_1 - bZ_2,$$

where

$$a = \frac{\sqrt{\lambda}}{\sqrt{\lambda+lpq}} \quad \text{and} \quad b = \frac{\sqrt{lpq}}{\sqrt{\lambda+lpq}}.$$

Because $a^2+b^2 = 1$, Z is a standard normal random variable. First, assume that $a>0$ and $b>0$. We have

$$\begin{aligned} d_K(W,Z) &\le d_K(a\hat{U}-b\hat{V}, a\hat{U}-bZ_2) + d_K(a\hat{U}-bZ_2, aZ_1-bZ_2) \\ &\le d_K(-b\hat{V}, -bZ_2) + d_K(a\hat{U}, aZ_1) \\ &= d_K(\hat{V}, Z_2) + d_K(\hat{U}, Z_1) \\ &\le 0.4785\Big(\frac{1}{\sqrt{\lambda}} + \frac{p^2+q^2}{\sqrt{lpq}}\Big). \end{aligned} \tag{76}$$

Recall that for a given $t \in [0,1)$, $A_{(0,t]}$ is a Poisson random variable with mean $\Lambda G(t)$, and $D_{(0,t]}$ is a binomial random variable with parameters $\big(z(n), \mu H(t)\big)$ given that $X(0) = n$. Replacing U and V with $A_{(0,t]}$ and $D_{(0,t]}$, and replacing λ, l, and p with $\Lambda G(t)$, $z(n)$, and $\mu H(t)$ in (76), respectively, we then get (33). If a (or b) equals zero, then (33) follows from (72) and (75) (or (74)).

Proof of Part (b) of Lemma 2. Let X_∞ denote steady-state midnight customer count. Then, $(N-X_\infty)^+$ represents the steady-state number of idle servers in the midnight. We can prove that

$$\mathbb{E}\Big[(N-X_\infty)^+\Big] = (1-\rho)N. \tag{77}$$

Equality (77) is intuitive because $1-\rho$ equals the average proportion of time a server is idle in the midnight in steady-state. We leave the proof of (77) to the end of this lemma. Fix a $\kappa > 1$. Employing the Markov inequality, we have

$$\begin{aligned}\sum_{n\le N/\kappa} \pi(n) &= \mathbb{P}\{X_\infty \le N/\kappa\} \\ &= \mathbb{P}\{(N-X_\infty)^+ \ge N(1-1/\kappa)\} \\ &\le \frac{\kappa}{(\kappa-1)N}\mathbb{E}\Big[(N-X_\infty)^+\Big] = \frac{\kappa}{\kappa-1}(1-\rho).\end{aligned}$$

Now,

$$\begin{aligned}\sum_{n=1}^{\infty} \frac{1}{\sqrt{z(n)}}\pi(n) &= \sum_{n\le N/\kappa} \frac{1}{\sqrt{n}}\pi(n) + \sum_{n>N/\kappa} \frac{1}{\sqrt{z(n)}}\pi(n) \\ &\le \sum_{n\le N/\kappa} \pi(n) + \frac{\sqrt{\kappa}}{\sqrt{N}} \sum_{n>N/\kappa} \pi(n) \\ &\le \frac{\kappa}{\kappa-1}(1-\rho) + \frac{\sqrt{\kappa}}{\sqrt{N}} \\ &\le \frac{\kappa}{\kappa-1}(1-\rho) + \frac{\sqrt{\kappa\mu}}{\sqrt{\Lambda}}. \end{aligned} \tag{78}$$

Choosing $\kappa = 2$, we prove the lemma.

Finally, to prove (77), we use the basic adjoint relationship (BAR) for the midnight stationary distribution π, namely, for any bounded function $f : \mathbb{Z}_+ \to \mathbb{R}$,

$$\sum_{n=0}^{\infty} f(n)\pi(n) = \sum_{n=0}^{\infty} \mathbb{E}_n[f(n+A_0-D_0)]\pi(n). \tag{79}$$

Here, $\mathbb{E}_n$ is the expectation under $\mathbb{P}_n$, the probability distribution given that the starting midnight count is n; A_0 is a Poisson variable with mean Λ; and D_0 (under $\mathbb{P}_n$) is a binomial random variable with parameters $(z(n), \mu)$ and is independent of A_0 (recall that $z(n) = \min(n, N)$ is defined in (6)). Fix a $b > 0$. Then $f^{(b)}$, defined via

$$f^{(b)}(x) = \min(x, b) \quad \text{ for } x \in \mathbb{Z}_+,$$

is a bounded function. One can check that $f^{(b)}$ is a Lipschitz continuous satisfying

$$\big|f^{(b)}(x) - f^{(b)}(y)\big| \le |x-y|, \ \forall x, \ y \in \mathbb{Z}_+. \tag{80}$$

Applying $f^{(b)}$ to the BAR in (79), one has

$$\sum_{n=0}^{\infty} f^{(b)}(n)\pi(n) = \sum_{n=0}^{\infty} \mathbb{E}_n[f^{(b)}(n+A_0-D_0)]\pi(n),$$

or

$$\sum_{n=0}^{\infty}\Big(f^{(b)}(n)-\mathbb{E}_n[f^{(b)}(n+A_0-D_0)]\Big)\pi(n)=0$$

for each $b>0$. Therefore,

$$\lim_{b\to\infty}\sum_{n=0}^{\infty}\Big(f^{(b)}(n)-\mathbb{E}_n[f^{(b)}(n+A_0-D_0)]\Big)\pi(n)=0. \tag{81}$$

One can check that for any given $n\in\mathbb{Z}_+$,

$$\begin{aligned}
&\big|f^{(b)}(n)-\mathbb{E}_n[f^{(b)}(n+A_0-D_0)]\big| = \big|\mathbb{E}_n[f^{(b)}(n)-f^{(b)}(n+A_0-D_0)]\big|\\
&\le \mathbb{E}_n\big|f^{(b)}(n)-f^{(b)}(n+A_0-D_0)\big|\\
&\le \mathbb{E}_n|A_0-D_0|\\
&\le \mathbb{E}_n(A_0+D_0)\le \Lambda+N\mu,
\end{aligned}$$

where the constant $\Lambda+N\mu$ is independent of $b>0$, and the second inequality comes from (80). By the dominated convergence theorem, the limit in (81) can be moved into the summation, leading to

$$\sum_{n=0}^{\infty}\lim_{b\to\infty}\Big(f^{(b)}(n)-\mathbb{E}_n[f^{(b)}(n+A_0-D_0)]\Big)\pi(n)=0. \tag{82}$$

For each $n\in\mathbb{Z}_+$, $\lim_{b\to\infty} f^{(b)}(n)=n$, and

$$\lim_{b\to\infty}\mathbb{E}_n[f^{(b)}(n+A_0-D_0)]=\mathbb{E}_n(n+A_0-D_0)=n+\Lambda-z(n)\mu,$$

where the first equality follows from the monotone convergence theorem. Therefore, (82) implies that

$$\sum_{n=0}^{\infty}\Big(n-\big(n+\Lambda-z(n)\mu\big)\Big)\pi(n)=0,$$

from which we have

$$\sum_{n=0}^{\infty}z(n)\pi(n)=\Lambda/\mu=N\rho. \tag{83}$$

The left side is the expected number of busy servers in steady state. Therefore, the expected number of idle servers in steady state is $N-N\rho=(1-\rho)N$. □

Appendix E: Proof of Theorem 3.

In this section, we use Stein's method to prove Theorem 3. The framework of Stein's method includes four components: Poisson equation, generator coupling, derivative bounds, and moment bounds. In Section E.1, we give the main proof of the theorem and show how to use Poisson equation to do generator coupling (which is a more complete version than what we gave in Section 4.3.3 of the main paper). Then, in Section E.2, we prove the derivative bounds (Lemma 4). In Section E.3, we prove the moment bounds (Lemmas 5 and 6). Finally, in Section E.4, we prove several extra lemmas that are needed during the proofs in Sections E.1 to E.3.

E.1. Main proof

Proof. Fix an $h \in \mathrm{Lip}(1)$. Let $f = f_h$ be one solution to the Poisson equation

$$G_Y f(x) = h(x) - \mathbb{E}[h(Y_\infty)], \quad x \in \mathbb{R}, \tag{84}$$

where

$$G_Y f(x) = \frac{1}{2}\sigma_0^2 f''(x) + b(x) f'(x), \quad x \in \mathbb{R} \tag{85}$$

has been defined in (47) in the main paper, with

$$\begin{aligned} b(x) &= \delta(-\beta + x^-), \\ \sigma_0^2 &= 2\delta. \end{aligned}$$

Recall that

$$G_{\tilde{X}} f(x) = \mathbb{E}_n[f(x + \delta(A_0 - D_0)) - f(x)] \quad \text{for } x = \delta(n - N) \text{ and } n \in \mathbb{Z}_+, \tag{86}$$

where $\mathbb{E}_n$, A_0 and D_0 have been explained in Section 4.3.3 of the main paper. Let $\tilde{X}_k = \delta(X_k - N)$ for $k \in \mathbb{Z}_+$. It follows from (3) that

$$\tilde{X}_{k+1} = \tilde{X}_k + \delta A_k - \delta D_k, \quad k \geq 0,$$

from which one can verify

$$G_{\tilde{X}} f(x) = \mathbb{E}[f(\tilde{X}_{k+1}) | \tilde{X}_k = x] - f(x).$$

Thus, $G_{\tilde{X}}$ is the generator of the *scaled* midnight count process $\{\tilde{X}_k : k = 0, 1, 2, \ldots\}$. It follows from Proposition 2 of [18] that

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

for any $f : \mathbb{R} \to \mathbb{R}$ satisfying

$$\mathbb{E}\Big| f(\tilde{X}_\infty) \Big| < \infty.$$

For each $h \in \mathrm{Lip}(1)$, we verify this condition is satisfied for $f = f_h$ in Appendix E.4.1. Then, using the Poisson equation (84), we have

$$\begin{aligned} \mathbb{E}[h(\tilde{X}_\infty)] - \mathbb{E}[h(Y_\infty)] &= \mathbb{E}[G_Y f(\tilde{X}_\infty)] \\ &= \mathbb{E}\big[G_Y f(\tilde{X}_\infty) - G_{\tilde{X}} f(\tilde{X}_\infty)\big]. \end{aligned} \tag{87}$$

We have shown in (52) of the main paper that, by doing Taylor expansion for $G_{\tilde{X}} f(x)$ for each given $x = \delta(n - N)$, we get

$$\begin{aligned} G_{\tilde{X}} f(x) = G_Y f(x) &+ \frac{1}{2}\delta^2 \Big[-1 - \beta(2 - \delta) + (1 - \delta)(\beta - x^-) + (\beta - x^-)^2 \Big] f''(x) \\ &+ \frac{1}{6}\delta^3 \mathbb{E}_n[f'''(\xi)(A_0 - D_0)^3], \end{aligned} \tag{88}$$

where

$$|\xi - x| \le \delta |A_0 - D_0|.$$

Recall that $\delta = 1/\sqrt{N} \le 1$. From (87) and (88), one has

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

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

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

The rest of the proof follows from Lemmas 4 to 6 below.

LEMMA 4. *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*

$$\|f_h''\| \le C_3/\delta \quad \text{and} \quad \|f_h'''\| \le C_3/\delta. \tag{89}$$

LEMMA 5. *There exists a constant $C_4 = C_4(\beta) > 0$ such that*

$$\mathbb{E}\left|\beta - \tilde{X}_\infty^-\right| \le C_4 \quad \text{and} \quad \mathbb{E}[(\beta - \tilde{X}_\infty^-)^2] \le C_4. \tag{90}$$

LEMMA 6. *There exists a constant $C_5 = C_5(\beta) > 0$ such that*

$$\mathbb{E}\big[\mathbb{E}_n |A_0 - D_0|^3\big] \le C_5 \delta^{-3/2}. \tag{91}$$

The proof for Lemmas 4 to 6 are detailed in Sections E.2 and E.3 of this appendix. □

Remark. It is known, for example [46], convergence in the Wasserstein metric implies convergence in distribution. Note that when the space Lip(1) in (46) is replaced by $\mathcal{H}_K = \{1_{(-\infty,b]}(x) : b \in \mathbb{R}\}$, the corresponding distance is the Kolmogorov distance, denoted by $d_K(U,V)$. When V has a density that is bounded by a constant $\bar{v} > 0$, it is known that

$$d_K(U,V) \le \sqrt{2\bar{v} d_W(U,V)}. \tag{92}$$

One can check that the density of Y_∞ is bounded. Thus, Theorem 3 implies that $\tilde{X}_\infty$ and Y_∞ are close in the Kolmogorov distance as well.

E.2. Proof for Lemma 4 (derivative bounds)

We first state some results on f'' and f'''. We consider a more general form of the Poisson equation (50) by using $\sigma^2(x)$ to denote a general variance term in G_Y instead of using the constant variance σ_0^2. This gives

$$\frac{1}{2}\sigma^2(x) f''(x) + b(x) f'(x) = h(x) - \mathbb{E}[h(Y_\infty)], \tag{93}$$

Dividing $\frac{1}{2}\sigma^2(x)$ from both sides of (93), we get

$$f''(x)+\frac{2b(x)}{\sigma^2(x)}f'(x)=\frac{2\left(h(x)-\mathbb{E}[h(Y_\infty)]\right)}{\sigma^2(x)}. \tag{94}$$

Recall that the density of $Y(\infty)$ is proportional to

$$q(x)=\frac{1}{\sigma^2(x)}u(x) \tag{95}$$

with

$$u(x)=\exp\left(\int_0^x\frac{2b(s)}{\sigma^2(s)}ds\right).$$

Thus, multiplying $u(x)$ on both sides of (94), we get

$$\Big(u(x)f'(x)\Big)'=2\left(h(x)-\mathbb{E}[h(Y_\infty)]\right)q(x).$$

Therefore,

$$f'(x)=-\frac{2}{u(x)}\int_x^\infty\left(h(z)-\mathbb{E}[h(Y_\infty)]\right)q(z)dz\quad\text{ for all } x\in\mathbb{R}. \tag{96}$$

is one unique solution (up to a constant) satisfying

$$\lim_{x\to\infty}u(x)f'(x)=0. \tag{97}$$

Because the density of $Y(\infty)$ is proportional to $q(x)$,

$$\int_{-\infty}^\infty\left(h(z)-\mathbb{E}[h(Y_\infty)]\right)q(z)dz=0.$$

Thus, we can express $f'(x)$ in the following equivalent form:

$$f'(x)=\frac{2}{u(x)}\int_{-\infty}^x\left(h(z)-\mathbb{E}[h(Y_\infty)]\right)q(z)dz\quad\text{ for all } x\in\mathbb{R}. \tag{98}$$

In the following proof, we will use one of these two expressions in (96) and (98) for $f'(x)$ as we see fit.

To bound $f''(x)$, we differentiate (94) and get

$$f'''(x)+\frac{2b(x)}{\sigma^2(x)}f''(x)=2g(x), \tag{99}$$

where

$$g(x)=\Big(\frac{h(x)-\mathbb{E}h(Y_\infty)}{\sigma^2(x)}\Big)'-\Big(\frac{b(x)}{\sigma^2(x)}\Big)'f'(x).$$

Therefore, multiplying $u(x)$ on both sides, we get

$$(u(x)f''(x))'=2u(x)g(x). \tag{100}$$

If

$$\lim_{x\to\infty}u(x)f''(x)=0, \tag{101}$$

we have

$$f''(x) = -\frac{2}{u(x)} \int_x^\infty u(w)g(w)dw \quad \text{for all } x \in \mathbb{R}. \tag{102}$$

If

$$\lim_{x\to-\infty} u(x)f''(x) = 0, \tag{103}$$

we have

$$f''(x) = \frac{2}{u(x)} \int_{-\infty}^x u(w)g(w)dw \quad \text{for all } x \in \mathbb{R}. \tag{104}$$

Once we have estimate of $f''(x)$ from either (102) or (104), we can obtain the estimate of $f'''(x)$ via

$$f'''(x) = 2g(x) - \frac{2b(x)}{\sigma^2(x)} f''(x), \tag{105}$$

which comes from (99).

In the rest of the proof, for the function $h \in \text{Lip}(1)$, without loss of generality, we assume $h(0) = 0$. Therefore,

$$|h(x)| \le |x|, \quad \forall x \in \mathbb{R}.$$

E.2.1. When $x \ge 0$ We have $b(x) = b(0) = -\delta\beta$ and $\sigma^2(x) = \sigma_0^2 = 2\delta$ as two constants. Let

$$\gamma = -\frac{2b(0)}{\sigma_0^2} = \beta. \tag{106}$$

Thus,

$$u(x) = \exp(-\gamma x).$$

We first check condition (101) holds. From (94) we have for $x \ge 0$,

$$f''(x) = \gamma f'(x) + \frac{2\left(h(x) - \mathbb{E}[h(Y_\infty)]\right)}{\sigma_0^2}. \tag{107}$$

Because $|h(x)| \le |x| = x$ and $u(x) > 0$ for any $x \ge 0$,

$$|u(x)\left(h(x) - \mathbb{E}[h(Y_\infty)]\right)| \le \exp(-\gamma x)\, x + \exp(-\gamma x)\, |\mathbb{E}[h(Y_\infty)]|,$$

which implies that $\lim_{x\to\infty} u(x)\left(h(x) - \mathbb{E}[h(Y_\infty)]\right) = 0$. Together with (97), we get

$$\lim_{x\to\infty} u(x)f''(x) = 0.$$

Then, to bound $|f''|$, we have

$$\begin{aligned} g(x) &= \Big(\frac{h(x) - \mathbb{E}h(Y_\infty)}{\sigma_0^2}\Big)' \\ &= \frac{h'(x)}{\sigma_0^2} \end{aligned}$$

and

$$\begin{aligned} f''(x) &= -\frac{2}{u(x)}\int_x^\infty u(w)g(w)dw \\ &= -\frac{2}{\sigma_0^2\exp(-\gamma x)}\int_x^\infty h'(x)\exp(-\gamma w)dw \end{aligned}$$

from (102).

Because $|h'(w)|\le 1$ for any $w\in\mathbb{R}$, we have

$$\begin{aligned} |f''(x)| &\le \frac{2}{\sigma_0^2}\left|\frac{\int_x^\infty \exp(-\gamma w)dw}{\exp(-\gamma x)}\right| \\ &= \frac{2}{\gamma\sigma_0^2} \\ &= \frac{1}{\beta}\frac{1}{\delta}, \quad \forall x\ge 0. \end{aligned} \tag{108}$$

Then, from (105), we get

$$\begin{aligned} |f'''(x)| &= \left|\frac{2h'(x)}{\sigma_0^2}+\gamma f''(x)\right| \\ &\le \frac{4}{\sigma_0^2} \\ &= \frac{2}{\delta}, \quad \forall x\ge 0. \end{aligned} \tag{109}$$

Again, we have used the fact that $|h'(x)|\le 1$ to get the inequality.

E.2.2. When $x<0$ Recall that when $x<0$, $b(x)=-\delta(\beta+x)$ and $\sigma^2(x)=\sigma_0^2=2\delta$. Then,

$$\begin{aligned} u(x) &= \exp\left(\int_x^0 \frac{2\delta(\beta+s)}{2\delta}ds\right) \\ &= \exp\left(\frac{1}{2}(s+\beta)^2|_x^0\right) \\ &= \exp\left(\frac{1}{2}\beta^2\right)\exp\left(-\frac{1}{2}(x+\beta)^2\right) \end{aligned}$$

and

$$q(x) = \frac{1}{\sigma^2(x)}u(x) = \frac{1}{2\delta}\exp\left(\frac{1}{2}\beta^2\right)\exp\left(-\frac{1}{2}(x+\beta)^2\right).$$

Let

$$\phi_\beta(x) = \exp\left(-\frac{1}{2}(x+\beta)^2\right).$$

Correspondingly,

$$\begin{aligned} |f'(x)| &= \left|\frac{2}{u(x)}\int_{-\infty}^x (h(s)-\mathbb{E}[h(Y_\infty)])\,q(s)ds\right| \\ &\le \frac{2}{2\delta}\frac{\int_{-\infty}^x |s|\phi_\beta(s)ds}{\phi_\beta(x)} + \frac{2\mathbb{E}[|Y_\infty|]}{2\delta}\left|\frac{\int_{-\infty}^x \phi_\beta(s)ds}{\phi_\beta(x)}\right|. \end{aligned} \tag{110}$$

Here, for the inequality, we have used the fact that $|h(s)| \le |s|$ for all $s < 0$.

Using results from the Erlang-C model [8] (see the proof of Lemma 12 there), we know that both

$$\frac{\int_{-\infty}^{x} |s|\phi_\beta(s)ds}{\phi_\beta(x)} \quad \text{and} \quad \frac{\int_{-\infty}^{x} \phi_\beta(s)ds}{\phi_\beta(x)}$$

can be bounded by two constants that are independent of δ. Thus, there exists a constant C that is independent of δ and satisfies

$$|f'(x)| \le \frac{C}{\delta}. \tag{111}$$

Next, we check condition (103) holds. Again, from (94) we have for $x < 0$,

$$f''(x) = -\frac{2b(x)}{\sigma_0^2} f'(x) + \frac{2\left(h(x) - \mathbb{E}[h(Y_\infty)]\right)}{\sigma_0^2} \tag{112}$$

$$= (\beta + x) f'(x) + \frac{2\left(h(x) - \mathbb{E}[h(Y_\infty)]\right)}{\sigma_0^2}. \tag{113}$$

Because $|h(x)| \le |x| = -x$, $|f'(x)| \le \frac{C}{\delta}$, and

$$|u(x)x| \le \exp\left(\frac{1}{2}\beta^2\right) \exp\left(-\frac{1}{2}(x+\beta)^2\right)(-x)$$
$$\to 0 \text{ as } x \to -\infty.$$

Thus, one can check that

$$\lim_{x \to -\infty} u(x) f''(x) = 0.$$

Correspondingly, we have

$$f''(x) = \frac{2}{u(x)} \int_{-\infty}^{x} u(w) g(w) dw \quad \text{for all } x \in \mathbb{R}, \tag{114}$$

where

$$g(x) = \left(\frac{h(x) - \mathbb{E}h(Y_\infty)}{\sigma_0^2}\right)' - \left(\frac{b(x)}{\sigma_0^2}\right)' f'(x) \tag{115}$$

$$= \frac{1}{2\delta} h'(x) + \frac{1}{2} f'(x). \tag{116}$$

Since $|h'(x)| \le 1$ and $|f'(x)| \le C/\delta$,

$$|f''(x)| \le \frac{1}{2\delta}(1 + C) \frac{\int_{-\infty}^{x} \phi_\beta(s)ds}{\phi_\beta(x)}.$$

Thus, there exists a constant C' that is independent of δ and satisfies

$$|f''(x)| \le C'/\delta. \tag{117}$$

Finally, for $f'''(x)$, we have

$$f'''(x) = 2g(x) - \frac{2b(x)}{\sigma_0^2} f''(x)$$
$$= 2g(x) + (x + \beta) f''(x),$$

where $|g(x)|$ is bounded by $(1+C)/(2\delta)$, and

$$|(x+\beta)f''(x)| \le (|x|+\beta)\frac{1}{2\delta}(1+C)\frac{\int_{-\infty}^{x}\phi_\beta(s)ds}{\phi_\beta(x)}.$$

Again, using results from the Erlang-C model [8] (see the proof of Lemma 12 there), we can bound

$$\frac{|x|\int_{-\infty}^{x}\phi_\beta(s)ds}{\phi_\beta(x)}$$

by some constant that is independent of δ. Eventually, there exits a constant C'' that is independent of δ and satisfies

$$|f'''(x)| \le \frac{C''}{\delta}. \tag{118}$$

With (108), (109), (117), (118), we have proved that there exists a constant $C_3 = C_3(\beta)$ such that $\|f''\| \le C_3/\delta$ and $\|f'''\| \le C_3/\delta$.

E.3. Moment bounds

E.3.1. Proof for Lemma 5.

Proof. To bound $\mathbb{E}[\left|\beta - \tilde{X}_\infty^-\right|]$ and $\mathbb{E}[(\beta - \tilde{X}_\infty^-)^2]$, it is sufficient to show

$$\mathbb{E}[\tilde{X}_\infty^2 \mathbb{1}(\tilde{X}_\infty < 0)] \le C_4' \tag{119}$$

for some constant $C_4' = C_4'(\beta)$, which also implies

$$\mathbb{E}[\left|\tilde{X}_\infty\right| \mathbb{1}(\tilde{X}_\infty < 0)] \le \sqrt{C_4'}$$

by Jensen's inequality.

Now, to prove (119), we consider the following Lyapunov function

$$V(x) = x^2.$$

For a given x such that $x = \delta(n-N)$ with $n \in \mathbb{Z}_+$, we have

$$\begin{aligned} G_{\tilde{X}}V(x) &= \mathbb{E}_n[V\big(x+\delta(A_0-D_0)\big)] - V(x) \\ &= \mathbb{E}_n\left[(x+\delta(A_0-D_0))^2 - x^2\right] \\ &= \mathbb{E}_n\left[2\delta(A_0-D_0)x + \delta^2(A_0-D_0)^2\right] \\ &= 2x\cdot\delta\mathbb{E}_n[A_0-D_0] + \delta^2\mathbb{E}_n[(A_0-D_0)^2]. \end{aligned}$$

Here, $\mathbb{E}_n$ is the expectation under the probability distribution given that the starting midnight count is n with $x = \delta(n-N)$; A_0 is a Poisson variable with mean Λ, and D_0 is a binomial random variable with parameters $(\min(n,N),\mu)$ and is independent of A_0. Correspondingly,

$$\begin{aligned} \delta\mathbb{E}_n[A_0-D_0] &= -\delta(\beta - x^-), \\ \delta^2\mathbb{E}_n[(A_0-D_0)^2] &= 2\delta + \delta^2\Big(-1-\beta(2-\delta) + (1-\delta)(\beta-x^-) + (\beta-x^-)^2\Big) \\ &= 2\delta + \delta^2(-1-\beta(2-\delta)) + \delta^2(1-\delta)(\beta-x^-) + \delta^2(\beta-x^-)^2. \end{aligned}$$

Thus,

$$G_{\tilde{X}}V(x) = -2\delta(\beta - x^-)x + \delta^2(1-\delta)(\beta - x^-) + \delta^2(\beta - x^-)^2 + 2\delta + \delta^2(-1-\beta(2-\delta)).$$

Dividing both sides by δ, we get

$$\begin{aligned}
G_{\tilde{X}}V(x)/\delta &= -2(\beta - x^-)x + \delta(1-\delta)(\beta - x^-) + \delta(\beta - x^-)^2 + 2 + \delta(-1-\beta(2-\delta)) \\
&= \mathbb{1}(x<0)\Big((\delta-2)x^2 + (1-\delta)(\delta-2\beta)x\Big) \\
&\quad + \mathbb{1}(x\geq 0)\Big(-2\delta\beta x\Big) \\
&\quad + 2 + \delta(-1-\beta(2-\delta)) + \delta(1-\delta)\beta + \delta\beta^2 \\
&\leq \mathbb{1}(x<0)\Big((\delta-\frac{3}{2})x^2 + \frac{1}{2}(1-\delta)^2(\delta-2\beta)^2\Big) \\
&\quad + \mathbb{1}(x\geq 0)\Big(-2\delta\beta x\Big) \\
&\quad + 2 + \delta(-1-\beta(2-\delta)) + \delta(1-\delta)\beta + \delta\beta^2, \qquad (120)
\end{aligned}$$

where the inequality comes from the fact that

$$\begin{aligned}
&(\delta-2)x^2 + (1-\delta)(\delta-2\beta)x \\
&\leq (\delta-2)x^2 + (1-\delta)(\delta-2\beta)x + \frac{1}{2}\Big(x - (1-\delta)(\delta-2\beta)\Big)^2 \\
&= (\delta-\frac{3}{2})x^2 + \frac{1}{2}(1-\delta)^2(\delta-2\beta)^2.
\end{aligned}$$

For the function $V(x) = x^2$, we can check that $\mathbb{E}\Big|V(\tilde{X}_\infty)\Big| < \infty$ is satisfied using results from Section E.4.1. Thus,

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

As a result, replacing x by $\tilde{X}_\infty$ in (120) and then taking expectation with respect to $\tilde{X}_\infty$ gives us

$$\begin{aligned}
0 &= \mathbb{E}[G_{\tilde{X}}V(\tilde{X}_\infty)/\delta] \\
&\leq (\delta-\frac{3}{2})\mathbb{E}\Big[\tilde{X}_\infty^2\mathbb{1}(\tilde{X}_\infty<0)\Big] \\
&\quad + \mathbb{E}\Big[\big(-2\delta\beta\tilde{X}_\infty\big)\mathbb{1}(\tilde{X}_\infty\geq 0)\Big] \\
&\quad + 2 + \delta(-1-\beta(2-\delta)) + \delta(1-\delta)\beta + \delta\beta^2 + \frac{1}{2}(1-\delta)^2(\delta-2\beta)^2.
\end{aligned}$$

Then, we have

$$\begin{aligned}
&(\frac{3}{2}-\delta)\mathbb{E}\Big[\tilde{X}_\infty^2\mathbb{1}(\tilde{X}_\infty<0)\Big] \\
&\leq 2 + \delta(-1-\beta(2-\delta)) + \delta(1-\delta)\beta + \delta\beta^2 + \frac{1}{2}(1-\delta)^2(\delta-2\beta)^2.
\end{aligned}$$

Since $\delta = 1/\sqrt{N} \leq 1$, we have

$$\begin{aligned}
&\mathbb{E}\Big[\mathbb{1}(\tilde{X}_\infty<0)(\tilde{X}_\infty)^2\Big] \\
&\leq 2\Big(2 + \delta(-1-\beta(2-\delta)) + \delta(1-\delta)\beta + \delta\beta^2 + \frac{1}{2}(1-\delta)^2(\delta-2\beta)^2\Big) \\
&\leq 2\Big(2 + \beta + \beta^2 + 2\beta^2\Big),
\end{aligned}$$

where the right-hand side is a constant that is independent of δ. □

E.3.2. Proof for Lemma 6.

Proof. We want to bound $\mathbb{E}[\mathbb{E}_n[|A_0 - D_0|^3]]$, where the inside expectation $\mathbb{E}_n$ is the conditional expectation given that the midnight count is n, while the outside expectation is taken with respect to all n.

First, for a given n, we have

$$\begin{aligned}\mathbb{E}_n[|A_0 - D_0|^3] &= \mathbb{E}_n[|(A_0 - \Lambda) + (z(n)\mu - D_0) + (\Lambda - z(n)\mu)|^3] \\ &\le 9\left(\mathbb{E}_n[|A_0 - \Lambda|^3] + \mathbb{E}_n[|D_0 - z(n)\mu|^3] + \mathbb{E}_n[|\Lambda - z(n)\mu|^3]\right) \\ &\le 36\left(\left(\sqrt{\Lambda}\right)^3 + \left(\sqrt{z(n)\mu}\right)^3\right) + 9|\Lambda - z(n)\mu|^3 + 72,\end{aligned}$$

where $z(n) = \min(n, N)$, the first inequality comes from the well-known Loeve's c_r-inequality, namely, $\mathbb{E}|X_1 + \cdots + X_n|^r \le n^{r-1}\sum_{i=1}^n \mathbb{E}|X_i|^r$, and the second inequality comes from Lemma 7 below.

Then, taking expectation over all possible n, we have

$$\begin{aligned}\mathbb{E}\left(\mathbb{E}_n[|A - D_n|^3]\right) &\le 36\left(\Lambda^{\frac{3}{2}} + (N\mu)^{\frac{3}{2}}\right) + 9\mathbb{E}\left[|\Lambda - z(n)\mu|^3\right] + 72 \\ &\le 72\left(\sqrt{N}\right)^{\frac{3}{2}} + 9\mathbb{E}\left[|z(n)\mu - \Lambda|^3\right] + 72 \\ &= 72\delta^{-\frac{3}{2}} + 9\mathbb{E}\left[|z(n)\mu - \Lambda|^3\right] + 72.\end{aligned}$$

The first inequality comes from the fact that $z(n) \le N$ for all n, and the second inequality comes from our assumptions that $\mu = 1/\sqrt{N}$ and $\Lambda = \sqrt{N} - \beta$.

For the term $\mathbb{E}\left[|z(n)\mu - \Lambda|^3\right]$, we have that

$$\begin{aligned}\mathbb{E}\left[|z(n)\mu - \Lambda|^3\right] &= \mathbb{E}\left[|N\mu - \Lambda - \mu(n-N)^-|^3\right] \\ &= \mathbb{E}\left[\left|\beta - \tilde{X}_\infty^-\right|^3\right] \\ &\le 4\beta^3 + 4\mathbb{E}\left[\left|\tilde{X}_\infty^-\right|^3\right] \\ &= 4\beta^3 + 4\mathbb{E}\left[\left|\tilde{X}_\infty\right|^3 \mathbb{1}(\tilde{X}_\infty < 0)\right].\end{aligned}$$

From Lemma 8 below, we know that $\mathbb{E}\left[|z(n)\mu - \Lambda|^3\right]$ can be bounded by some constant times $N^{\frac{3}{8}}$, which is of a smaller order than $\delta^{-\frac{3}{2}} = N^{\frac{3}{4}}$. Thus, there exists a constant C_5 such that $\mathbb{E}\left[\mathbb{E}_n|A_0 - D_0|^3\right] \le C_5\delta^{-3/2}$.

We end this section by proving Lemmas 7 and 8.

Lemma 7. *Let A be a Poisson random variable with mean Λ, and B be a binomial random variable with parameters $(z(n), 1/\sqrt{N})$. Then,*

$$\mathbb{E}\left(|A - \Lambda|\right)^3 \le (2\Lambda)^{3/2} + 4 \quad \textit{for all } \Lambda \ge 0, \tag{121}$$

$$\mathbb{E}\left(\left|B - z(n)/\sqrt{N}\right|\right)^3 \le \left(4\big(z(n)\big)^2/N\right)^{3/4} + 4 \quad \textit{for all } z(n) \ge 0 \textit{ and } N \ge 1. \tag{122}$$

LEMMA 8. *Therefore exists a constant* $C_6 = C_6(\beta) > 0$ *such that*

$$\mathbb{E}[\left|\tilde{X}_\infty\right|^3 \mathbb{1}(\tilde{X}_\infty < 0)] \le C_6 N^{\frac{3}{8}}.$$

The proof of Lemma 8 is in Section E.4.2. To prove Lemma 7, we use the facts that, see for example Page 539 of [42],

$$\mathbb{E}[(A-\Lambda)^4] = \Lambda + 3\Lambda^2,$$

and

$$\mathbb{E}[(B - z(n)p)^4] = 3\big(z(n)\big)^2 p^2 q^2 + z(n)pq(1-6pq),$$

where $p = 1/\sqrt{N}$ and $q = 1-p$. When $\Lambda \ge 1$,

$$\mathbb{E}[(A-\Lambda)^4] \le 4\Lambda^2,$$

and thus, $\mathbb{E}[(A-\Lambda)^3] \le (2\Lambda)^{3/2}$ from applying Jensen's inequality. When $\Lambda < 1$, we have $\mathbb{E}[(A-\Lambda)^4] \le 4$ and correspondingly, $\mathbb{E}[(A-\Lambda)^3] \le (4)^{3/4} \le 4$, which completes the proof for (121). One can prove (122) in a similar way. □

E.4. Proofs for lemmas in Sections E.1 to E.3

E.4.1. Check condition (49). To check (49) is satisfied, we use Proposition 2 in [18]. Fix an $h \in \mathrm{Lip}(1)$. We need to prove that for the Poisson equation solution $f = f_h$, the following condition

$$\mathbb{E}\left|f(\tilde{X}_\infty)\right| < \infty \tag{123}$$

holds, where $\tilde{X}_\infty$ is the steady-state customer count for the scaled midnight count process $\{\frac{X_k - N}{\sqrt{N}}, k \ge 0\}$.

To prove (123), it is sufficient to prove

$$\mathbb{E}[X_\infty^2] < \infty. \tag{124}$$

To see this, we first note that (124) implies that $\mathbb{E}[X_\infty] < \infty$ and $\mathbb{E}[\tilde{X}_\infty^2] < \infty$. The latter implies $\mathbb{E}\left|\tilde{X}_\infty\right| < \infty$. Setting $\mu_X = \mathbb{E}[\tilde{X}_\infty]$, one has

$$\begin{aligned} |f(x)| &= |f(\mu_X + x - \mu_X)| \\ &= \left|f(\mu_X) + f'(\mu_X)(x-\mu_X) + f''(\xi)(x-\mu_X)^2\right| \\ &\le |f(\mu_X)| + |f'(\mu_X)||x-\mu_X| + |f''(\xi)|(x-\mu_X)^2. \end{aligned}$$

Thus,

$$\mathbb{E}\left|f(\tilde{X}_\infty)\right| \le |f(\mu_X)| + |f'(\mu_X)|\mathbb{E}\left|\tilde{X}_\infty - \mu_X\right| + \mathbb{E}\left[|f''(\xi)|(\tilde{X}_\infty - \mu_X)^2\right],$$

from which and Lemma 4, one proves (123).

We now prove (124). For that, we consider the following Lyapunov function

$$V(n) = n^3 + \frac{3v_2}{2v_1}n^2, \quad \text{for } n \in \mathbb{Z}_+,$$

where

$$v_1 = N\mu - \Lambda, \quad v_2 = \Lambda + N\mu(1-\mu) + v_1^2$$

are two positive constants. Note that $v_1 > 0$ because $\rho < 1$. For a given $n \geq N$, we have

$$\begin{aligned}
&\mathbb{E}_n[V(n + A_0 - D_0)] - V(n) \\
&= \mathbb{E}_n\left[(n + A_0 - D_0)^3 + \frac{3v_2}{2v_1}(n + A_0 - D_0)^2 - n^3 - \frac{3v_2}{2v_1}n^2\right] \\
&= \mathbb{E}_n\left[3(A_0 - D_0)n^2 + 3(A_0 - D_0)^2 n + (A_0 - D_0)^3 + \frac{3v_2}{2v_1}2(A_0 - D_0)n + \frac{3v_2}{2v_1}(A_0 - D_0)^2\right].
\end{aligned}$$

As introduced for (79), $\mathbb{E}_n$ is the expectation under the probability distribution given that the starting midnight count is n; A_0 is a Poisson variable with mean Λ, and D_0 is a binomial random variable with parameters $(\min(n, N), \mu)$ and is independent of A_0. Since $n \geq N$, D_0 is simply a binomial(N, μ) here. Correspondingly,

$$\begin{aligned}
\mathbb{E}_n[A_0 - D_0] &= \Lambda - N\mu = -v_1, \\
\mathbb{E}_n[(A_0 - D_0)^2] &= \Lambda + N\mu(1-\mu) + (\Lambda - N\mu)^2 = v_2.
\end{aligned}$$

Also,

$$\mathbb{E}_n[(A_0 - D_0)^3] = v_3$$

is a constant for all $n \geq N$. As a result,

$$\begin{aligned}
&\mathbb{E}_n[V(n + A_0 - D_0)] - V(n) \\
&= -3v_1 n^2 + 3v_2 n + v_3 + \frac{3v_2}{2v_1}2(-v_1)n + \frac{3v_2}{2v_1}v_2 \\
&= -3v_1 n^2 + \left(v_3 + \frac{3v_2^2}{2v_1}\right), \quad n \geq N.
\end{aligned} \tag{125}$$

For $n = 0, 1, \ldots, N-1$, let

$$M^* = \max_{n=0,1,\ldots,N-1}\Big(\mathbb{E}_n[V(n + A_0 - D_0)] - V(n)\Big). \tag{126}$$

Then, combining (125) and (126), we have that, for any state $n \geq 0$,

$$\begin{aligned}
\mathbb{E}_n[V(n + A_0 - D_0)] - V(n) &\leq \max(M^*, 0) + \max\left(v_3 + \frac{3v_2^2}{2v_1}, 0\right) + 3v_1 N^2 - 3v_1 n^2 \\
&= M^\dagger - f(n),
\end{aligned}$$

where

$$\begin{aligned}
f(n) &= 3v_1 n^2, \\
M^\dagger &= \max(M^*, 0) + \max\left(v_3 + \frac{3v_2^2}{2v_1}, 0\right) + 3v_1 N^2 > 0,
\end{aligned}$$

and $f(\cdot)$ is a nonnegative function. Then, applying the comparison theorem (see for example, Theorem 14.2.2 in [41]), we get

$$\mathbb{E}[f(X_\infty)] \le M^\dagger,$$

which implies that

$$\mathbb{E}[X_\infty^2] \le \frac{M^\dagger}{3v_1},$$

completing the proof. □

E.4.2. Third moment partial bounds. To prove Lemma 8, we need:

$$\mathbb{E}[\left|\tilde{X}_\infty\right|^3 \mathbb{1}(\tilde{X}_\infty < 0)] \le C_6 N^{\frac{3}{8}}. \tag{127}$$

It is sufficient to prove

$$\mathbb{E}[(\tilde{X}_\infty)^4 \mathbb{1}(\tilde{X}_\infty < 0)] \le C_6\sqrt{N} \tag{128}$$

and then apply Jensen's inequality.

To prove (128), we consider the following function

$$V(x) = x^4 + a_1 x^3 + a_2 x^2, \tag{129}$$

where a_1 and a_2 are two constants that will be determined via (130) and (131) below.

The proof is similar to that in Section E.3.1. For a given x such that $x = \delta(n-N)$ with $n \in \mathbb{Z}_+$, we have

$$\begin{aligned}
G_{\tilde{X}}V(x) &= \mathbb{E}_n[V\big(x+\delta(A_0-D_0)\big)] - V(x)\\
&= \mathbb{E}_n\Big[(x+\delta(A_0-D_0))^4 + a_1(x+\delta(A_0-D_0))^3 + a_2(x+\delta(A_0-D_0))^2\\
&\quad - x^4 - a_1x^3 - a_2x^2\Big]\\
&= \mathbb{E}_n\Big[4\delta(A_0-D_0)x^3 + 6\delta^2(A_0-D_0)^2x^2 + 4\delta^3(A_0-D_0)^3x + \delta^4(A_0-D_0)^4\\
&\quad + 3a_1\delta(A_0-D_0)x^2 + 3a_1\delta^2(A_0-D_0)^2x + a_1\delta^3(A_0-D_0)^3\\
&\quad + 2a_2\delta(A_0-D_0)x + a_2\delta^2(A_0-D_0)^2\Big]\\
&= 4x^3\cdot\delta\mathbb{E}_n[A_0-D_0] + \Big(6\delta^2\mathbb{E}_n[(A_0-D_0)^2] + 3a_1\delta\mathbb{E}_n[(A_0-D_0)]\Big)x^2\\
&\quad + \Big(4\delta^3\mathbb{E}_n[(A_0-D_0)^3] + 3a_1\delta^2\mathbb{E}_n[(A_0-D_0)^2] + 2a_2\delta\mathbb{E}_n[(A_0-D_0)]\Big)x\\
&\quad + \delta^4\mathbb{E}_n[(A_0-D_0)^4] + a_1\delta^3\mathbb{E}_n[(A_0-D_0)^3] + a_2\delta^2\mathbb{E}_n[(A_0-D_0)^2].
\end{aligned}$$

Again, $\mathbb{E}_n$ is the conditional expectation given that the starting midnight count is n with $x = \delta(n-N)$, and as we obtained before,

$$\delta\mathbb{E}_n[A_0-D_0] = -\delta(\beta - x^-).$$

Also, when $x \geq 0$, $n \geq N$, and D_0 is binomial (N, μ), independent of the value of x. We write the conditional expectation $\mathbb{E}_n$ as $\mathbb{E}_N$ for all $n \geq N$. Correspondingly, we can choose a_1 and a_2 appropriately such that for all $x \geq 0$ ($n \geq N$),

$$6\delta^2 \mathbb{E}_N[(A_0 - D_0)^2] + 3a_1\delta \mathbb{E}_N[(A_0 - D_0)] = 0 \tag{130}$$

and

$$4\delta^3 \mathbb{E}_N[(A_0 - D_0)^3] + 3a_1\delta^2 \mathbb{E}_N[(A_0 - D_0)^2] + 2a_2\delta \mathbb{E}_N[(A_0 - D_0)] = 0. \tag{131}$$

Moreover, one can check that $a_1 = 2\delta \mathbb{E}_N[(A_0 - D_0)^2]/\mathbb{E}_N[(D_0 - A_0)] \to 4/\beta$ when $N \to \infty$, and $a_2 = \Big(4\delta^2 \mathbb{E}_N[(A_0 - D_0)^3] + 3a_1\delta \mathbb{E}_N[(A_0 - D_0)^2]\Big)/2\mathbb{E}_N[(D_0 - A_0)]$ also converges to a constant that only depends on β when $N \to \infty$.

Then, we have

$$\begin{aligned} G_{\bar{X}}V(x) = & -4\delta(\beta - x^-)x^3 + \Big(6\delta^2 \mathbb{E}_n[(A_0 - D_0)^2] + 3a_1\delta \mathbb{E}_n[(A_0 - D_0)]\Big)x^2 \mathbb{1}(x < 0) \\ & + \Big(4\delta^3 \mathbb{E}_n[(A_0 - D_0)^3] + 3a_1\delta^2 \mathbb{E}_n[(A_0 - D_0)^2] + 2a_2\delta \mathbb{E}_n[(A_0 - D_0)]\Big)x \mathbb{1}(x < 0) \\ & + \delta^4 \mathbb{E}_n[(A_0 - D_0)^4] + a_1\delta^3 \mathbb{E}_n[(A_0 - D_0)^3] + a_2\delta^2 \mathbb{E}_n[(A_0 - D_0)^2]. \end{aligned}$$

Dividing δ from both sides, we get

$$\begin{aligned} G_{\bar{X}}V(x)/\delta = & -4(\beta - x^-)x^3 + \Big(6\delta \mathbb{E}_n[(A_0 - D_0)^2] + 3a_1 \mathbb{E}_n[(A_0 - D_0)]\Big)x^2 \mathbb{1}(x < 0) \\ & + \Big(4\delta^2 \mathbb{E}_n[(A_0 - D_0)^3] + 3a_1\delta \mathbb{E}_n[(A_0 - D_0)^2] + 2a_2 \mathbb{E}_n[(A_0 - D_0)]\Big)x \mathbb{1}(x < 0) \\ & + \delta^3 \mathbb{E}_n[(A_0 - D_0)^4] + a_1\delta^2 \mathbb{E}_n[(A_0 - D_0)^3] + a_2\delta \mathbb{E}_n[(A_0 - D_0)^2]. \end{aligned} \tag{132}$$

Note that for a given x,

$$\begin{aligned} -4(\beta - x^-)x^3 &= -4(\beta x^3 + x^4)\mathbb{1}(x < 0) - 4\beta x^3 \mathbb{1}(x \geq 0) \\ &\leq -4(\beta x^3 + x^4)\mathbb{1}(x < 0), \end{aligned}$$

because $4\beta x^3$ is positive when $x \geq 0$. Moreover, we have

$$\begin{aligned} -4(\beta x^3 + x^4)\mathbb{1}(x < 0) &\leq -4(\beta x^3 + x^4)\mathbb{1}(x < 0) + (x + \beta)^4 \mathbb{1}(x < 0) \\ &= (-3x^4 + 6\beta^2 x^2 + 4\beta^3 x + \beta^4)\mathbb{1}(x < 0). \end{aligned}$$

Thus,

$$\begin{aligned} G_{\bar{X}}V(x)/\delta \leq & -3x^4 \mathbb{1}(x < 0) + \Big(6\delta \mathbb{E}_n[(A_0 - D_0)^2] + 3a_1 \mathbb{E}_n[(A_0 - D_0)] + 6\beta^2\Big)x^2 \mathbb{1}(x < 0) \\ & + \Big(4\delta^2 \mathbb{E}_n[(A_0 - D_0)^3] + 3a_1\delta \mathbb{E}_n[(A_0 - D_0)^2] + 2a_2 \mathbb{E}_n[(A_0 - D_0)] + 4\beta^3\Big)x \mathbb{1}(x < 0) \\ & + \delta^3 \mathbb{E}_n[(A_0 - D_0)^4] + a_1\delta^2 \mathbb{E}_n[(A_0 - D_0)^3] + a_2\delta \mathbb{E}_n[(A_0 - D_0)^2] + \beta^4. \end{aligned} \tag{133}$$

Because $\Lambda = \sqrt{N} - \beta \le \sqrt{N}$ and $\mu = \delta = 1/\sqrt{N}$, we have

$$\begin{aligned}
\mathbb{E}_n[|A_0 - D_0|] &\le \Lambda + N\mu \le 2\sqrt{N},\\
\delta\mathbb{E}_n[(A_0 - D_0)^2] &\le \delta(\Lambda + \Lambda^2 + N\mu(1-\mu) + N^2\mu^2)\\
&\le \frac{1}{\sqrt{N}}(\sqrt{N} + N + \sqrt{N} + N)\\
&\le 4\sqrt{N},\\
\delta^2\mathbb{E}_n[|A_0 - D_0|^3] &\le 4\delta^2(\mathbb{E}_n[A_0^3] + \mathbb{E}_n[D_0^3])\\
&\le C_7\delta^2 N^3\mu^3 = C_7\sqrt{N},\\
\delta^3\mathbb{E}_n[(A_0 - D_0)^4] &\le 8\delta^3\left(\mathbb{E}_n[A_0^4] + \mathbb{E}_n[D_0^4]\right)\\
&\le C_8\delta^3 N^4\mu^4 = C_8\sqrt{N},
\end{aligned}$$

where C_7 and C_8 are two constants. Note that the bound for $\mathbb{E}_n[|A_0 - D_0|^3]$ here is much looser than that stated in Lemma 6. Since we have proved that $\mathbb{E}[\tilde{X}_\infty^2 \mathbb{1}(\tilde{X}_\infty < 0)]$ and $\mathbb{E}[\tilde{X}_\infty \mathbb{1}(\tilde{X}_\infty < 0)]$ are bounded by some constants that are independent of δ or N, let $M_N = C\sqrt{N}$ with C large enough, we then get

$$0 = \mathbb{E}[G_{\tilde{X}} V(\tilde{X}_\infty)/\delta] \le -3\mathbb{E}[\tilde{X}_\infty^4 \mathbb{1}(\tilde{X}_\infty < 0)] + M_N \tag{134}$$

by replacing x with $\tilde{X}_\infty$ and taking expectation for (133). As a result, there exist a constant C_6 such that $\mathbb{E}[\tilde{X}_\infty^4 \mathbb{1}(\tilde{X}_\infty < 0)] \le C_6\sqrt{N}$. Note that to use the basic adjoint relation $\mathbb{E}[G_{\tilde{X}} V(\tilde{X}_\infty)] = 0$ for $V(x) = x^4 + a_1 x^3 + a_2 x^2$, one just need to verify $\mathbb{E}[\tilde{X}_\infty^4] < \infty$, and the proof is similar to that in Section E.4.1 and is omitted here.

Appendix F: Computational complexity of the exact analysis

In this section, we investigate the computational complexity for evaluating the mean waiting time $\mathbb{E}_\infty[W(t)]$. We compare between the exact analysis introduced in Section 2 and the normal approximation developed in Section 4.

To evaluate the mean waiting time $\mathbb{E}_\infty[W(t)]$, we use the following integral:

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

The key is to evaluate $\mathbb{P}_\infty(W(t) > x)$, which is equivalent to (18), i.e.,

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

Conditioning on the value of $X(0) = n$, we have

$$\mathbb{P}_\infty(W(t) > x) = \sum_{n=0}^\infty \mathbb{P}_\infty(D_{(0,t+x]} - A_{(0,t]} \le n - N | X(0) = n)\pi(n),$$

where $\pi(n) = \mathbb{P}_\infty(X(0) = n)$. Note that when the stationary distribution of the midnight count, π is already computed, the only difference between the exact analysis and the normal approximations lies in how to evaluate the conditional probability

$$r_\infty(t, x, n) = \mathbb{P}_\infty(D_{(0,t+x]} - A_{(0,t]} \le n - N | X(0) = n). \tag{135}$$

F.1. Exact analysis

To evaluate $\mathbb{P}_\infty(W(t) > x)$ with the exact analysis, we use the formulas specified in Proposition 8 in Section C.3 and consider $0 \le t < 1$.

F.2. Normal approximations

We only need to compute the cdf of a standard normal distribution $\Phi(\cdot)$ to approximate the conditional probability in (135). We approximate the unconditional probability $\mathbb{P}_\infty(W(t) > x)$ by (34) and (35) in the main paper, depending on the value of $t + x$.

F.3. Implementation and comparison between exact analysis and normal approximations

For the exact analysis, we see from (59), (61), and (64), that computing the pmf and cdf of the binomial and Poisson distributions could be time-consuming. *To reduce unnecessary repetitive computations*, we create lookup tables and pre-store all values for the needed pmf and cdf when $X(0) = n$, $t \in [0, 1)$ and $x \ge 0$ are given. As a result, when doing the multiplications in (59), (61), and (64), we can directly obtain the values of $J_{t+x}(\cdot), g(\cdot),$ and $f_t(\cdot)$ from the lookup tables, and it takes $\mathcal{O}(N)$, $\mathcal{O}(N^2)$ and $\mathcal{O}(N^3)$ arithmetic operations to evaluate (59), (61), and (64), respectively. Clearly, the computational time for the exact analysis increases when the value of $t + x$ increases, which we can observe from Table 4, Column 2. For the normal approximations, although the formula inside $\Phi(\cdot)$ depends on the value of $t + x$, the computation is almost instantaneous so that the difference is negligible; see Column 3 in Table 4.

We implement the algorithms for both the exact analysis and the normal approximations using a program written in C++ and test them on a laptop with a 1.8 GHz Intel Core i5 processor and 8 GB memory. Table 4 compares the computational times between the exact analysis and the normal approximations for $N = 500$. We compare the computational performance on both the conditional probability in (135) given $X(0) = N$ and the unconditional probability $\mathbb{P}_\infty(W(t) > x)$. Table 4 clearly shows the computational advantages of normal approximations: the exact analysis takes more than 20 minutes to compute $P(W(t) > x)$ when $t + x = 3.625$, whereas the normal approximation takes negligible time. Note that to compute $\mathbb{E}[W(t)]$, we need to integrate $P(W(t) > x)$ over a number of x. Although we have not done the experiments, we expect the exact analysis would take several hours to compute $\mathbb{E}[W(t)]$, given its performance on computing $P(W(t) > x)$.

The advantage of normal approximations becomes even more prominent when N is larger. Table 5 compares the computational performance between exact analysis and normal approximations for $N = 980$. We observe that the computational time for normal approximations remains negligible, but the computational time for the exact analysis increases dramatically. When $N = 980$, for $t + x = 3.625$, computing $P(W(t) > x)$ now takes almost 4 hours, which indicates that evaluating $\mathbb{E}[W(t)]$ becomes painful.

	$P(W(t)>x\|X(0)=500)$		$P(W(t)>x)$	
$t+x$	exact analysis (sec)	normal approx (sec)	exact analysis (sec)	normal approx (sec)
0.625	0	0.002	1.30	0.002
1.625	0.01	0.002	2.02	0.003
2.625	0.55	0.002	872.63	0.003
3.625	0.82	0.002	1350.30	0.003

Table 4 Computational performance for the exact analysis and normal approximations when the stationary distribution π is given and $N=500$. Here, $\Lambda=90.95$, and we set the mean LOS at 5.30 days. In all experiments, we fix $t=9/24$ (9am), and vary x from 0.25 (6 hours) to 3.25 (3 days and 6 hours). The stationary distribution π is truncated at a size of $n=2000$.

	$P(W(t)>x\|X(0)=980)$		$P(W(t)>x)$	
$t+x$	exact analysis (sec)	normal approx (sec)	exact analysis (sec)	normal approx (sec)
0.625	0.02	0.004	5.49	0.004
1.625	0.04	0.004	9.08	0.004
2.625	2.41	0.004	7905.27	0.004
3.625	4.66	0.004	13774.28	0.004

Table 5 Computational performance for the exact analysis and normal approximations when the stationary distribution π is given and $N=980$. Here, $\Lambda=181.91$, and we set the mean LOS at 5.30 days. In all experiments, we fix $t=9/24$ (9am), and vary x from 0.25 (6 hours) to 3.25 (3 days and 6 hours). The stationary distribution π is truncated at a size of $n=3000$.

Appendix G: Alternative non-Poisson arrival processes

In many real systems, the arrival processes may not be Poisson but show over-dispersion [38], which means that the variance is greater than the mean for arrivals in certain time intervals, violating the Poisson process assumption. In this section, we consider an alternative arrival process that could be over-dispersed and show how to adapt our developed algorithms to this non-Poisson arrival process.

For the alternative arrival process, we use the one introduced in [49] for non-ED admitted patients. That is, at the beginning of day k, we first generate a random number of patients, A_k to arrive within day k, and then generate the arrival time on day k for each of these A_k patients following a distribution with cdf $G(\cdot)$. Note that A_k does not necessarily follow a Poisson distribution, but that if it does, the arrival process then becomes a non-homogeneous Poisson process [31]. We assume the mean of A_k is still Λ, but the variance now equals $\mathrm{Var}(A_k)$.

Following the two-time-scale framework, we compute the stationary distribution for the midnight count first. If using the exact Markov chain analysis, we just need to change the distribution of A_k from Poisson to $G(\cdot)$ and then solve π accordingly. To use the approximate midnight distribution from Stein's method, we need to revise $\sigma^2(x)$ in (38). Recall that when deriving (53), the term Λ in the second equality there reflects the variance of the daily arrival under the Poisson arrival assumption. Thus, replacing Λ by $\mathrm{Var}(A_k)$ gives us a modified variance

$$\begin{aligned}\tilde{\sigma}^2(x) &= \delta^2\Lambda\left(\frac{\mathrm{Var}(A_k)}{\Lambda}+(1-\mu)\right)+b^2(x)-\delta(1-\mu)b(x)\\ &= \delta^2\Lambda\left(\mathrm{DS}_a+(1-\mu)\right)+b^2(x)-\delta(1-\mu)b(x),\end{aligned}$$

where $\mathrm{DS}_a = \mathrm{Var}(A_k)/\Lambda$ is the index of dispersion (see, for example, Page 72 of [13]) of the arrival process. Then, plugging the new $\tilde{\sigma}^2(x)$ to (40), we can obtain the new approximation formula to approximate midnight count distribution under the alternative arrival process.

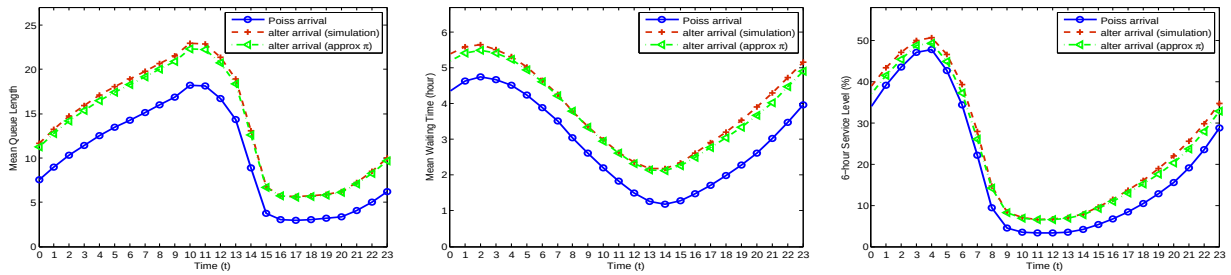
Next, we show how to adapt the normal approximations to predict the time-dependent performance. Similar to the adjustment on the approximate midnight distribution, we only need to change the $\Lambda G(t)$ item in the denominators (within the square root sign) of (27), (34), and (35) to

$$\Lambda G(t)\big(1-G(t)\big) + \mathrm{Var}(A_k)G(t)^2;$$

this is based on the fact that the number of arrivals from 0 to t, $A_{(0,t]}$ follows a binomial distribution with parameter $(m, G(t))$ when conditioning on the value of the daily arrival equal to m.

Figure 12 shows some numerical results for the alternative arrival process with $N = 500$. In the numerical experiments, we choose the distribution of A_k to have a similar shape as the empirical one for the ICU-admitted patients (see Figure 18(b) in [50]). We made some adjustments so that the daily arrival rate $\Lambda = 90.95$ matches what we used in Section 5. However, the index of dispersion DS_a is 1.48, consistent with the empirical observations. The arrival time distribution $G(\cdot)$ remains the same as the one used in Section 5.

In Figure 12, we plot the time-dependent performance curves under the alternative arrival process obtained from (i) simulation estimates, and (ii) normal approximations with π approximated by (41). We focus on the case with the approximate π because the Stein's method is more robust to the arrival process (it only requires estimates of the mean and variance of the daily arrival A_k). We can see that the approximation is still quite accurate: the performance curves predicted from using the approximate π are close to the simulation curves. For comparison, we also plot the performance curves under the Poisson arrival process with the same arrival rate. We can see from Figure 12 that because the variance of the arrival process is larger now, the system becomes more congested with higher queue length and longer waiting time compared to the Poisson arrival case. This increase in the system congestion is consistent with the prediction from Kingman's formula for single-server queueing systems [30].



(a) $\mathbb{E}_\infty(Q(t))$ (b) $\mathbb{E}_\infty(W(t))$ (c) $P_\infty(W(t) > 6/24)$

Figure 12 **Time-dependent performance curves when using the alternative arrival process ($N = 500$). Here, $\Lambda = 90.95$, mean LOS is 5.30 days, and we use the baseline discharge distribution. In each subfigure, the solid curve is from using the non-homogenous Poisson arrival process; the two dashed curves are from using the alternative arrival process with the red curves from simulation estimates and the green curves from normal approximations using π from (41).**