Queues with Time-Varying Arrivals and Inspections with Applications to Hospital Discharge Policies

Carri W. Chan
Decision, Risk, and Operations, Columbia Business School cwchan@columbia.edu

Jing Dong
Industrial Engineering and Management Sciences, Northwestern University jing.dong@northwestern.edu

Linda V. Green
Decision, Risk, and Operations, Columbia Business School lvg1@columbia.edu

In order for a patient to be discharged from a hospital unit, a physician must first perform a physical examination and review the pertinent medical information to determine that the patient is stable enough to be transferred to a lower level of care or be discharged home. Requiring an inspection of a patient's 'readiness for discharge' introduces an interesting dynamic where patients may occupy a bed longer than medically necessary. Motivated by this phenomenon, we introduce a queueing system with time-varying arrival rates in which servers who have completed service cannot be released until an inspection occurs. We examine how such a dynamic impacts common system measures such as stability, expected number of customers in the system, probability of waiting and expected waiting time. Leveraging insights from an infinite-server model, we're able to optimize the timing of inspections and find via theoretical and numerical analysis that 1) optimizing a single inspection time could lead to significant improvements in system performance when the amplitude of the arrival rate function is large, 2) multiple inspections should be uniformly distributed throughout the day, and 3) the marginal improvements of adding additional inspection times is decreasing.

Key words: Queueing, Time-varying arrival rates, Infinite server queues, healthcare

1. Introduction

Standard queueing models assume that a server is free to serve a new job as soon as the current job in process completes service. However, such an assumption may not be reasonable in some systems. Motivated by a hospital setting, we introduce a new queueing system which captures hospital discharge dynamics that, to the best of our knowledge, have not been considered in the past. Our analysis of this model provides some insights into how the frequency and timing of physician rounds, which determine the timing of patient discharge orders, impact patient flow dynamics.

We consider a standard practice in hospital care: a physician must check on a patient to assess whether she is ready to be transferred to a lower level of care or sent home before a discharge order can be placed. That is, while a patient (customer) may be medically ready to be discharged, she will continue to occupy a bed (server) until a physician checks on her and writes a transfer or discharge order. Therefore, a patient's length-of-stay (LOS) is likely to be longer than is medically necessary due to the need for a physician's

'inspection'. Most hospitals have rounds once a day, typically in the morning, at which point patients are examined to determine if they are ready for discharge. Such rounds occur in the ICU when making decisions to transfer patients to step-down units or the general wards. They also occur in the wards when deciding to discharge patients home.

While queueing models have been used extensively to examine patient flows in hospitals (e.g. McManus et al. (2004), Yankovic and Green (2011), Saghafian et al. (2012) among others), to the best of our knowledge, none have considered the need for physicians to inspect patients to determine whether they are 'ready for discharge'. In other service systems, such as call-centers, such a phenomenon does not exist as servers typically correspond to humans who will know immediately when a customer has completed service. In contrast, in the healthcare setting, a physician must review lab tests, imaging results, vital signs, etc. before a patient can be moved from a hospital unit. Typically, nurses execute the care plan dictated by the patient's physician; however, even if a nurse suspects a patient may be ready for discharge, this cannot occur without physician approval. Physicians have many demands on their time (e.g. surgeries, office visits, follow up with outpatients, charting, etc.) and are not generally available during the day. So physician 'inspections' most often occur on a one-time basis–early in the morning when rounds take place, which in teaching hospitals, also provide an educational opportunity for medical residents and students. In this paper, we examine the impact of the timing as well as what gains could be achieved by increasing the frequency of 'inspections'.

Motivated by the above discussion, we develop a queueing system where a server who completes service can only be freed at pre-specified inspection times. If service has not completed prior to the inspection time, the customer continues to occupy the server until at least the next inspection time. In the hospital setting, which serves as our primary motivation, a customer corresponds to a patient; each server corresponds to a bed; and the service time of a customer corresponds to the time of physical admission to a bed until the time the patient is medically ready to be discharged or transferred to another unit. Note that in our system, the service time may be *shorter* than the total time a patient spends in the bed. We assume that the periodic arrival rate varies depending on the time of the day. We refer to this system as an $M_t/M(\mathbf{T})/s$ model. M_t refers to the time varying Poisson arrival process, which is not only a fairly standard queueing assumption, but also has been shown to be a reasonable model for many healthcare applications (see for example Kim and Whitt (2014) and Green et al. (2006)). $M(\mathbf{T})$ denotes the fact that service times are exponentially distributed but the servers are only released at inspection times defined by the vector $\mathbf{T}$. In this work, we examine how the required inspection alters the queueing dynamics as well as assess when such inspections should occur to optimize system performance. Our main contributions can be summarized as follows:

- **Stability:** We analyze the stability of the $M_t/M(\mathbf{T})/s$ system and introduce a notion of an 'effective service time', which reflects the inflation of time spent with a server due to the need for inspection. We show that the stability condition is invariant to the timing of the inspection.

- **Equilibrium analysis:** We analyze the periodic equilibrium behavior of the $M_t/M(\mathbf{T})/s$ system and provide explicit characterizations of the queue length and the waiting time processes.
- **Optimizing performance:** We leverage insight from an infinite-server model to optimize the timing of inspection times. In particular, we identify inspection times to minimize the expected mean and maximum number of customers in the system as well as the probability of waiting. The gains which can be achieved by optimizing the timing of inspections depend on the amplitude of the arrival rate function. For example, when assuming a sinusoidal arrival rate function and a single inspection time per cycle, the reduction in the time-average expected occupancy level can range from 0 (when the arrival rate is constant over time) up to 10% (when the amplitude is maximal).
- **Impact of additional inspection times:** We find that system performance improves (e.g. stability region increases, congestion decreases) when adding more inspection times; however, the marginal improvements are decreasing in the number of inspections. We also find that when there are multiple equally spaced inspection times, the system performance is very robust with respect to the timing of inspections.
- **Model Extensions:** We check the robustness of our analysis under different modeling assumptions. More specifically, we examine the reliability of the infinite server approximation, when using a more general class of service time distributions, including the widely used log-normal distribution. In addition, we consider scenarios where inspection times do not have to be equally spaced. Next, we examine our results in the case of more general arrival rate functions; specifically, we consider the case of the empirical arrival rate to Emergency Departments in the United States. Lastly, we analyze the dynamics of an $M_t/M(\mathbf{T})/s/s$ system where customers are blocked (rerouted) when the system is full. In most cases, the infinite server system approximation still enables us to select the optimal inspection schedule to minimize the occupancy level and/or probability of blocking in the loss system and all of our prior insights translate to these extended models. However, in some instances (e.g. deterministic service times) the system dynamics can differ substantially.

The rest of the paper is structured as follows. We conclude this section with a brief review of related literature. We formally introduce the $M_t/M(\mathbf{T})/s$ queueing model in Section 2. In Section 3, we derive the stability condition for our new system. Section 4 derives expressions for various time-dependent performance measures. In Section 5, we introduce an infinite-server model, which we leverage to provide insights into optimizing the inspection times in the finite server situation. We are also able to relax the assumption of exponential service times in this analysis. In Section 6 & 7, we consider several model extensions. Finally, we summarize our conclusions and provide discussion for further research in Section 8. All proofs of the results are given in the appendix.

1.1. Related Literature

In this paper, we introduce a queueing system with time-varying arrival rates which is prevalent in hospitals as well as many other service systems (e.g. Armony et al. (2015), Shi et al. (2016)). While the assumption

of time-varying arrivals is better able to capture reality, such models tend to be much harder to analyze than their time-homogenous counterparts. A naive approach to approximate system performance is to eliminate the time-variation and use the steady-state distribution of a time homogeneous queue whose arrival rate is the long-run average arrival rate. However, such an approach is unable to capture the variability in performance over time, and because the variation can be quite significant in some cases, insights extracted from such an approach may be very misleading (Green et al. 1991, Jennings et al. 1996).

When the time scale of variation in the arrival rate is much larger than the service time, the pointwise stationary approximation, which uses the steady-state distribution of a time homogeneous queue with arrival rate $\lambda(t)$ to approximate the performance of the time varying queue at time t, works quite well (Green and Kolesar 1991). To cope with the case when the time scale of variation is of approximately the same order or much smaller than the service times, Jennings et al. (1996) propose an infinite-server approximation, which seems to be reasonably accurate. In this work, we use a similar infinite-server approximation in our analysis, albeit for a very different queueing model, and show numerically that it is able to capture some of the main features of our system. In our analysis, we leverage results for $M_t/M/\infty$ queues from Eick et al. (1993a) and Eick et al. (1993b).

There has been quite a bit of work on staffing and capacity management policies under time-varying arrival rates (see, for example, Jennings et al. (1996), Green et al. (2007) and Liu and Whitt (2012)). In contrast to these papers, we consider the case where the number of servers, which in our model corresponds to the number of hospital beds, is fixed. As such, our focus is on characterizing the impact of alternative inspection (discharge) policies and then optimizing over them.

The notion of inspection policies is relatively new in the queueing literature. Zazanis (2004) analyzes infinite server queues with synchronized departures driven by a single point process. Dobson et al. (2010) uses performance analysis of a c dimensional discrete time Markov Chain to examine the impact of 'early discharges' from the intensive care unit (ICU). They assume the patient's length of stay is discrete (in days) and bounded. In addition, patients can be discharged earlier if there are not enough servers (beds) in the ICU. Chan et al. (2012) develops an optimization framework to select patients for such early discharges. In our setting, we assume that patients can only be discharged once they are medically ready.

Inspection in queues and other stochastic models has been analyzed in the manufacturing setting (see for example Ohnishi et al. (1986), Jewkes (1995) and Yao and Zheng (1996)). The primary focus in these papers is quality control and/or detection of malfunctioning machines. As such, the system dynamics are quite different from our setting. For example, in most manufacturing settings, a faulty product will be detected during inspection which may lead to additional processing procedures; however, it is unlikely the product will continue to occupy the servers (machines) after it finishes service while it awaits inspection. In the context of scheduling inspections in order to detect malfunctioning machines, it is the status of the machine (broken or not) which determines whether a new job can begin, rather than the status of the job and

knowledge of this status which determines when the server is released. There is also an extensive literature on 'vacation models' where servers can temporarily leave the system (e.g. Federgruen and Green (1986), Doshi (1986)). In this case, servers can go on vacation (rather than require maintenance) under various conditions, such as when the system is empty. The majority of these papers focus on deriving steady-state performance measures given different vacation dynamics for a single server.

There has also been some work on queues with batch departures (e.g. Foster and Nyunt (1961), Foster and Perera (1964)). Here, the goal is to examine steady state analysis when jobs are served in batches. In such a setting, the effective time in system increases for jobs as they must wait until a full batch can be served. However, in contrast to our model, the jobs depart asynchronously because, although they begin service in batches, their service times are not identical.

Most related to our work is Powell et al. (2012), Shi et al. (2016), and Dai and Shi (2014), which all consider modifying discharge timing in a hospital setting. Powell et al. (2012) and Shi et al. (2016) apply a data-driven approach to build a queuing model and use simulation to test the performance of different discharge scenarios, while Dai and Shi (2014) use asymptotic analysis to examine the same. All of these papers focus on the impact on patient "boarding" (waiting for an available bed) in the Emergency Department (ED) when patient discharges are shifted earlier in the day. The argument is that moving patients out of beds earlier will reduce the waiting time of new patients. Dai and Shi (2014) assumes that medical needs dictate patient LOS on a daily level, while operational factors dictate admission and discharge times which operate on an hourly level. As such, an implicit assumption in these papers is that a patient is ready for discharge at any time during the day of discharge, and thus, they conclude that to minimize delays, discharging everyone at midnight would be optimal. However, because such a discharge policy is not feasible, they consider alternative policies. In contrast, our model takes the perspective that the recovery of patients is a continuous process. For example, a patient who is not yet ready to be discharged in the morning may become well enough to be discharged later in the afternoon. This implies that earlier discharge times do not necessarily lead to better performance. While we also consider discharge timing, we aim to optimize this depending on when a patient is 'ready for discharge'. This notion of inspection is not considered in the prior works.

2. The $M_t/M(\mathbf{T})/s$ Model

In order to understand the impact of requiring a physician's review before a patient can be discharged, we introduce a queueing model where customers (patients) complete service but continue to occupy the servers (beds) until an inspection occurs. Our model is depicted in Figure 1.

Jobs arrive to the system according to a time-varying Poisson process with rate given by $\lambda(t)$ (measured per hour). We define $\lambda(t)$ for $t \in \mathbb{R}$. We assume $\lambda(\cdot)$ varies within a day but follows the same intraday pattern across days (periodic pattern). Specifically, $\lambda(t+24) = \lambda(t)$. We let

$$\Lambda(t) = \int_0^t \lambda(u)du$$

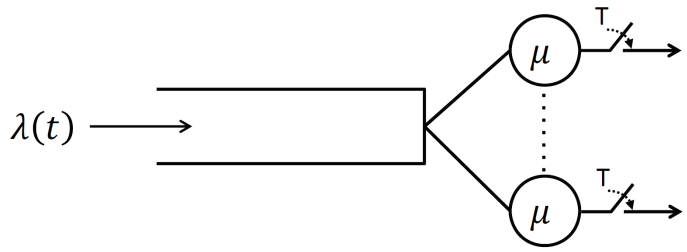


Figure 1 $M_t/M(\mathbf{T})/s$ **queueing model with 'readiness to discharge' inspections: customers can only be removed from the system after an inspection which occurs at time(s)** $\mathbf{T}$**.**

denote the cumulative arrival rate function. Specifically, $\Lambda(24)$ denotes the *daily* arrival rate. Without loss of generality, we also assume time zero is the beginning of a day. If the period and time-scale of variation do not occur according to a 24-hour period, a simple re-scaling of time, t, can incorporate such dynamics. The periodic assumption allows us to measure the long-run average performance via the notion of a periodic equilibrium, as we detail in Section 4.

There are s identical servers in the system. If a customer arrives and a server is available, it will begin service immediately. If there are no servers available, the customer must wait in the (infinite buffer) queue until a server becomes available. In order to focus on the impact of inspections, we consider a homogenous customer population where each customer's service time is exponentially distributed with rate μ (measured per hour). The reciprocal of μ can be thought of as the mean service time for the average patient. Upon completion of service, the customer will *still occupy the server*. The server is only released once an inspection occurs and the customer has completed service *prior* to the inspection time. We let $\mathbf{T} \in [0,24]^N$, with $T_1 < T_2 < \cdots < T_N$, denote the daily $N \geq 1$ inspection times. Thus, the first inspection occurs each day at T_1, the second at T_2, etc.

Suppose a customer arrives at time 0 and finds an available server. This customer immediately begins processing. It will complete service at time t_0, where t_0 is exponentially distributed with mean $1/\mu$. The server this customer occupies will be freed at time k on day d, where k and d are defined as:

$$k = \begin{cases} \inf_{i\in\{1,2,\ldots,N\}}\{T_i : T_i \geq \mod(t_0,24)\}, & \text{if} \quad \mod(t_0,24) \leq T_N; \\ T_1, & \text{if} \quad \mod(t_0,24) > T_N. \end{cases}$$

$$d = \begin{cases} \lfloor t_0/24 \rfloor, & \text{if} \quad \mod(t_0,24) \leq T_N; \\ \lfloor t_0/24 \rfloor + 1, & \text{if} \quad \mod(t_0,24) > T_N. \end{cases}$$

Specifically, a customer will leave the server at the first inspection time following completion of service. If a customer finishes service after the last inspection time on a specific day, it needs to wait until the first inspection time on the next day. No new customers may begin service with this server until it is freed. We refer to such a system as an $M_t/M(\mathbf{T})/s$ system. Our goal is to understand the role $\mathbf{T}$ plays in system dynamics.

As a benchmark for comparison, an $M_t/M/s$ system is a system in which inspections occur continuously, so that a customer is released as soon as it completes service. We note that there exists an $M_t/M(\mathbf{T})/s$ system with $N = \infty$ that is equivalent to an $M_t/M/s$ system (e.g, when inspections are evenly spaced throughout the day, $T_i = 24/N$).

For tractability, we assume that inspections take place instantaneously and that it is possible to inspect all jobs simultaneously. In relating our model back to practice, a physician typically spends a few minutes with each patient during rounds. While this is neither simultaneous nor instantaneous, the scale of the duration is small enough that we believe the time required to complete rounds is not a first order effect compared to the impact of the inspection timing. Additionally, we assume that completed jobs leave the servers immediately following inspection. In practice, after a physician approves a patient's discharge, the patient will continue to occupy the bed until the entire discharge or transfer process - which includes paperwork, the conveyance of discharge instructions, arrangement for transport etc., is completed, and the bed will not be available for another patient until it is made ready by housekeeping.

3. Stability conditions

We start by considering the impact of the inspection times on the stability of the $M_t/M(\mathbf{T})/s$ system. We find that the need for inspection prior to releasing a server alters the stability condition from that of a more standard $M_t/M/s$ system; moreover, the number of inspection times plays a substantial role.

Recall that with continuous inspection, customers are released immediately upon completion, so that we have an $M_t/M/s$ system. It is well-known that the stability condition for this system is simply:

$$\Lambda(24) < s24\mu.$$

where $\Lambda(24)$ is the daily arrival rate (Heyman and Whitt 1984). In contrast, the stability condition for the $M_t/M(\mathbf{T})/s$ system with a finite number of inspection points ($N < \infty$) is more nuanced. While stability is a relatively coarse measure of system performance, it is a starting point to understand the impact of the inspection times. In what follows we establish the stability condition based on a discrete time Markovian system descriptor.

3.1. One inspection per day: $N = 1$

We start by considering the most extreme case of the $M_t/M(\mathbf{T})/s$ system. Here we assume there is only a single inspection each day, so that $N = 1$ and T is simply a scalar time of daily inspection.

We denote by Q_n the number of customers in the system immediately following the inspection on day n, which occurs at $T + 24 \times (n-1)$. It is straightforward to see that the dynamics of Q_n can be described by the following recursion:

$$Q_{n+1} = Q_n + A_{n+1} - D_{n+1},$$

where $A_{n+1} \sim Poisson(\Lambda(24))$ is the number of new arrivals between the inspection time on day n (excluded) and the inspection time on day $n+1$ (included) and D_{n+1} is the number of discharges at the inspection time on day $(n+1)$. That is, D_{n+1} is the number of customers who have completed service between time $T+(n-1)\times 24$ and time $T+n\times 24$. Under the exponential service time assumption, $\{Q_n, n \geq 1\}$ is a discrete time Markov chain.

The service time, in hours, of each customer in service is exponentially distributed with rate μ so that the rate in days is 24μ. Since there are s servers, an upper bound on the number of departures in a day is given by the number of departures if there were s customers in service to start with. Specifically, if we let $\bar{D}_{n+1} \sim Binomial(s, 1-e^{-24\mu})$, then $D_{n+1} \leq_{st} \bar{D}_{n+1}$, where $\leq_{st}$ denotes stochastic dominance.

We have the following lemma about the stochastic stability conditions when having only one inspection per day:

Lemma 1 *Under the one inspection time per day policy, the system is stable if*

$$s(1-e^{-24\mu}) > \Lambda(24).$$

The system is unstable if

$$s(1-e^{-24\mu}) < \Lambda(24).$$

Intuitively, we can view $(1-e^{-24\mu})$ as the *effective service rate* per day of each server. We then notice that $(1-e^{-24\mu}) < 24\mu$, i.e. the effective service rate under the one inspection per day policy is smaller than the effective service rate per day with continuous inspection. Thus, the inspection time artificially inflates the 'service requirement' of each customer. Interestingly, the stability condition does not depend on the actual inspection time T, nor does it depends on the nature of the time-variability. It does, however, depend on the periodicity assumption of the arrival rate. While this is true when considering stability, when we consider other performance measures, such as the number of customers in the system, we will see that the inspection time can have a substantial impact. Moreover, optimizing the timing of the inspections is only relevant in a time-varying environment; when arrival rates are time-homogenous, the system performance is invariant to the precise timing of inspections.

3.2. Multiple inspections per day

We now consider the case where there are multiple inspections per day.

To demonstrate the basic idea, we start by discussing the case of $N=2$ as an example. We denote the inspection times as $0 \leq T_1 < T_2 < 24$, which are the times by which the servers that finished service can be freed. The number of customers in the system right after the second inspection time on day n, Q_n, has the following dynamic:

$$Q_{n+1} = Q_n + A^{(1)}_{n+1} - D^{(1)}_{n+1} + A^{(2)}_{n+1} - D^{(2)}_{n+1}$$

where $A_{n+1}^{(1)} \sim Poisson(\int_{T_2-24}^{T_1} \lambda(t)dt)$, is the number of new arrivals between the second inspection time on day n (excluded) and the first inspection time on day $n+1$ (included). $D_{n+1}^{(1)}$ is the number of discharges at the first inspection time on day $(n+1)$. Likewise, $A_{n+1}^{(2)} \sim Poisson(\int_{T_1}^{T_2} \lambda(t)dt)$, is the number of new arrivals between the first inspection time on day $(n+1)$ (excluded) and the second inspection time on the same day (included). $D_{n+1}^{(2)}$ is the number of discharges at the second inspection time on day $(n+1)$.

Let $\Delta_2^{(1)} = 24 + T_1 - T_2$ and $\Delta_2^{(2)} = T_2 - T_1$. Similar to the single inspection case, we let $\bar{D}_{n+1}^{(i)} \sim Binomial(s, 1 - e^{-\Delta_2^{(i)}\mu})$ for $i = 1, 2$. Then $D_{n+1}^{(i)} \leq_{st} \bar{D}_{n+1}^{(i)}$ for $i = 1, 2$. Intuitively, the system is stable if the arrival rate is less than the maximum achievable service rate, i.e. $s(1 - e^{-\Delta_2^{(1)}\mu}) + s(1 - e^{-\Delta_2^{(2)}\mu}) > \Lambda(24)$, and the system is unstable if the arrival rate is more than the maximum achievable service rate, i.e. $s(1 - e^{-\Delta_2^{(1)}\mu}) + s(1 - e^{-\Delta_2^{(2)}\mu}) < \Lambda(24)$.

We can generalize this reasoning to derive the stability condition for general N.

Theorem 1 *Consider the policy with inspections that occur at times $T_1 < T_2 < \cdots < T_N$ every day. Let $\Delta_N^{(1)} = 24 + T_1 - T_N$, $\Delta_N^{(i)} = T_i - T_{i-1}$ for $i = 2, \cdots, N$. The system is stable if*

$$\sum_{i=1}^{N} s(1 - e^{-\Delta_N^{(i)}\mu}) > \Lambda(24).$$

The system is unstable if

$$\sum_{i=1}^{N} s(1 - e^{-\Delta_N^{(i)}\mu}) < \Lambda(24).$$

As in the single discharge per day case, the established stability condition does not depend on the precise timing of the inspections. Rather its dependence on the inspection times only appears via the duration of time between them. We now consider what separation of inspection times would result in the largest stability region. A direct result from Theorem 1 is that the maximum effective service rate is achieved with evenly spaced inspection times.

Lemma 2 *$\sum_{i=1}^{N}(1 - e^{-\Delta_N^{(i)}\mu})$ attains its maximum when $\Delta_N^{(i)} = 24/N$ for $i = 1, 2, \cdots, N$.*

Intuitively, this is because the effective service rate is dictated by the duration of time between inspections. Specifically, we are concerned with the number of jobs which complete service between two consecutive inspection times. Because the effective service rate is concave in $\Delta_N^{(i)}$'s, one can use a *Lagrangian argument* to demonstrate the duration between inspection times should be equal throughout the day (i.e. consider $L(\Delta_N^{(i)}, \nu) = \sum_{i=1}^{N}(1 - e^{-\Delta_N^{(i)}\mu}) - \nu(24 - \sum_{i=1}^{N} \Delta_N^{(i)})$).

When we have continuous inspections, the effective service rate per day is 24μ. We see that requiring an inspection before a server is freed reduces the effective service rate in our system. In particular, let $\Upsilon(N) :=$

$N(1-e^{-\mu 24/N})$ denote the effective service rate per day per server with the N evenly spaced inspections per day. One can easily verify that $\Upsilon'(x)>0$ and $\Upsilon''(x)<0$ for $x\geq 0$, so that $\Upsilon(N)$ is increasing in N and

$$\Upsilon(N+1)-\Upsilon(N)<\Upsilon(N)-\Upsilon(N-1).$$

This suggests that more inspections per day will improve the efficiency of the system (the effective service rate increases as N increases), but the improvement is diminishing. We also notice that $\lim_{N\to\infty}\Upsilon(N)=24\mu$.

We let $\Lambda_{\max}(N)=sN(1-e^{-24\mu/N})$ be the constraint on the daily arrival rate for stability for evenly spaced N inspections per day. We have that adding one more inspection increases $\Lambda_{\max}(N)$ by:

$$s\left[(N+1)(1-e^{-24\mu/(N+1)})-N(1-e^{-24\mu/N})\right],$$

which is linearly increasing in s for fixed N. Similarly, we can deduce that adding one more server increases $\Lambda_{\max}$ by:

$$N\left(1-e^{-24\mu/N}\right),$$

which is a constant for fixed N. Thus, there exists a threshold $\bar{s}(N)$ such that for $s>\bar{s}(N)$ adding an additional inspection will increase the effective capacity of the system *more* than adding a server. For $s<\bar{s}(N)$, adding more servers is more beneficial.

Since most systems currently have $N=1$ inspections, we consider what $\bar{s}(1)$ would be, which is a direct corollary to Theorem 1.

Corollary 1 *For*

$$\bar{s}(1)=\frac{1-e^{-24\mu}}{2(1-e^{-12\mu})-(1-e^{-24\mu})},$$

if $s>\bar{s}(1)$, then adding one more inspection 12 hours after the first inspection, for a total of 2 inspections, will increase the stability region more than adding an additional server. If $s<\bar{s}(1)$, then adding an additional server is more effective.

For $\mu=1/75$ and $1/130$ (which are typical service rates in the ICU and internal wards respectively), we have that $\bar{s}=12.5267$ and 21.6820, respectively. This suggests that for small systems, adding more beds would be more effective, while for large systems, adding more inspection rounds would be more effective.

Interestingly, we see that the time-variation of the arrival rate does not have any impact on the stability condition. This is because stability is a fairly coarse measure of system performance.

Note that our stability results rely on the assumption of exponential service times. This is imposed to precisely quantify the distribution of the number of departures per day. In Section 6, we will relax this assumption and consider general service time distributions. We find that when the coefficient of variation of the service time distribution is not too different that of the exponential distribution, the above insights generally carry over. However, when the service time distribution is very different, e.g. deterministic, we will see that the system dynamics can be quite different.

4. Time dependent performance measures

In our stability analysis, we did not consider the precise nature of time-variation of the arrival process. However, for other performance measures, the arrival rate has a substantial impact. We now turn our attention to measures–the number of customers in system, the probability of waiting, and the expected waiting time–which have time dependent dynamics induced by both the time-varying arrival rates and the inspection schedule defined by $\mathbf{T}$. We are interested in these measures as minimizing them can result in increased access to care for patients as well as reduced operational costs (potentially by requiring fewer beds to meet demand). In what follows, we will assume there are enough servers for the system to be stable so that these performance measures are well defined and finite.

For simplicity of exposition, in this section, we focus on the single inspection per day case. Similar results hold for the case of multiple inspections per day. Let τ_n denote the inspection time on day n. Then $\tau_1 = T$ and $\tau_{n+1} = \tau_n + 24$. We divide the analysis into two time scales (discrete v.s. continuous).

Following the stability analysis in Section 3, we denote Q_n as the number of customers in the system right after the inspection on day n, A_{n+1} as the number of new arrivals between τ_n (excluded) and τ_{n+1} (included), and D_{n+1} as the number of discharges at τ_{n+1}. Q_n is a discrete time Markov chain on the state space $\mathbb{Z}^+$. Under the stability condition in Theorem 1, it has a unique stationary distribution, which we denote as $\pi(\cdot)$.

We next define some continuous time system descriptors. Let $Q = \{Q(t) : t \geq 0\}$ denote the number of customers in the system at time t. We assume the sample paths of Q are right continuous with left limit. Thus, an arrival at time t will be counted in $Q(t)$, and discharges at time t will not be counted in $Q(t)$. Let $W = \{W(t) : t \geq 0\}$ denote the waiting time process, i.e. $W(t)$ is the time a customer would have to wait if he arrives to the system at time t. For simplicity of notation, we also write $A(s,t)$ as the cumulative number of arrivals on the interval $(s,t]$. Then we have

$$Q_n = Q(\tau_n) \text{ and } A_n = A(\tau_k, \tau_{k+1}).$$

As no discharges occur between inspection times, for $\tau_n < t < \tau_{n+1}$,

$$Q(t) = Q_n + A(\tau_n, t)$$

We also have the following simple relation between $Q(t)$ and $W(t)$. For $\tau_n \leq t < \tau_{n+1}$,

1. If $Q(t) \geq s$, then $W(t) \geq \tau_{n+1} - t$. That is, if the system is full at t, a customer, that arrives at time t, needs to wait at least until the next inspection time before beginning service.

2. If $(Q(t) - s)^+ \geq \sum_{i=n+1}^{n+k} D_i$, then $W(t) \geq 24k + (\tau_{n+1} - t)$ for $k = 1, 2, \ldots$. That is, if there are more than $\sum_{i=n+1}^{n+k} D_i$ customers waiting at time t, a new customer that arrives at time t needs to wait at least until the $(n+k+1)$-th discharge time before beginning service.

As Q_n is Harris recurrent under the stability condition, $Q(t)$ and $W(t)$ are regenerative processes (e.g. we can pick $\{\tau_n : Q(\tau_n) = j\}$, for any $j \in \mathbb{Z}^+$, as the regeneration points). We next consider the notion of periodic equilibrium as in Heyman and Whitt (1984), and then characterize the periodic equilibrium for $Q(t)$ and $W(t)$ based on the stationary distribution of Q_n (Theorem 2).

Definition 1 *We say $\{X(t) : t \geq 0\}$ is in periodic equilibrium with period κ, if $X(t)$ has the same distribution as $X(t+\kappa)$ for $t \geq 0$.*

Theorem 2 *If $Q_n \sim \pi$, then $\{Q(\tau_n + t) : t \geq 0\}$ and $\{W(\tau_n + t) : t \geq 0\}$ are in periodic equilibrium with period equal to 24. For $0 < t < 24$, the conditional distributions of $Q(\tau_n + t)$ and $W(\tau_n + t)$, given that $Q(\tau_n)$ is distributed according to π, are*

$$\begin{aligned}
P(Q(\tau_n + t) = m | Q_n \sim \pi) &= \sum_{l=0}^{m} \pi(l) P(l + A(\tau_n, \tau_n + t) = m), \\
P(W(\tau_n + t) > 0 | Q_n \sim \pi) &= \sum_{l=0}^{\infty} \pi(l) P(l + A(\tau_n, \tau_n + t) \geq s), \\
P(W(\tau_n + t) > 24k | Q_n \sim \pi) &= \sum_{l=0}^{\infty} \pi(l) P\left(l + A(\tau_n, \tau_n + t) - \sum_{i=n+1}^{n+k} D_i \geq s \right).
\end{aligned}$$

Assuming now the system starts operating from the infinite past from an empty state, we denote $(\Omega, \mathcal{F}, P_\infty)$ as the probability space where $\{Q(t) : t \in \mathbb{R}\}$ and $\{W(t) : t \in \mathbb{R}\}$ are defined (see Thorisson (1985) for theoretical support for initializing non-stationary models at $t = -\infty$). Without loss of generality, we assume $\tau_0 := 0$. We also denote τ_n, $n \geq 0$, as the n-th inspection time, counting forwards in time, after time 0 ; and τ_{-n}, $n \geq 0$, as the n-th inspection time, counting backwards in time, before time 0. Then $\{Q_n = Q(\tau_n) : n \geq 0\}$ would be in steady state. We also have that $\{Q(t) : t \geq 0\}$ and $\{W(t) : t \geq 0\}$ would be in periodic equilibrium. We also denote E_∞ as the expectation with respect to P_∞, which denotes the equilibrium distribution. Using time-varying Little's law (Bertsimas and Mourtzinou 1997), we have the following Proposition.

Proposition 1 *For $0 \leq t < 24$,*

$$\begin{aligned}
E_\infty[(Q(T+t) - s)^+] &= \int_T^{T+t} P_\infty(W(u) > 0) \lambda(u) du + \sum_{k=1}^{\infty} \int_T^{T+24} P_\infty(W(u) > 24k) \lambda(u) du, \\
E_\infty[W(T+t)] &= P_\infty(W(T+t) > 0)(24 - t) + 24 \sum_{k=1}^{\infty} P_\infty(W(T+t) > 24k).
\end{aligned}$$

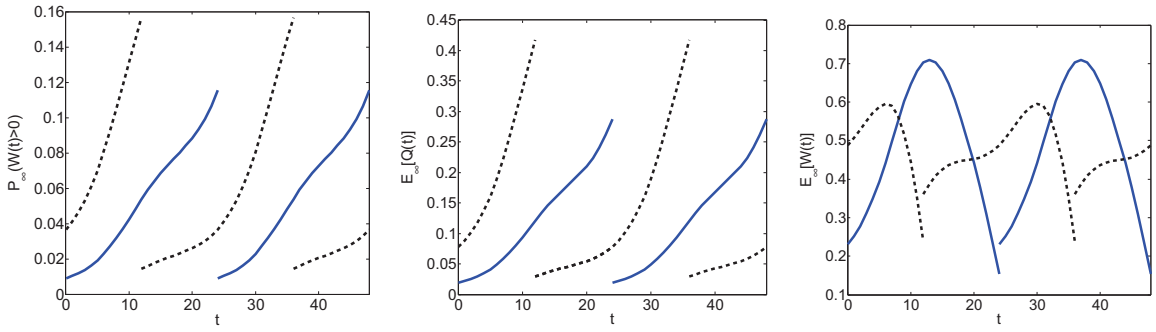
Remark 1 *In the hospital setting, it is unusual that a patient waits more than a day for a bed. When the probability of such a delay is very small, $P_\infty(W(T+t) > 24) \approx 0$, we have*

$$\begin{aligned}
E_\infty[(Q(T+t) - s)^+] &\approx \int_T^{T+t} P_\infty(W(u) > 0) \lambda(u) du, \\
E_\infty[W(T+t)] &\approx P_\infty(W(T+t) > 0)(24 - t).
\end{aligned}$$

As t increases, more customers arrive and no customers can be discharged until the next discharge time. Thus, both $P_\infty(W(T+t)>0)$ and $E_\infty[(Q(T+t)-s)^+]$ are increasing in t for $0 \le t < 24$. This monotonicity in t does not hold in general for $E_\infty[W(T+t)]$. Due to the discharge dynamics, on the one hand, an arriving patient may end up waiting less than a patient who has already arrived if both patients are admitted at the same discharge point; on the other hand, a new patient may wait longer than a previous patient if he cannot be admitted at the following discharge point, but the earlier patient can.

Figure 2 plots $P_\infty(W(t)>0)$, $E_\infty[Q(t)]$ and $E_\infty[W(t)]$ as a function of t for different inspection times T. Note these inspection times were chosen arbitrarily for illustrative purposes and we use a sine function as the arrival rate function. We observe that $P_\infty(W(t)>0)$ and $E_\infty[Q(t)]$ are increasing in t for $t \in [T, T+24)$. That is, both the probability of waiting and the expected number in system are non-decreasing following the inspection time, T, until the next inspection time 24 hours later at $T+24$. On the other hand, $E_\infty[W(t)]$ is not monotonic in t. As we will see in Section 5, this creates challenges in estimating how the inspection time(s) will impact the expected waiting time. We also observe the periodicity of the performance measures and the discontinuity at inspection times.

Figure 2 $M_t/M(\mathbf{T})/s$ **queueing system: Time dependent performance measures (**$\lambda(t) = 0.25 + 0.125\sin(\pi/12t)$**,** $\mu = 1/75$**,** $s = 31$**)**
$-: T = 0$**,** $\cdots: T = 12$

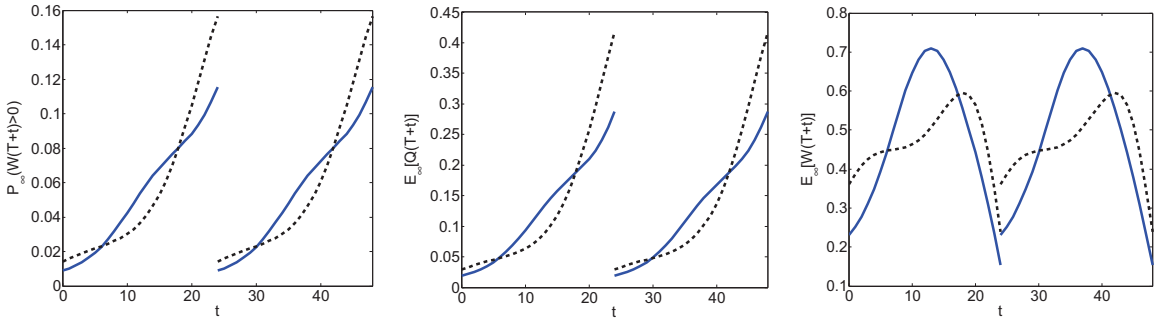


(a) $P_\infty(W(t)>0)$ (b) $E_\infty(Q(t)-s)^+$ (c) $E_\infty W(t)$

In comparing the time-varying performance for different inspection times, it is not clear which inspection time results in better performance. To better facilitate comparison, Figure 3 plots the time shifted version of the performance measures, so we can compare performance measures at identical times following the inspection times. Specifically, we plot $P_\infty(W(T+t)>0)$, $E_\infty[Q(T+t)]$ and $E_\infty[W(T+t)]$ for the same systems in Figure 2. While it is clear that different discharge times ($T=0$ v.s. $T=12$) result in different performance outcomes, it is not obvious which inspection time is better. Specifically, there is no clear dominance between the two curves. For example, although the probability of waiting right before and after inspection is smaller when $T=0$ compared to when $T=12$, 10 hours after inspection $P(W(T+10)>0)$ is

larger when $T = 0$ compared to when $T = 12$. This motivates us to look at alternative average performance measures in considering how to optimize over inspection times in Section 5.

Figure 3 $M_t/M(\mathbf{T})/s$ **queueing system: Time dependent performance measures with shifted starting point** ($\lambda(t) = 0.25 + 0.125\sin(\pi/12t)$, $\mu = 1/75$, $s = 31$)
$-: T = 0, \cdots: T = 12$



(a) $P_\infty(W(T+t) > 0)$ (b) $E_\infty Q(T+t) - s)^+$ (c) $E_\infty W(T+t)$

5. When should inspections occur?

In this section, we analyze how inspection times should be scheduled to provide high service quality.

As before, we will assume that there are enough servers given the inspection schedule such that the system is stable. Based on the stochastic stability analysis, we will focus on policies that have equally spaced inspection times. For simplicity of notation, we write $\Delta_N = 24/N$, which is the time interval between successive inspection times when we have N evenly spaced inspections per day.

We define an arrival rate weighted time average as

$$\bar{V}(t) := \Lambda(t)^{-1} \int_0^t X(u)\lambda(u)du$$

where $X(u)$ is some nonnegative function of $Q(u)$ or $W(u)$. As we examine various arrival rate weighted time average performance measures, the following results will be useful.

Proposition 2 *If* $\sup_t E[X(t)] < \infty$, $\bar{V}(t) \to \bar{V}(\infty)$ *with probability (w.p.) 1, where*

$$\bar{V}(\infty) = \Lambda(24)^{-1} \int_0^{24} E_\infty[X(u)]\lambda(u)du.$$

We also define the average performance seen by arriving customers as

$$\bar{U}(t) := \frac{\int_0^t X(u)dA(s)}{A(t)}$$

The following theorem was proved in (Wolff 1982) and is known as the time-varying version of "Poisson Arrivals See Time Averages".

Theorem 3 $\bar{V}(t) \to \bar{V}(\infty)$ *w.p. 1 if and only if* $\bar{U}(t) \to \bar{V}(\infty)$ *w.p. 1, as* $t \to \infty$.

5.1. An infinite server model

In order to understand the role of the inspection times on system performance, we wish to understand how the inspection times impact the number of customers in the system. To do this, we start by examining an infinite server queueing model with inspection times: $M_t/M(\mathbf{T})/\infty$. This will provide insights into the role of the inspection times and the delayed release of servers. Though there is no waiting in the case of an infinite server model, the timing of inspections can impact server occupancy. If we wish to minimize server occupancy levels, it is desirable to schedule inspections at times when many jobs have recently completed service in order to quickly clear the system of these jobs. We will also extend the analysis to general service time distributions in Section 6.1. As we are interested in long-run average performance measures, we work with the periodic equilibrium distribution of the stochastic system, i.e. we assume the system starts operating from the infinite past and we analyze its performance on $\mathbb{R}^+$.

We distinguish between two classes of customers: class (1) customers that have not finished service, and class (2) customers that have finished service but are still in the system (waiting for inspection). Let $Z_i(t)$ denote the number of class i customers in the system at time t for $i = 1, 2$. Then the total number of customers in the system is $Q(t) = Z_1(t) + Z_2(t)$. Regardless of the discharge policy, in equilibrium, $Z_1(t)$ follows a Poisson distribution with mean

$$m(t) = \int_{-\infty}^{t} \bar{F}(t-s)\lambda(s)\, ds,$$

where $\bar{F}$ denotes the complimentary cumulative distribution function of the service time distribution (Eick et al. 1993b).

For t between two adjacent discharge times, τ_i and τ_{i+1}, $Q(t)$ is monotonically increasing in t. This is because $Q(t)$ includes all of the customers who had not finished service prior to the previous inspection time, τ_i, (Γ in Figure 4), and the new arrivals between τ_i and t (H in Figure 4). In particular, $Q(t)$ follows a Poisson distribution with mean $m(\tau_i) + \int_{\tau_i}^{t} \lambda(u)du$ in stationarity. We notice that

1. The time-averaged expected occupancy level over $[\tau_i, \tau_{i+1})$ is

$$\frac{\int_{\tau_i}^{\tau_{i+1}} E_\infty[Q(t)]\lambda(t)dt}{\int_{\tau_i}^{\tau_{i+1}} \lambda(t)dt} = \frac{\int_{\tau_i}^{\tau_{i+1}} \left(m(\tau_i) + \int_{\tau_i}^{t} \lambda(u)du\right)\lambda(t)dt}{\int_{\tau_i}^{\tau_{i+1}} \lambda(t)dt} = m(\tau_i) + \frac{1}{2}\int_{\tau_i}^{\tau_{i+1}} \lambda(t)dt.$$

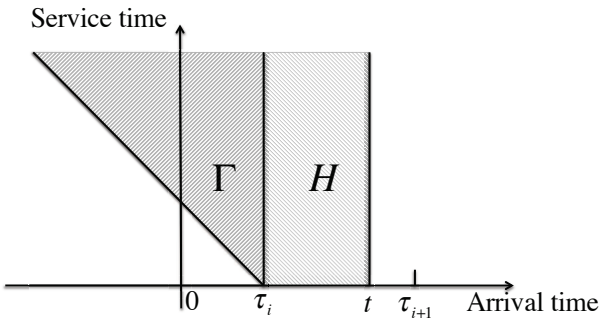
2. The maximum expected occupancy level on $[\tau_i, \tau_{i+1})$ is

$$E_\infty[Q(\tau_{i+1}-)] = m(\tau_i) + \int_{\tau_i}^{\tau_{i+1}} \lambda(t)dt,$$

where $E_\infty[Q(t-)] = \lim_{s\uparrow t} E_\infty[Q(s)]$.

In what follows, we look at two performance measures for the infinite server model. The first is the time-averaged expected occupancy level over a day, which we denote as $\zeta_N(T)$ when there are N equally spaced inspections per day with the first inspection at time T. While making operational decisions based on this

Figure 4 **Point process representation of infinite server queue. The horizontal axis depicts possible arrival times, while the vertical axis depicts service times. Thus, the point (t_i, V_i) represents a job which arrives at time t_i with service time realization of V_i. Any arrivals in the shaded region $\Gamma \cup H$, will remain in the system at time t. The jobs in Γ are those who arrived before τ_i but haven't completed service before τ_i. Those jobs in H are those who have arrived following inspection at time τ_i. So at the very least, the jobs in the shaded region will remain in the system until the next inspection at time τ_{i+1}.**



measure may lead to reasonable and sometimes very good performance in some service systems, it may fall short in settings where it is essential to ensure very speedy access to service with waiting times much shorter than the mean service time (e.g. hospital care). As such, we also consider a second measure: the maximum expected occupancy level within a day, which we denote as $\eta_N(T)$. Our primary focus will be on $\eta_N(T)$ as it provides insight into how many servers are necessary to provide *high* service quality.

$M_t/M/\infty$: If we allow continuous inspection (discharge a customer whenever he finishes service), then the system evolves as a standard infinite server queueing system with time-inhomogeneous arrival rates with $Q(t) = Z_1(t)$. Thus, the time-average expected occupancy level is:

$$\zeta_\infty := \Lambda(24)^{-1} \int_0^{24} m(t)\lambda(t)dt,$$

and the maximum expected occupancy level is

$$\eta_\infty := \max_{0 \le t \le 24} m(t).$$

$N = 1$: If we have one inspection per day at time T, then

$$\zeta_1(T) := \frac{\int_T^{T+24} \left(m(T) + \int_T^t \lambda(u)du \right) \lambda(t)dt}{\int_T^{24+T} \lambda(t)dt} = m(T) + \frac{1}{2}\Lambda(24),$$

and

$$\eta_1(T) := m(T) + \Lambda(24).$$

General $N > 1$: If we have N equally spaced inspections per day with the first inspection at time T, then

$$\zeta_N(T) = \sum_{k=1}^{N} \left(m(T + (k-1)\Delta_N) + \frac{1}{2} \int_{T+(k-1)\Delta_N}^{T+k\Delta_N} \lambda(t)dt \right) \frac{\int_{T+(k-1)\Delta_N}^{T+k\Delta_N} \lambda(t)dt}{\Lambda(24)}.$$

If we denote $\eta_N^{(i)}(T)$ as the expected occupancy level right before the i-th inspection time for $1 \leq i \leq N$. Then $\eta_N^{(i)}(T) = m(T + (i-2)\Delta_N) + \int_{T+(i-2)\Delta_N}^{T+(i-1)\Delta_N} \lambda(t)\, dt$, and

$$\eta_N(T) := \max_{1 \leq i \leq N} \eta_N^{(i)}(T).$$

Remark 2 *We can also construct performance measures based on the number of servers occupied by customers who have already finished service (class 2). These measures may be of independent/additional interests to hospital managers. In particular, $Z_2(t)$, for $t \in [\tau_i, \tau_{i+1})$, follows a Poisson distribution with mean*

$$E[Z_2(t)|t \in [\tau_i, \tau_{i+1})] = m(\tau_i) + \int_{\tau_i}^{t} \lambda(u)du - m(t) = -\int_{-\infty}^{t} \lambda(s)(\bar{F}(\tau_i - s) - \bar{F}(t-s))ds.$$

For illustrative purposes, we consider the following specific form of the arrival rate function

Assumption 1 *The arrival rate takes the following form:*

$$\lambda(t) = \bar{\lambda} + \beta \sin(\gamma t)$$

where $0 < \beta \leq \bar{\lambda}$ and $\gamma = 2\pi/24$.

Sinusoidal arrival rate functions are commonly assumed in the literature to capture periodic behavior. In this case, β measures the variability of the arrival rate function, and is called the amplitude of the arrival rate.

Under Assumption 1, the following relationship was shown in Eick et al. (1993a):

$$m(t) = \int_{-\infty}^{t} \lambda(s)\bar{F}(t-s)\, ds = \frac{\bar{\lambda}}{\mu} + \beta\left(\frac{\mu}{\mu^2+\gamma^2}\sin(\gamma t) - \frac{\gamma}{\mu^2+\gamma^2}\cos(\gamma t)\right)$$

As a direct result of this relationship, we have that

$$m(t) = \frac{\bar{\lambda}}{\mu} + \frac{\beta}{\sqrt{\mu^2+\gamma^2}}\sin(\psi + \gamma t)$$

where ψ satisfies i) $-\pi < \psi \leq \pi$, ii) $\sin\psi = -\gamma/\sqrt{\mu^2+\gamma^2}$ and iii) $\cos\psi = \mu/\sqrt{\mu^2+\gamma^2}$. Thus, the maximum number of customers in the system who have not finished service is given by $\max_t m(t) = \bar{\lambda}/\mu + \beta/\sqrt{\mu^2+\gamma^2}$. Depending on the relative values of μ (time scale of service time) and γ (time scale at which we see variability in arrival rate), $m(t)$ could attain its maximum anywhere on the interval $[18, 24]$.

In the hospital setting, we are interested in the case where the time-scale for service (days) is much longer than the time-scale of fluctuation in the arrival rate (hours). Chan et al. (2014) found that in such a regime, ignoring the time-variability of arrivals does not significantly alter queueing dynamics. However, we find this not to be the case when considering the discharge dynamics of our model. It is precisely the time-variation in arrival rates which makes this problem interesting and challenging. In fact, without time-variation of the arrival rates, the timing of the inspection times (beyond being evenly spaced throughout the day) does not matter at all. We now consider the following scenario:

Assumption 2 *The service rate and period of the arrival rate function are related as:*

$$\mu \ll \gamma.$$

Based on Assumptions 1 and 2, we have the following approximation of $m(t)$:

$$m(t) \approx m_0(t) = \frac{\bar{\lambda}}{\mu} - \frac{\beta}{\gamma}\cos(\gamma t)$$

In what follows, we work under the above two assumptions (Assumption 1 & 2) and use the approximation $m_0(t)$ to replace $m(t)$. In particular, the closed form expressions we developed for $\zeta_N(T)$'s and $\eta_N(T)$'s are based on the approximation $m_0(t)$.

Lemma 3 *When applying the approximation $m_0(t)$ for $m(t)$, the time-average expected occupancy level for a single inspection time is:*

$$\zeta_1(T) = \bar{\lambda}(1/\mu + 12) - \frac{\beta}{\gamma}\cos(\gamma T),$$

and for $N \geq 2$

$$\zeta_N(T) = \bar{\lambda}(1/\mu + \Delta_N/2).$$

Lemma 3 suggests having more inspection times will reduce the time-average expected occupancy level but the marginal gains from adding more inspections is decreasing. Moreover, as the number of inspection times increases, the performance of the $M_t/M(\mathbf{T})/\infty$ system converges to that of the $M_t/M/\infty$ system:

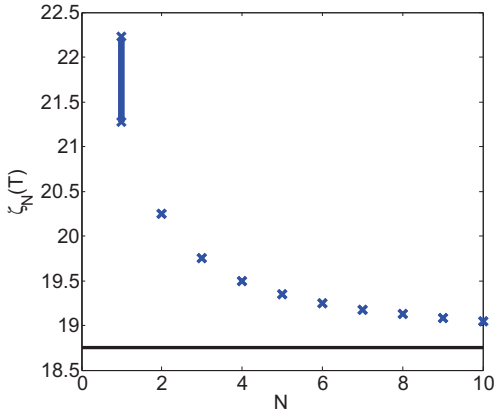
$$\lim_{N\to\infty} \zeta_N(T) = \bar{\lambda}/\mu.$$

We also observe that, while a well chosen T results in improvement in performance measured by the time-average expected occupancy level when there is only one inspection per day, the timing of inspections does not affect the time-average expected occupancy level when there is more than one inspection per day ($N \geq 2$).

When there is only one inspection per day, the improvement we can gain from carefully choosing the inspection time depends on the value of β, which measures the magnitude of variability of the arrival rate. The larger the value of β, the more we gain from carefully selecting the inspection times. For our particular arrival rate function, the peak arrival rate is at $t = 6$. Then, the inspection time that leads to the minimum time-average expected occupancy level is 6 hours before the peak arrival and the inspection time that leads to the maximum time-average expected occupancy level is 6 hours after the peak arrival.

Figure 5 plots the time-average expected occupancy level, $\zeta_N(T)$, for different values of N. The interval at $N = 1$ indicates the range of values that can be achieved by different inspection times when having only one inspection per day. The horizontal line is at $\zeta_\infty(T) = \bar{\lambda}/\mu$. Notice that the reduction in time-average expected occupancy level due to optimizing a single inspection time (from 22.23 to 21.27) can

Figure 5 **Time-average expected occupancy level of an $M_t/M(\mathbf{T})/\infty$ queueing system: $\zeta_N(T)$ for different values of N ($\bar{\lambda} = 0.25$, $\beta = 0.5\bar{\lambda}$, $\mu = 1/75$)**



be comparable in magnitude to the reduction that can be achieved by adding in a second inspection time (from 21.27 to 20.25). Additionally, we see that having more inspection times improves performance, but the marginal gains are diminishing.

While the timing of inspections does not affect the time-average expected occupancy levels when there are multiple inspections per day, this is not the case for the maximum expected occupancy level. We now consider $\eta_N(T)$. We first notice that $\max_t m_0(t) = \frac{\bar{\lambda}}{\mu} + \frac{\beta}{\gamma}$. When substituting $m_0(t)$ for $m(t)$,

$$\begin{aligned}\eta_N^{(i)}(T) &= \frac{\bar{\lambda}}{\mu} - \frac{\beta}{\gamma}\cos(\gamma(T + (i-2)\Delta_N)) + \int_{T+(i-2)\Delta_N}^{T+(i-1)\Delta_N} \lambda(t)\,dt \\ &= \frac{\bar{\lambda}}{\mu} + \bar{\lambda}\Delta_N - \frac{\beta}{\gamma}\cos(\gamma(T + (i-1)\Delta_N)) \qquad (1)\end{aligned}$$

We observe from (1) that we can gain some improvement in performance, measured by $\eta_N(T)$, by appropriately choosing the inspection time T. The magnitude of improvement varies and is determined by the value of β. The larger the daily variation in arrival rate, the more we gain from a carefully chosen inspection time. This suggests that the greater the variability in the daily arrival rate, the more important it is to optimize the timing of the inspections. When there is little variability, system managers can select inspection times based on 'convenience' as it will have little impact on system performance as measured by the maximum expected occupancy level.

Let $\Xi(N) = \min_T \eta_N(T)$ denote the minimum maximum expected occupancy level when there are N equally spaced inspection times per day.

Lemma 4 *When substituting $m_0(t)$ for $m(t)$, $\arg\min_T \eta_N(T) = 12 - \Delta_N/2 - \lceil N/2 - 1\rceil\Delta_N$, and*

$$\Xi(N) = \frac{\bar{\lambda}}{\mu} + \bar{\lambda}\Delta_N - \frac{\beta}{\gamma}\cos\left(\pi - \frac{\pi}{N}\right).$$

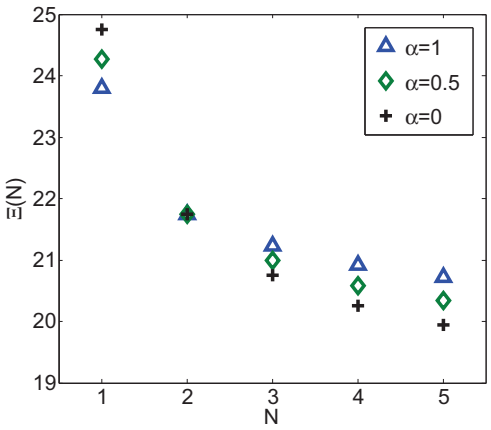
Lemma 4 indicates that the optimal inspection times are symmetrically distributed around $t = 12$ (the peak of $-\beta/\gamma\cos(\gamma T)$). In particular, if there is an odd number of inspection times, there will always be one inspection 6 hours before the peak arrival, so that there exists some i such that $T_i = 0$.

As with the stability condition, having more inspection times helps, but the marginal effect is decreasing. Specifically, as $\Xi'(N) < 0$ and $\Xi''(N) > 0$ for $N \geq 0$, $\Xi(N)$ is decreasing in N and

$$\Xi(N+1) - \Xi(N+2) < \Xi(N) - \Xi(N+1).$$

This suggests that having more inspection times will improve the system performance, but the improvement is diminishing. Figure 6 plots $\Xi(N)$ for different values of $\alpha = \beta/\bar{\lambda}$.

Figure 6 **Maximum expected occupancy level in an $M_t/M(\mathbf{T})/\infty$ queueing system after optimizing the inspection time(s) $\mathbf{T}$: $\Xi(N)$ for different values of N ($\bar{\lambda} = 0.25$, $\mu = 1/75$)**



We have now seen how the inspection time(s) determines the maximum expected occupancy level in our infinite server model. Moreover, we can optimize over this expression to minimize the maximum expected occupancy level. Doing so will likely result in high service quality given the capacity constraint in our original finite server model.

5.2. Optimal inspection times

In this section, we numerically test how the choice of inspection times affects the performance of the $M_t/M(\mathbf{T})/s$ queue and whether the insights we gained from the infinite server model can be carried over to the finite server case. For our numerical illustrations, we use the arrival rate function $\lambda(t) = \bar{\lambda} + \beta\sin(\gamma t)$, where $\beta = \bar{\lambda}/2$, $\gamma = \pi/12$, and service rate $\mu = 1/75$. Thus, the mean nominal service time–a lower bound to the time a customer occupies a server–is about three times longer than the arrival rate period (75 hours versus 24 hours). Our choice of service time is consistent with the typical average ICU length of stay (LOS) (Chan et al. 2014). In the internal wards, the typical LOS is 127 hours (Dai and Shi 2014). As our results rely on the assumption that $\mu \ll \gamma$, the smaller the service rate, the more accurate we expect our approximations

will be. Indeed, while we only present results for $\mu = 1/75$, numeric studies suggest our approximations are even more accurate when $\mu = 1/130$. Additionally, we vary the value of $\bar{\lambda}$ to see how the system scale affects performance.

5.2.1. Measures for the number of customers in the system We start by examining the measures we analyzed for the infinite server model, the time-average expected occupancy level, $\zeta_N(T)$, and the maximum expected occupancy level, $\eta_N(T)$.

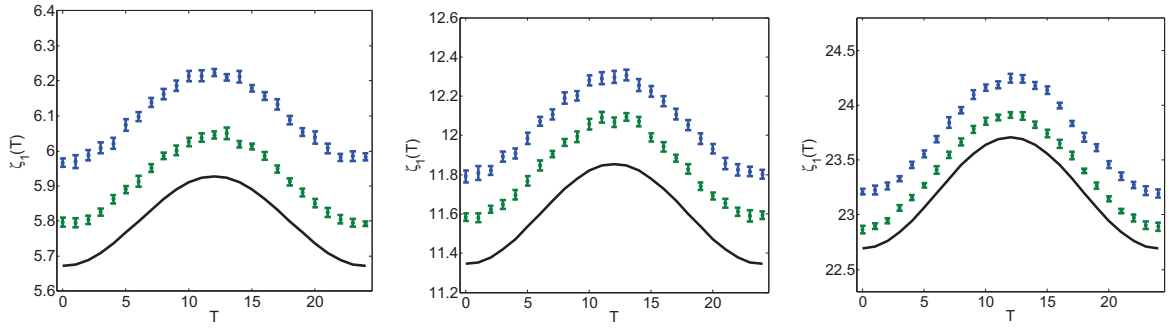
For the infinite server model, recall from Lemma 3 and Equation (1) that when there is only one inspection per day, both the time-average expected occupancy level, $\zeta_1(T)$, and the maximum expected occupancy level, $\eta_1(T)$, are functions of the inspection time T. When there are multiple inspections per day, the average occupancy level $\zeta_N(T)$ is a constant, while the maximum expected occupancy level, $\eta_N(T)$ does vary with the inspection time T.

Turning back to the finite server model, let $\rho = \bar{\lambda}/(s\mu)$ denote the nominal load for the system–this ignores the inflation of time spent occupying a server which is introduced by the need for inspection prior to discharge. We also denote $\rho_e(N) = \lambda/(sN(1 - e^{-\mu 24/N})/24)$ as the effective load when we have N equally spaced inspections per day. Notice that $\rho_e(N) > \rho$ for the same staffing level s as the inspection time required for discharge prolongs the actual length of stay of each patient.

Figure 7 illustrates how the average occupancy levels, $\zeta_N(T)$, change with the inspection time using simulation. We use the method of batch means with 20 batches and each batch containing arrivals for about 5000 days. We plot both the mean values and the 95% confidence intervals. We compare the performance between one inspection per day and two inspections per day. The numerical experiments are conducted for staffing levels (nominal occupancy rates). As our analysis is based on an infinite server system, we expect it to be more accurate as the system scale grows. Thus, it is important to also verify our insights translate to small systems. As such, we also consider system scale parameters.

As with most queueing systems, we observe economies of scale as the system size increases. In particular, as the system scale gets larger, we can increase the system load to achieve similar performance seen in smaller systems with lower loads. When $\bar{\lambda} = 5/75 = 0.0667$, we set $s = 9$ and 10, which correspond to the nominal load and, correspondingly the nominal occupancy level, of 55.56% and 50%, respectively. When $\bar{\lambda} = 10/75 = 0.1333$, we set $s = 16$ and 17, which correspond to the nominal occupancy level of 62.5% and 58.82% respectively. When $\bar{\lambda} = 20/75 = 0.2667$, we set $s = 30$ and 32, which correspond to the nominal occupancy level of 66.67% and 62.5% respectively. The effective occupancy level is much higher than the nominal occupancy level. For example, when $\bar{\lambda} = 0.0667$, $s = 9$ and there is only one inspection per day, the effective (simulated) occupancy level is $\rho_e(1) = \bar{\lambda}/(s(1 - e^{-24\mu})/24) = 64.95\%$ (v.s. $\rho = 55.56\%$), and when $\bar{\lambda} = 0.2667$, $s = 30$ and there is only inspection per day, the effective occupancy level is $\rho_e(1) = 77.91\%$ (v.s. $\rho = 66.67\%$).

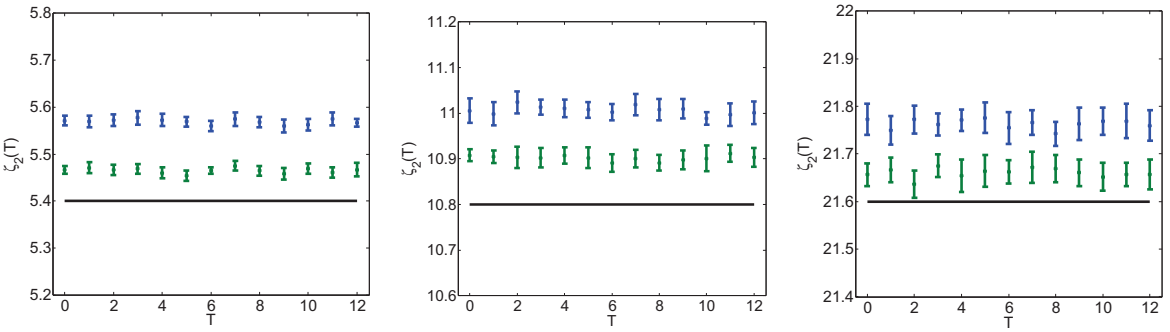
Figure 7 $M_t/M(\mathbf{T})/s$ **queueing system: Time-average expected occupancy level as a function of** T **for different staffing levels**



(a) $\bar{\lambda} = 0.0667$ (upper: $s = 9$, middle: $s = 10$, lower: $s = \infty$)

(b) $\bar{\lambda} = 0.1333$ (upper: $s = 16$, middle: $s = 17$, lower: $s = \infty$)

(c) $\bar{\lambda} = 0.2667$ (upper: $s = 30$, middle: $s = 32$, lower: $s = \infty$)



(d) $\bar{\lambda} = 0.0667$ (upper: $s = 9$, middle: $s = 10$, lower: $s = \infty$

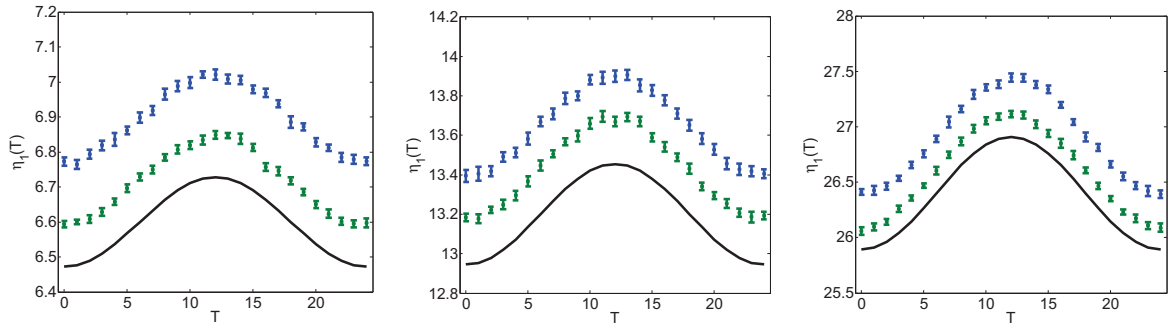
(e) $\bar{\lambda} = 0.1333$ (upper: $s = 16$, middle: $s = 17$, lower: $s = \infty$)

(f) $\bar{\lambda} = 0.2667$ (upper: $s = 30$, middle: $s = 32$, lower: $s = \infty$)

We note that despite the small size of the systems we simulate, the *shape* of the average occupancy levels are very much aligned with that of the infinite server analysis (depicted by a solid line). Specifically, when there is a single inspection time, the inspection time T which minimizes (maximizes) the average occupancy level is the same as that given by the analysis of the infinite server model. When there are two inspection times, we see that, as suggested by our analysis in Section 5.1, the average occupancy level is almost invariant to the time the inspections occur (the fluctuations of the simulated $\eta_2(T)$, for different T, are less than 0.5% of $\eta_2(0)$). Certainly, as the system becomes less congested, the finite server system behaves more similarly to the infinite server system, so we see that the average occupancy level approaches (and will eventually converge to) the solid line given by the infinite server analysis.

We also observe that adding a second inspection improves system performance significantly. For example, when $\bar{\lambda} = 0.1333$, $s = 16$, adding a second inspection reduces the average occupancy level from 11.8 (when $T = 0$) to 11, which is a decrease of 7.27%. Moreover, adding a second inspection time makes the system performance more robust across different choices of inspection times. Indeed, for the average occupancy level, adding a second inspection makes the choice of inspection time almost irrelevant.

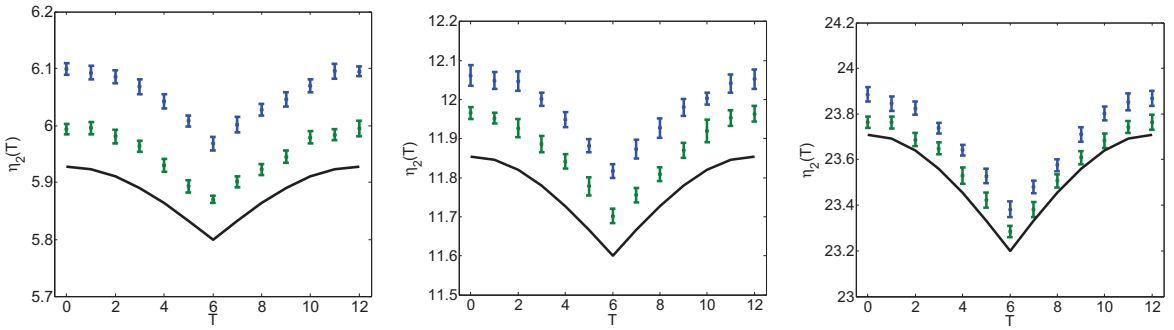
Figure 8 $M_t/M(\mathbf{T})/s$ **queueing system: Maximum expected occupancy level as a function of** T **for different staffing levels**



(a) $\bar{\lambda} = 0.0667$ (upper: $s = 9$, middle: $s = 10$, lower: $s = \infty$)

(b) $\bar{\lambda} = 0.1333$ (upper: $s = 16$, middle: $s = 17$, lower: $s = \infty$)

(c) $\bar{\lambda} = 0.2667$ (upper: $s = 30$, middle: $s = 32$, lower: $s = \infty$)



(d) $\bar{\lambda} = 0.0667$ (upper: $s = 9$, middle: $s = 10$, lower: $s = \infty$)

(e) $\bar{\lambda} = 0.1333$ (upper: $s = 16$, middle: $s = 17$, lower: $s = \infty$)

(f) $\bar{\lambda} = 0.2667$ (upper: $s = 30$, middle: $s = 32$, lower: $s = \infty$)

We next consider maximum occupancy levels, $\eta_N(T)$. Figure 8 demonstrates how $\eta_N(T)$ depends on the inspection time(s) based on simulation. We use the same set of parameters as in Figure 7. Specifically, we investigate the system performance for different system scales and staffing levels. We also compare performance between the one inspection per day policy and the two inspections per day policy. Similar to the case for $\zeta_N(T)$, we observe that $\eta_N(T)$, when plotted as a function of the inspection time T, takes the same shape as for the infinite server queue model. The actual value is a vertical upward shift of the curve suggested by the infinite sever model; the less heavily loaded the system (increasing the number of servers, s), the smaller the scale of the upward shift. Additionally, we see that the second inspection time again improves system performance. While the maximum occupancy level does depend on the precise schedule of the inspection times, the amount of variation in performance between the optimal and worst inspection times is less when there are two inspections compared to one. This trend continues as we add more inspection times. For the sake of space, we do not include the figures.

5.2.2. Measures of waiting When considering the number of customers in the system, we find that the insights from the infinite server model translate very well to the finite server setting. On the other hand, one

may want to consider other performance measures. In particular, in service settings–especially in healthcare systems–customer (patient) waits are of great interest. Unfortunately, when considering an infinite server system, there is never any waiting involved. Still, we wish to explore how our insights from the infinite server system may provide some insight into measures of waiting in a finite server setting.

We start by considering the mean probability of waiting of each arriving customer:

$$\bar{V}_p(N,T) := \lim_{t\to\infty} \frac{1}{\Lambda(t)} \int_0^t 1\{Q(u) \geq s\}\lambda(u)du = \frac{\int_0^{24} P_\infty(W(u) > 0)\lambda(u)du}{\Lambda(24)}.$$

From Theorem 3 we have

$$\lim_{t\to\infty} \frac{\int_0^{\Lambda(t)} 1\{Q(u) \geq s\}dA(u)}{A(t)} = \bar{V}_p(N,T).$$

Using the same simulation depicted in Figures 7 and 8, Figure 9 plots the simulated probability of waiting, $\bar{V}_p(N,T)$, as a function of the first inspection time, T. We see that the probability of waiting, $\bar{V}_p(N,T)$, follows practically the same shape as the average occupancy level, $\zeta_N(T)$. This is because the probability of waiting, $P_\infty(Q \geq s)$, is very closely related to the queue length process. Thus, if our goal were to simultaneously minimize the average occupancy level and the probability of waiting, we could utilize the analysis of an infinite server model to select the optimal inspection time(s).

A separate measure of waiting which may be of interest is the mean waiting time as seen by arriving customers. Specifically, we define:

$$\bar{V}_w(N,T) := \lim_{t\to\infty} \frac{1}{\Lambda(t)} \int_0^t W(u)\lambda(u)du = \frac{\int_0^{24} E_\infty[W(u)]\lambda(u)du}{\Lambda(24)}$$

again, from Theorem 3 we have

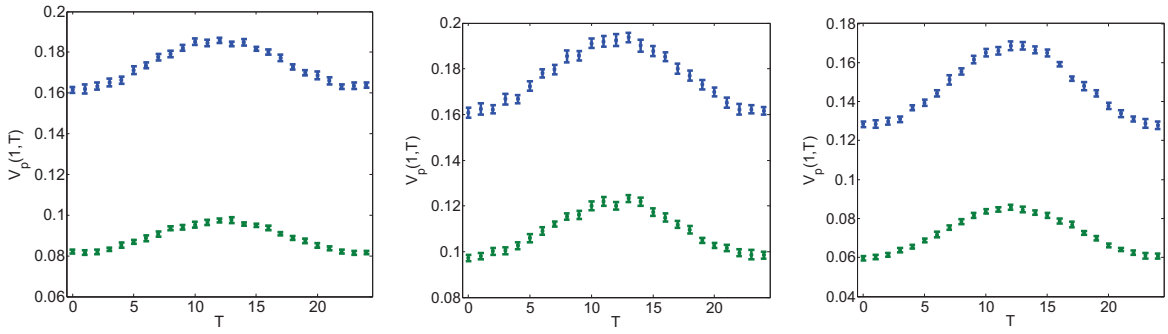
$$\lim_{t\to\infty} \frac{\int_0^t W(u)dA(u)}{A(t)} = \bar{V}_w(N,T)$$

Figure 10 plots the average waiting time, $\bar{V}_w(N,T)$, as a function of the first inspection time, T. We start by considering the case of a single inspection time. We can see that, unlike the probability of waiting, the expected waiting time, $\bar{V}_w(N,T)$, looks quite different from the average occupancy level, $\zeta_N(T)$. In Section 4, we observed that, for the time-varying performance measures (rather than the arrival rate weighted time averages considered in this section), the average waiting time, $E_\infty[W(T+t)]$, behaves quite differently from the average queue length, $E_\infty[Q(T+t)]$, and the probability of waiting $P_\infty(W(T+t)>0)$. We aim to give an intuitive explanation for this phenomenon. If the system is very heavily loaded, then almost every incoming customer will need to wait until the next inspection time to be admitted into service. As such,

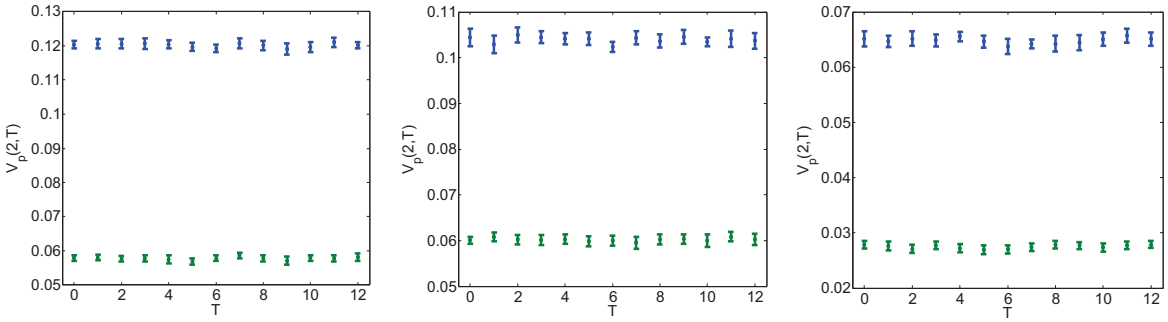
$$E\left[\int_T^{T+24} (T+24-t)\, dA(t)\right] = \int_T^{T+24} (T+24-t)\lambda(t)\, dt = \frac{24^2}{2}\bar{\lambda} + 24\frac{\beta}{\gamma}\cos(\gamma T).$$

which takes exactly the opposite shape of $\zeta_1(T) = \bar{\lambda}(\frac{1}{\mu}+12) - \frac{\beta}{2\gamma}\cos(\gamma T)$.

Figure 9 $M_t/M(\mathbf{T})/s$ **queueing system: Probability of waiting as a function of** T **for different staffing levels**



(a) $\bar{V}_p(1,T)$: $\bar{\lambda}=0.0667$ (upper: $s=9$, lower: $s=10$) (b) $\bar{V}_p(1,T)$: $\bar{\lambda}=0.1333$ (upper: $s=16$, lower: $s=17$) (c) $\bar{V}_p(1,T)$: $\bar{\lambda}=0.2667$ (upper: $s=30$, lower: $s=32$)



(d) $\bar{V}_p(2,T)$: $\bar{\lambda}=0.0667$ (upper: $s=9$, lower: $s=10$) (e) $\bar{V}_p(2,T)$: $\bar{\lambda}=0.1333$ (upper: $s=16$, lower: $s=17$) (f) $\bar{V}_p(2,T)$: $\bar{\lambda}=0.2667$ (upper: $s=30$, lower: $s=32$)

We notice that this phenomenon does not seem to exist when there are two inspection times. In fact, as was the case for the average occupancy level, the expected waiting time is practically invariant to the timing of the inspections. Thus, we see again that with more inspection times, the system performance is more robust to the actual choices of the inspection times as long as they are equally spaced.

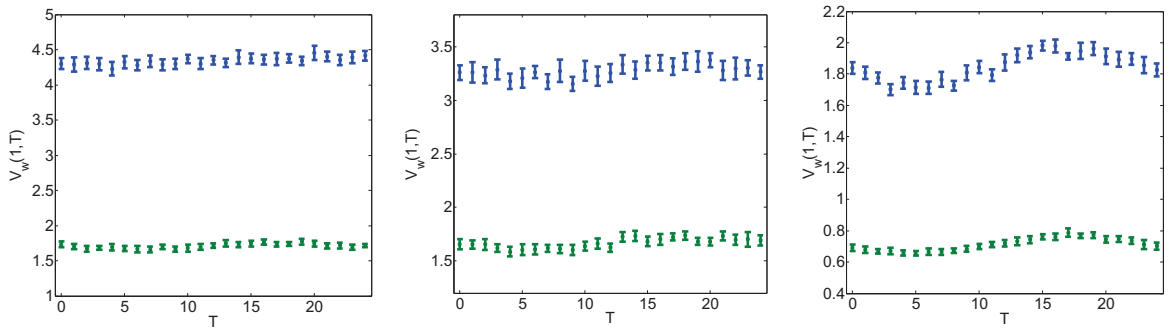
6. Model Robustness

In this section, we consider several generalizations of the results in the previous sections. Specifically, we will relax the following assumptions: exponential service times, equally spaced inspection times and a sinusoidal arrival rate function. We find that, in many cases, the infinite server approximation continues to provide useful insights.

6.1. General service time distributions

While the exponential assumption on service times is useful for analysis and necessary for our stability results in Section 3, it may be desirable to relax this assumption when considering certain systems. Fortunately, we find that our analysis using the infinite server system can be easily extended to general service time distributions. For the analysis in this subsection, we keep the sinusoidal assumption on the arrival rate

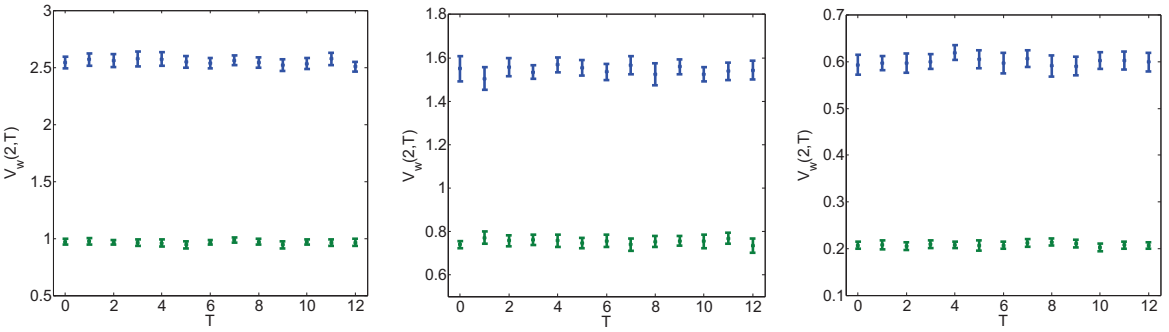
Figure 10 $M_t/M(\mathbf{T})/s$ **queueing system: Expected waiting time as a function of** T **for different staffing levels**



(a) $\bar{V}_w(1,T)$: $\bar{\lambda} = 0.0667$ (upper: $s = 9$, lower: $s = 10$)

(b) $\bar{V}_w(1,T)$: $\bar{\lambda} = 0.1333$ (upper: $s = 16$, lower: $s = 17$)

(c) $\bar{V}_w(1,T)$: $\bar{\lambda} = 0.2667$ (upper: $s = 30$, lower: $s = 32$)



(d) $\bar{V}_w(2,T)$: $\bar{\lambda} = 0.0667$ (upper: $s = 9$, lower: $s = 10$)

(e) $\bar{V}_w(2,T)$: $\bar{\lambda} = 0.1333$ (upper: $s = 16$, lower: $s = 17$)

(f) $\bar{V}_w(2,T)$: $\bar{\lambda} = 0.2667$ (upper: $s = 30$, lower: $s = 32$)

function.

6.1.1. Optimizing inspection times Let $\mu = 1/E[V]$. Then for general service times,

$$\begin{aligned} m(t) &= \frac{\bar{\lambda}}{\mu} + \frac{\beta}{\mu}\int_0^\infty \sin(\gamma(t-v))f_e(v)dv \\ &= \frac{\bar{\lambda}}{\mu} + \frac{\beta}{\mu}\left(E[\cos(\gamma V_e)]\sin(\gamma t) - E[\sin(\gamma V_e)]\cos(\gamma t)\right) \end{aligned}$$

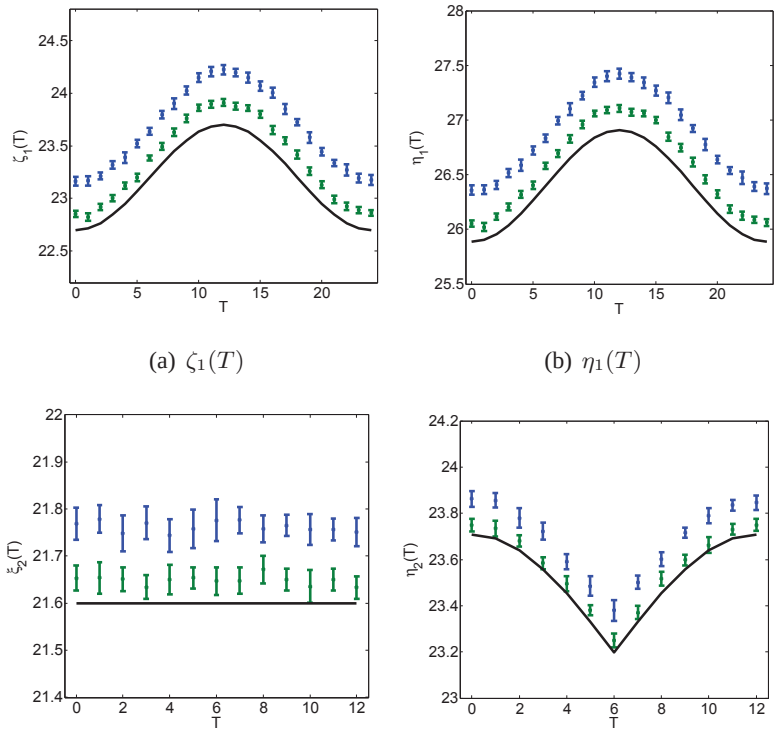
where f_e is the pdf of V_e, the equilibrium distribution of the service times. Based on this explicit expression for $m(t)$, we can solve the optimization problem, at least numerically, to find the optimal discharge policies for the infinite server queue model.

Regardless of the service time distribution, we observe from our numerical experiments that the corresponding infinite server queues provide useful insight into the dynamics of the queue length processes. Figure 11 provides a specific example where the service times are lognormally distributed.

Furthermore, approximating $m(t)$ with

$$m_0(t) = \frac{\bar{\lambda}}{\mu} - \frac{\beta}{\gamma}\cos(\gamma t)$$

Figure 11 $M_t/G(\mathbf{T})/s$ **queueing system with lognormal service time distribution: The average and maximum expected occupancy level as a function of the discharge time** T **for different staffing levels (**$\bar{\lambda} = 0.2665$**,** $\log V \sim N(3.818, 1)$**, upper:** $s = 30$**, middle:** $s = 32$**, lower** $s = \infty$**)**



(a) $\zeta_1(T)$ (b) $\eta_1(T)$

(c) $\zeta_2(T)$ (d) $\eta_2(T)$

is still appropriate for a number of service time distributions. As we saw earlier, this can greatly simplify the calculations. Figure 12 shows a few examples where $m_0(t)$ serves as a good approximation. This can occur when the service rate is a continuous random variable (i.e. $\bar{F}(v)$ is differentiable) and $\mu \ll \gamma$, because

$$\begin{aligned} m(t) &= \frac{\bar{\lambda}}{\mu} + \beta \int_0^\infty \sin(\gamma(t-v))\bar{F}(v)dv \\ &= \frac{\bar{\lambda}}{\mu} - \frac{\beta}{\gamma}\cos(\gamma t) - \frac{\beta}{\gamma}\int_0^\infty \cos(\gamma(t-v))f(v)dv \end{aligned}$$

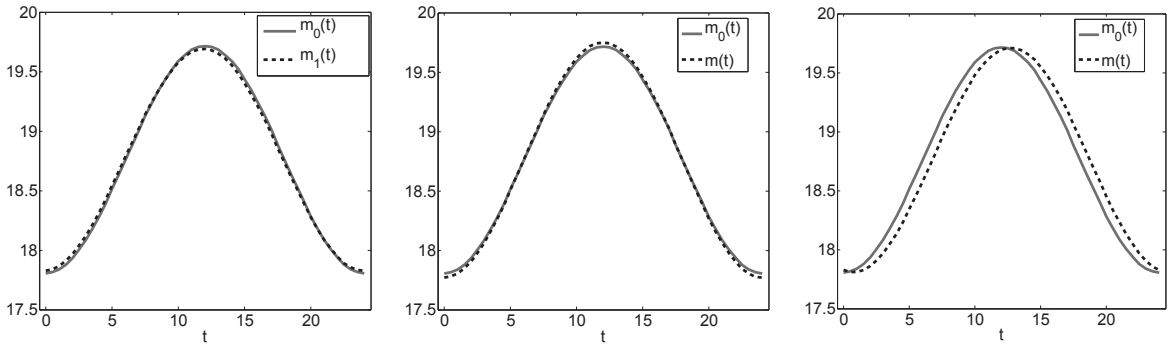
with the last term, $(\beta/\gamma)\int_0^\infty \cos(\gamma(t-v))f(v)dv \approx 0$, in a number of instances.

That said, it is also true that using $m_0(t)$ can be a very poor approximation for some service time distributions. An extreme example is the case when the service time is deterministic, i.e. $V = a$ for some constant a. In this case,

$$m(t) = \frac{\bar{\lambda}}{\mu} - \frac{\beta}{\gamma}\cos(\gamma t) + \frac{\beta}{\gamma}\cos(\gamma(t-a))$$

Depending on the value of a, it could take exactly the opposite shape as $m_0(t)$. In general when the service time distribution has very small variance, $m_0(t)$ will not serve as a good approximation. In the hospital setting, this could be the case for some surgical units where patients undergo very similar procedures that

Figure 12 Approximation for $m(t)$ with different service time distribution ($\bar{\lambda} = 0.25$,$\beta = \bar{\lambda}$)



(a) Uniform ($(0, 150.08)$) (b) Lognormal ($\log V \sim N(3.818, 1)$) (c) Pareto ($\alpha = 2, x_m = 37.52$)

are determined by evidence-based protocols. In such settings, it would be advisable to leverage information about the service time distribution to determine when inspections should occur, and when the service time is approximately deterministic, the formula for deterministic service times might serve as a good approximation.

When the service distribution has moderate variance (e.g. exponential and lognormal), $m_0(t)$ serves as a good approximation of $m(t)$. Then the essential information we need to know is the mean of the service time and the amplitude of the arrival rate to optimize the inspection times. In particular, we see that the shape of the time-averaged and maximum expected occupancy for the lognormal service time distribution is almost the same as that of the exponential service time distribution with the same mean.

6.1.2. Stability analysis Our results in Section 3 required the exponential assumption for the service time distributions. That said, we conjecture that when $m_0(t)$ is a good approximation for $m(t)$, these insights will still hold. This is, in general, the case for service time distributions with moderate variance. However, it is easy to see that this does not hold for all service time distributions. Consider a simple example with deterministic service times, D, and an arrival rate pattern where the arrival rate is 0 everywhere, except that it is $\Lambda(24)/2$ patients/hour from 0 to 1 and again from 2 to 3. With 2 inspection times, it is easy to see that to maximize the stability region, the inspections should occur at $T_1 = 1 + D + \epsilon$ and $T_2 = 3 + D + \epsilon$. Thus, we see that equally spaced inspection times is *not* an optimal strategy in this case.

6.2. Unequally spaced inspection times

Even though equally spaced inspection times maximizes the stability region when service times are exponentially distributed, it may not necessarily optimize system performance as measured by the occupancy level when the system is stable. Additionally, we just saw that equally spaced inspection times do not maximize the stability region in general. As such, we next consider optimizing the inspection times when they are not evenly spaced.

In order to gain some insight into the possible gains achieved by relaxing the requirement of equally spaced inspection, we consider the case of two inspections per day. Moreover, we assume a sinusoidal arrival rate function and exponential service times. In this case, $m_0(t)$ is a good approximation for $m(t)$. We start by analyzing the maximum expected occupancy level of the corresponding infinite server model.

As before, we let T denote the first discharge time in a day and Δ denote the distance between the first and second inspection time. For the maximum expected occupancy level, we denote the expected occupancy level right before the first and second inspection times as

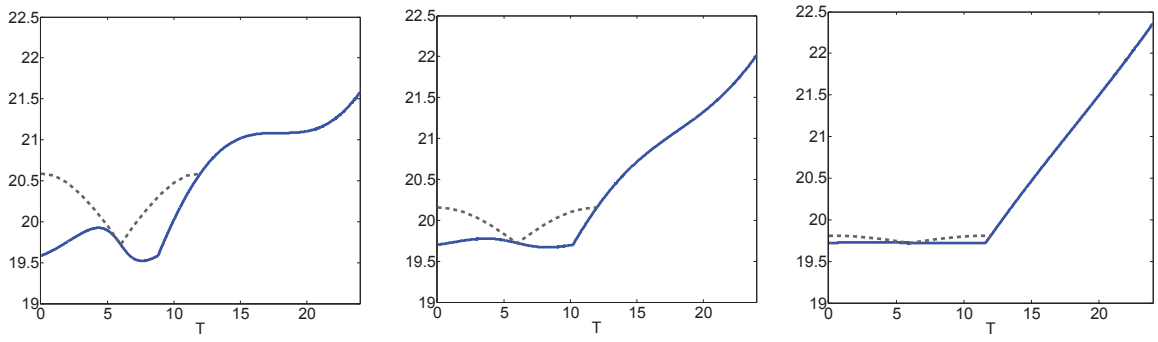
$$g_1(T,\Delta) := \frac{\bar{\lambda}}{\mu} + \bar{\lambda}(24-\Delta) - \frac{\beta}{\gamma}\cos(\gamma T) \text{ and } g_2(T,\Delta) := \frac{\bar{\lambda}}{\mu} + \bar{\lambda}\Delta - \frac{\beta}{\gamma}\cos(\gamma(T+\Delta)),$$

respectively. We observe that for fixed T, $g_1(T,\Delta)$ is decreasing in Δ and $g_2(T,\Delta)$ is increasing in Δ. We also observe that $g_1(T,0) > g_2(T,0)$ and $g_1(T,24) < g_2(T,24)$. Thus, for each T, there is a unique Δ, which is denoted as $\Delta(T^*)$, that solves $g_1(T,\Delta) = g_2(T,\Delta)$. Let $g^*(T) := \min_{0\le\Delta\le 24-T} \max\{g_1(T,\Delta), g_2(T,\Delta)\}$. Then

$$g^*(T) = \begin{cases} g_1(T,\Delta^*(T)) & \text{if } \Delta^*(T) \le 24-T; \\ g_1(T,24-T) = \bar{\lambda}T - \frac{\beta}{\gamma}\cos(\gamma T) & \text{if } \Delta^*(T) > 24-T. \end{cases}$$

The value of $\Delta^*(T)$ and $g^*(T)$ depends on the ratio between $\bar{\lambda}$ and β. Figure 13 plots $g^*(T)$ for different values of β. The dotted line represents the maximum expected occupancy level when inspections are equally spaced, $\eta_2(T)$. We observe that $\min_T g^*(T) - \eta_2(6)$ is in general fairly small and the difference decreases as β decreases.

Figure 13 $g^*(T)$ **(solid line) v.s.** $\eta_2(T)$ **(dotted line) for the infinite server model for different values of** β **(**$\bar{\lambda} = 0.2667$**,** $\mu = 1/75$**)**



(a) $\beta = \lambda$ (b) $\beta = 0.5\lambda$ (c) $\beta = 0.1\lambda$

Though the optimal inspection times may not be equally spaced, we are interested in understanding how the system performance is impacted by requiring equally spaced inspections. We do this via simulation. We use exponential service time distribution with rate $\mu = 1/75$, sinusoidal arrival rate function with $\bar{\lambda} =$

0.2665 and $s = 30$ servers. We use $\tilde{\eta}_2(T, \Delta)$ to denote the maximum mean occupancy level when there are two inspections at T and $T + \Delta$ per day. Table 1 summarizes our simulation results with the corresponding 95% confidence interval. We notice that the gaps in performance between the two policies are less than $.1\%$. We also see the difference decreases as β decreases. In fact, when $\beta = 0.1\lambda$, the difference is not even statistically significant at the 95% level. This robustness also translates to the maximum mean waiting time and maximum probability of waiting. Thus, it seems that requiring equally spaced inspections does not significantly impact system performance, while also having the benefit of being easy to convey to system administrators.

Table 1 The maximum mean occupancy level $\pm$ 95% confidence interval with two inspections per day

	Unevenly Spaced Inspections		Evenly Spaced Inspections
β	(T^*, Δ^*)	$\tilde{\eta}_2(T^*, \Delta^*(T^*))$	$\tilde{\eta}_2(6, 12)$
λ	$(7.61, 14.45)$	23.18 ± 0.03	23.41 ± 0.03
0.5λ	$(8.31, 13.32)$	23.31 ± 0.03	23.39 ± 0.04
0.1λ	$(8.73, 12.26)$	23.34 ± 0.03	23.36 ± 0.02

6.3. Empirical arrival rate function

Thus far, we have assumed a sinusoidal arrival rate function which has been amenable to deriving closed form expressions of interests. We now conduct a case study using the empirical arrival rate functions estimated from real hospital data. We take the scaled arrival rate function of all ED visits in the US during 2010 which is depicted in Figure 14(a) (Centers for Disease Control and Prevention 2010). We assume the service times are exponentially distributed with $\mu = 1/75$.

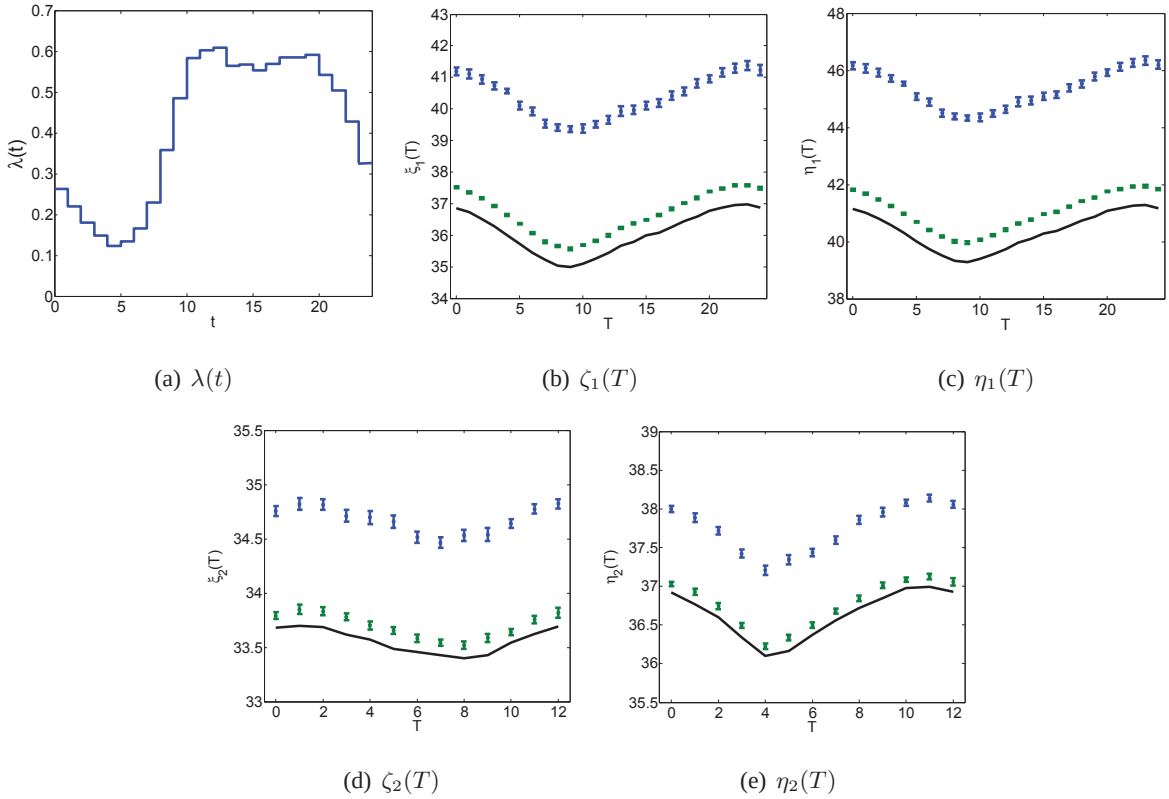
We plot the time-averaged and maximum expected occupancy level for different inspection policies and staffing levels in Figure 14. We use a relatively high effective traffic intensity. (Note that we find that the results are very similar for lower traffic intensities). Specifically, $\Lambda(24)/(40(1 - e^{-24\mu})) = 0.91$ and $\Lambda/(45(1 - e^{-24\mu})) = 0.81$. Notice that in this case, the arrival rate function no longer takes a sinusoidal form. If we assume there is a single peak of the arrival rate function (take the middle point of the two small peaks), then the peak is around $15 - 16$. We make the following observations from this case study:

a) The infinite server model still serves as a good approximation to determine optimal inspection time (i.e. the curves take the same shape).

b) For the time-averaged expected occupancy level, when there is a single inspection per day, the optimal inspection time is at 9, which is about $6 - 7$ hours before the peak arrival–as was the case for sinusoidal arrival rates. When there are two equally spaced inspections per day, although $\zeta_2(T)$ is changing over T, the difference for different inspection policies is less than 0.5 ($< 1\%$ of $\zeta_2(0)$), suggesting there is still a certain level of robustness to the inspection schedule when there are $N > 1$ inspections.

c) For the maximum expected occupancy level, when there is a single inspection per day, $\eta_2(T)$ follows the same shape as $\zeta_1(T)$ with the optimal inspection time at 9; when there are two equally spaced inspections per day, the optimal inspection times are at 4 and $4+12=16$. In particular, the first inspection is around $11-12$ hours before the peak and the second one is $0-1$ hour after the peak. Note that, qualitatively, this is essentially the same as the case of sinusoidal arrival rate, where the optimal inspection times take place 12 hours before the peak and right at the peak.

Figure 14 **Time-averaged and maximum expected occupancy level with fitted arrival rate function (upper: $s=40$, middle: $s=45$, lower: $s=\infty$)**



(a) $\lambda(t)$ (b) $\zeta_1(T)$ (c) $\eta_1(T)$

(d) $\zeta_2(T)$ (e) $\eta_2(T)$

7. The $M_t/M(\mathbf{T})/s/s$ Model

In this section, we consider a system where new arrivals are blocked (and sent elsewhere) if there is no available server (bed) upon arrival. Recent evidence suggests that when the ICU is busy, patients who are waiting for ICU admission may be rerouted to lower levels of care (Kim et al. 2015). Allon et al. (2013) found that when inpatient units (including the ICU) were busy, patients were more likely to be sent to other hospitals via ambulance diversion. In this section, we analyze properties of an $M_t/M(\mathbf{T})/s/s$ model, which can also be referred to as a loss model. To differentiate it from the $M_t/M(\mathbf{T})/s$ model, we denote the queue length process of the loss model as $\{Q^L(t): t\in\mathbb{R}\}$.

Following the same line of analysis as in Theorem 2, we have the following theorem characterizing the periodic equilibrium distribution of the loss system when there is one discharge per day, which takes place at τ_n on day n. We denote π^L as the stationary distribution of $Q_n^L := Q^L(\tau_n)$ for $n \in \mathbb{Z}$, and P_∞ as the periodic equilibrium distribution of $\{Q^L(t) : t \in \mathbb{R}\}$.

Theorem 4 *If $Q_n^L \sim \pi^L$, then $Q^L(\tau_n + t)$ is periodic in distribution with period equal to 24. Specifically, for $0 < t < 24$ and $m < s$,*

$$P_\infty(Q^L(\tau_n + t) = m) = \sum_{l=0}^{m} \pi^L(l) P(l + A(\tau_n, \tau_n + t) = m),$$

and

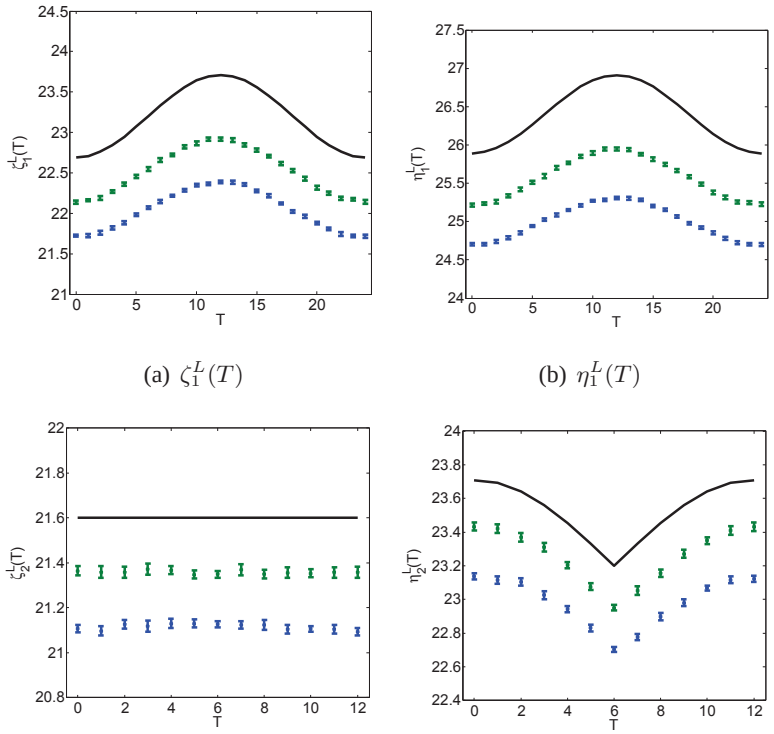
$$P_\infty(Q^L(\tau_n + t) = s) = \sum_{l=0}^{s} \pi^L(l) P(l + A(\tau_n, \tau_n + t) \geq s).$$

We observe from Theorem 4 that the probability of blocking, $P_\infty(Q^L(\tau_n + t) = s)$ is increasing in t for $0 < t < 24$. The proof of this result is practically identical to that of Theorem 2, so is omitted.

Similar to the $M_t/M(\mathbf{T})/s$ model case, the corresponding $M_t/M(\mathbf{T})/\infty$ model still serves as a reasonable approximation of the number in system dynamics of the loss model for different discharge policies. Moreover, as was the case with the $M_t/M(\mathbf{T})/s$ model, the infinite server model is a reasonable approximation for *general* service time distributions. Figure 15 plots the average occupancy level, $\zeta_N^L(T)$, and the maximum expected occupancy level, $\eta_N^L(T)$, $N = 1, 2$, for different values of T. We also plot the two measures of the corresponding infinite server queue model in the same figure as a solid line. We observe that both $\zeta_N^L(T)$ and $\eta_N^L(T)$ of the loss model take the same shape as the infinite server model, but the actual values are smaller than the corresponding infinite server model. Additionally, the higher the staffing level s, the closer the values are to those of the infinite server queue.

We observe that the *shape* of the occupancy level of an $M_t/M(\mathbf{T})/\infty$ infinite server model is very similar to that of the $M_t/M(\mathbf{T})/s$ and $M_t/M(\mathbf{T})/s/s$ models. When there are s servers and an infinite queue, the infinite server model underestimates the occupancy level, as customers (patients) have to wait in the $M_t/M(\mathbf{T})/s$ system. On the other hand, when there is no queue, the infinite server model overestimates the occupancy level, as customers (patients) who arrive to a busy system are lost in an $M_t/M(\mathbf{T})/s/s$ system, resulting in a lower effective arrival rate. In the hospital setting, many facilities have policies under which there is a maximum number of patients who may wait in the ED for admission to a specific unit such as the ICU, while others may either be placed in another unit or rerouted to another hospital via ambulance diversion (e.g. Allon et al. (2013)). This suggests that in these situations the infinite server model may be an even more accurate approximation to the occupancy level as its behavior seems to be more aligned with an $M_t/M(\mathbf{T})/s/(s+k)$ system. Figure 16 compares the time-average expected occupancy level of some $M_t/M(\mathbf{T})/s/(s+k)$ systems to the corresponding infinite server system when there is one discharge

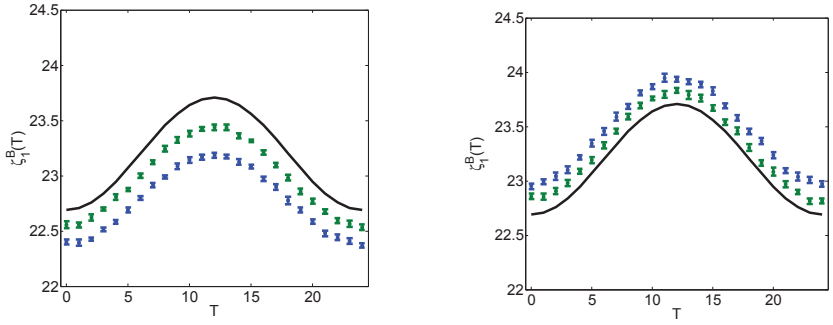
Figure 15 $M_t/M(\mathbf{T})/s/s$ **queueing system: The time-averaged and maximum expected occupancy level as a function of the discharge time** T **for one and two discharges per day (**$\lambda = 0.2667$**,** $\mu = 1/75$**, upper:** $s = \infty$**, middle:** $s = 32$**, lower** $s = 30$**)**



(a) $\zeta_1^L(T)$ (b) $\eta_1^L(T)$

(c) $\zeta_2^L(T)$ (d) $\eta_2^L(T)$

Figure 16 $M_t/M(\mathbf{T})/s/s+k$ **queueing system: The time-average expected occupancy level as a function of the discharge time** T **for different staffing levels,** s**, and waiting room sizes,** k **(**$\lambda = 0.2667$**,** $\mu = 1/75$**)**



(a) $k = 3$ (upper: $s = \infty$, middle: $s = 32$, lower: $s = 30$)

(b) $k = 9$ (upper: $s = 30$, middle: $s = 32$, lower: $s = \infty$)

per day. We observe that the actual performance curve may be above or below the infinite server curve depending on the value of k.

We also investigate the weighted average probability of blocking, $\bar{V}_B(N,T)$, and the stationary maximum

probability of blocking, $V_{B_m}(N,T)$, when there are N equally spaced inspections per day and the first inspection happens at time T. Specifically, the two performance measures are defined as follows:

$$\bar{V}_B(N,T) := \lim_{t\to\infty} \frac{1}{\Lambda(t)} \int_0^t 1\{Q^L(u) = s\}\lambda(u)du$$

and

$$V_{B_m}(N,T) := \max_k \{P_\infty(Q^L(T + k\Delta_N)) = s\}.$$

There are two approximations of the probability of blocking commonly used for the $M_t/G/s/s$ model (see Massey (2002) and references therein): the modified offer load (MOL) and the pointwise stationary (PS) approximation. When the time scale of variability of the arrival rate function is of the same scale or much smaller than the average service time, the MOL approximation is superior to the PS approximation. As this is the scenario which we consider, we focus on the MOL approximation. The approximated blocking probability is then given by the Erlang-C formula, which is the steady state probability of blocking for an $M/G/s/s$ model, but with the traffic intensity replaced by the mean queue length of the corresponding infinite server model, i.e.

$$P(Q^L(t) = s) \approx \frac{E[Q_\infty(t)]^s/s!}{\sum_{k=0}^s (E[Q_\infty(t)]^k/k!)},$$

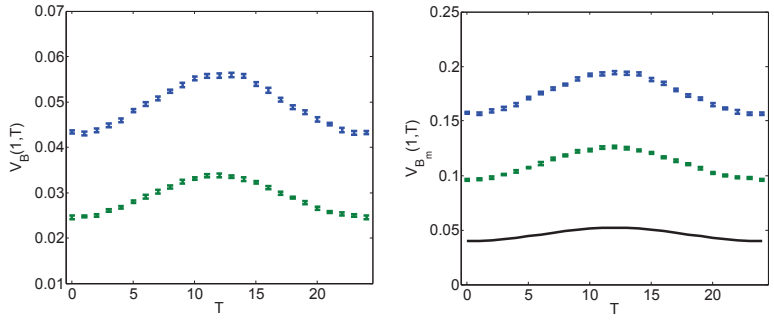
where $Q_\infty(t)$ is the queue length process of the corresponding infinite server queue. As the MOL approximation is monotonically increasing in $E[Q_\infty(t)]$, we test this approximation for the stationary maximum probability of blocking in our model. Note that we do not consider using the MOL approximation for the stationary average probability of blocking since this would involve integrating the Erlang-C formula multiplied by the arrival rate function, for which we are unable to derive closed-form solutions.

Figure 17 plots $\bar{V}_B(N,T)$ and $V_{B_m}(N,T)$ for different values of T and $N = 1, 2$. We also plot the MOL approximation for the stationary maximum probability of blocking in the same figure as a solid line. We observe that these two performance measures take the same shape as the corresponding queue length process. The MOL approximation is smaller than the actual stationary maximum probability of blocking. This is because of the unique dynamics introduced by the inspection feature of our model. Unlike a traditional loss model where the queue length process can fluctuate around level s (total number of servers), once the $M_t/M(\mathbf{T})/s/s$ model reaches capacity s, it remains there until the next inspection time. As such, the resulting probability of blocking is much larger than what would be seen in an $M_t/M/s/s$ model.

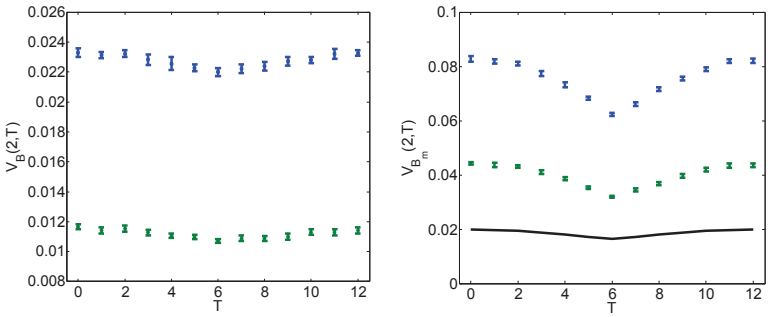
8. Conclusions and Further Research

Motivated by the need for physician examinations to discharge patients from hospital beds, we introduce a multi-server queueing system where a customer can only be released from the server when an inspection occurs and the customer has completed service prior to the inspection time. We analyze how the number and timing of the inspections affects system performance. Our analysis incorporates the feature that the arrival

Figure 17 $M_t/M(\mathbf{T})/s/s$ **queueing system: The average and maximum probability of blocking as a function of the discharge time** T **for one and two discharges per day (**$\lambda = 0.2667$**,** $\mu = 1/75$**, upper:** $s = \infty$**, middle:** $s = 32$**, lower** $s = 30$**)**



(a) $\bar{V}_B(1,T)$ (b) $V_{B_m}(1,T)$



(c) $\bar{V}_B(2,T)$ (d) $V_{B_m}(2,T)$

rate is time varying and the service time is of a much larger scale than the time scale of variability in the arrival rate. These are features commonly observed in the healthcare setting.

We introduce the concept of effective service rate based on stability analysis of the system and find that when having N inspections per day, evenly spaced inspections maximize the effective service rate. The effective service rate is increasing in N but the increments are decreasing in N. Given the other demands on physician time, it is neither practical, nor desirable, to allow for frequent discharges throughout the day. Fortunately, our analyses indicate that the most substantial gains are achieved by increasing from one to two inspection times per day.

We also characterize the periodic equilibrium of the time dependent performance measures. Because the number of servers is fixed while the arrival rate is time-varying and because the service time is much longer than the period of the arrival rate, traditional steady state approximations fail in our setting. Another challenge is that the size of the system can be very small, so we cannot rely on heavy-traffic approximations, which tend to be more accurate for very large systems. As such, we leverage the tractability of an infinite server model where the optimal inspection times can be characterized in closed form. We also show through numerical experiments that the insights from the infinite server model carry over to the finite server case. In

particular, we observe that adding a second inspection time not only improves system performance but also adds a lot of robustness to the performance with respect to the actual timing of the inspections. With two or more evenly spaced inspection times, the performance of the system is practically invariant to the precise schedule of inspection. We also consider model extensions of general service times and a loss model. Again, we see that the insights from the infinite server model also carry over to these cases.

In this work, we focus on performance measures related to the queue length process. These measures are very important to hospital administrators who are concerned with timely access to care for patients in order to improve clinical outcomes, and the cost-effective use of resources such as beds and nurses (e.g. Renaud et al. (2009), Rincon et al. (2010)). Specifically, we find that our infinite server model provides insight into how one can simultaneously minimize the average occupancy level and the probability of waiting. In fact, such a goal is likely highly attractive to hospital administrators. Typically, measures of delayed admission by the medical community have a binary notion: delayed versus not delayed (e.g. Chalfin et al. (2007), Renaud et al. (2009)). Moreover, minimizing the average occupancy level will ensure timely access to care for patients and could potentially allow for reductions in unit sizes (i.e. fewer beds and staff) thereby reducing operating costs. In our analysis of these system performance metrics, we have seen that precisely quantifying the expected waiting time in our setting is quite challenging. An interesting area of future research would be to extend the time-varying version of Little's Law to our queueing system.

From our analysis, we can ascertain a number of specific insights for practice: 1) Our analysis suggests that with a single inspection time, it *should take place* 6 hours before the peak in arrivals. In a hospital the peak arrival typically occurs in the afternoon, likely between 1-4pm. Thus, it is likely that the current scheduling of rounds, which typically takes place in the early morning, say at about 7am, is very good in terms of minimizing the average and maximum occupancy levels. 2) We find that it is possible to schedule inspections in order to simultaneously minimize occupancy and probability of waiting (or blocking); however, minimizing the expected waiting times will likely require different inspection times. 3) When adding additional inspection times, there are many benefits (e.g. increased stability region and robustness) to scheduling them at evenly spaced intervals throughout the day. Moreover, we find that while it might be worth considering adding one additional inspection, additional inspections have decreasing marginal gains and are probably not worthwhile.

In the ICU setting, administrators are quite concerned about patients who are ready to be discharged but are still occupying beds because they have not yet been examined by a physician who can issue the discharge order. This 'boarding' in the ICU may have strong implications as to the availability of beds for new ICU patients waiting to be admitted. Alternatively, patients who are ready to be discharged from the ICU may remain in ICU beds even after a physician's approval for discharge because there are no beds available in downstream units. Such a phenomenon also occurs in other hospital units. As such, extending our analysis

to a network perspective in order to gain insights on the impact of each unit's inspection/discharge policies would be an interesting area for further exploration.

There are number of extensions of our model and analysis which would be interesting to explore as future research. For instance, one could consider the impact of a (possibly random) delay between inspection time and the freeing of a server to capture the tasks necessary to complete between the time a doctor places a discharge order and the time the bed is ready for a new patient. Additionally, one could consider the duration of time necessary to complete an inspection.

Another potentially interesting area for future research would examine capacity decisions in light of the dynamics that arise from the need for inspections. Simple numerical explorations suggest that standard square-root staffing rule concepts (e.g. Kolesar and Green (1998)) to determine the number of beds needed to satisfy desired performance benchmarks do not immediately translate to this setting.

In a healthcare setting, patients are highly heterogenous with different care needs and priorities. Another potentially interesting future direction would be to incorporate priorities into a queueing system with inspections. It would be interesting to see how the optimal inspection times depend on patient mix and how different priority rules could impact congestion.

Finally, from a theoretical standpoint, it could be interesting to consider dynamic inspection schedules. For instance, one could consider a policy which has a baseline inspection only once a day but schedules an additional inspection every time the queue length exceeds a given threshold. It would be interesting to understand how various system primitives impact that optimal threshold. Of course, such dynamic policies may be difficult to implement in practice, but analysis of these policies could provide insight into what is potentially lost by requiring predictable and consistent inspection schedules.

Acknowledgments

The authors would like to thank the anonymous reviewers and associate editor, and the area editor for their valuable feedback and suggestions. The research by Carri W. Chan was partially supported by a National Science Foundation CAREER award (CMMI-1350059).

References

Allon, G., S. Deo, W. Lin. 2013. The impact of hospital size and occupancy of hospital on the extent of ambulance diversion: Theory and evidence. *Operations Research* **61** 554–562.

Armony, M., S. Israelit, A. Mandelbaum, Y.N. Marmor, Y. Tseytlin, G.B. Yom-tov. 2015. Patient flow in hospitals: A data-based queueing-science perspective. *Stochastic Systems* **5**(1) 146–194.

Bertsimas, D., G. Mourtzinou. 1997. Trasient laws of non-stationary queueing systems and their applications. *Queueing Systems* **25** 115–155.

Centers for Disease Control and Prevention. 2010. National hospital ambulatory medical care survey (accessed 7/21/2015). URL `ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/dataset_documentation/nhamcs/stata/`.

Chalfin, D. B., S. Trzeciak, A. Likourezos, B. M. Baumann, R. P. Dellinger. 2007. Impact of delayed transfer of critically ill patients from the emergency department to the intensive care unit. *Critical Care Medicine* **35** 1477–1483.

Chan, C. W., V. F. Farias, N. Bambos, G. Escobar. 2012. Optimizing ICU Discharge Decisions with Patient Readmissions. *Operations Research* **60** 1323–1342.

Chan, C. W., G. Yom-Tov, G. Escobar. 2014. When to use speedup: An examination of service systems with returns. *Operations Research* **62** 462–482.

Dai, J.G., P. Shi. 2014. A two-time-scale approach to time-varying queues for hospital inpatient flow management. Working paper.

Dobson, G., H.-H. Lee, E. Pinker. 2010. A Model of ICU Bumping. *Operations Research* **58** 1564–1576.

Doshi, B.T. 1986. Queueing systems with vacations - a survey. *Queueing Systems* **1** 29–66.

Eick, S.G., W.A. Massey, W. Whitt. 1993a. $M_t/G/\infty$ queues with sinusoidal arrival rate. *Management Science* **39**(2).

Eick, S.G., W.A. Massey, W. Whitt. 1993b. The physics of the $M_t/G/\infty$ queue. *Operations Research* **41**(4).

Federgruen, A., L. Green. 1986. Queueing systems with service interruptions. *Operations Research* **34**(5) 752–768.

Foster, F. G., K. M. Nyunt. 1961. Queues with Batch Departures I. *The Annals of Mathematical Statistics* **32**(4) 1324–1332.

Foster, F. G., A. G. A. D. Perera. 1964. Queues with Batch Departures II. *The Annals of Mathematical Statistics* **35**(3) 1147–1156.

Green, L., P. Kolesar, A. Svoronos. 1991. Some effects of nonstationarity on multiserver Markovian queueing systems. *Operations Research* **39** 502–511.

Green, L.V., P.J. Kolesar. 1991. The pointwise stationary approximation for queues with nonstationary arrivals. *Management Science* **37** 84–97.

Green, L.V., P.J. Kolesar, W. Whitt. 2007. Coping with time-varying demand when setting staffing requirements for a service system. *Production and Operations Management* **16**(1) 13–39.

Green, L.V., J. Soares, J.F. Giglio, R.A. Green. 2006. Using queueing theory to increase the effectiveness of emregency department provider staffing. *Academic Emergency Medicine* **13** 61–68.

Heyman, D., W. Whitt. 1984. The asymptotic behavior of queues with time-varying arrival rates. *Journal of Applied Probability* **21**(1) 143–156.

Jennings, O.B., A. Mandelbaum, W.A. Massey, W. Whitt. 1996. Server staffing to meet time-varying demand. *Management Science* **42**(10) 1383–1394.

Jewkes, E.M. 1995. Optimal inspection effort and scheduling for a manufacturing process with repair. *European Journal of Operational Research* **85** 340–351.

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

Kim, S-H, C. W. Chan, M. Olivares, G. Escobar. 2015. ICU Admission Control: An Empirical Study of Capacity Allocation and its Implication on Patient Outcomes. *Management Science* **61** 19–38.

Kolesar, P. J., L. V. Green. 1998. Insights on service system design from a normal approximation to erlang's delay formula. *Production and Operations Management* **7**(3) 282–293.

Liu, Y., W. Whitt. 2012. Stabilizing customer abandonment in many-server queues with time-varying arrivals. *Operations Research* **60**(2) 1551–1564.

Massey, W.A. 2002. The analysis of queues with time-varying rates for telecommunication models. *Telecommunication Systems* **21** 173–204.

McManus, M.L., M. C. Long, A. Cooper, E. Litvak. 2004. Queuing theory accurately models the need for critical care resources. *Anesthesiology* **100** 1271–1276.

Meyn, S., R.L. Tweedie. 2009. *Markov Chains and Stochastic Stability*. 2nd ed. Cambridge University.

Ohnishi, M., H. Kawai, H. Mine. 1986. An optimal inspection and replacement policy for a deteriorating system. *Journal of Applied Probability* **23** 973–988.

Powell, Emilie S., Rahul K. Khare, Arjun K. Venkatesh, Ben D. Van Roo, James G. Adams, Gilles Reinhardt. 2012. The relationship between inpatient discharge timing and emergency department boarding. *The Journal of Emergency Medicine* **42**(2) 186 – 196. doi:http://dx.doi.org/10.1016/j.jemermed.2010.06.028. URL http://www.sciencedirect.com/science/article/pii/S0736467910006530.

Renaud, B., A. Santin, E. Coma, N. Camus, D. Van Pelt, J. Hayon, M. Gurgui, E. Roupie, J. Hervé, M.J. Fine, C. Brun-Buisson, J. Labarère. 2009. Association between timing of intensive care unit admission and outcomes for emergency department patients with community-acquired pneumonia. *Critical Care Medicine* **37**(11) 2867–2874.

Rincon, F., S.A. Mayer, J. Rivolta, J. Stillman, B. Boden-Albala, M.S V. Elkind, R. Marshall, J.Y. Chong. 2010. Impact of delayed transfer of critically ill stroke patients from the emergency department to the neuro-icu. *Neurocritical Care* **13** 75–81.

Saghafian, S., W. Hopp, M. Van Oyen, J. Desmond, S.L. Kronick. 2012. Patient Streaming as a Mechanism for Improving Responsiveness in Emergency Departments. *Operations Research* **60** 1080–1097.

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

Thorisson, H. 1985. On regenerative and ergodic properties of k-server queue with nonstationary Poisson arrivals. *Journal of Applied Probability* **22** 893–902.

Wolff, R.W. 1982. Poisson arrivals see time average. *Operations Research* **30**(2) 223–231.

Yankovic, N., L. Green. 2011. Identifying Good Nursing Levels: A Queuing Approach. *Operations Research* **59** 942–955.

Yao, D.D., S. Zheng. 1996. Coordinated production and inspection in a tandem system. *Queueing Systems* **24** 59–82.

Zazanis, M.A. 2004. Infinite server queues with synchronized departures driven by a single point process. *Queueing Systems* **48** 309–338.

Appendix A: Proof of stability conditions

PROOF: [Proof of Lemma 1] We start by recognizing that $\{Q_n\}$ is a Markov Chain on a countable state space $\mathcal{S} = \mathbb{Z}^+$. Moreover, it is irreducible.

Stability: We analyze the stability condition first. We will use the following Lyapunov function: $V(q) = |q|$. Let $\epsilon = s(1-e^{-24\mu}) - \Lambda(24)$. By assumption $s(1-e^{-24\mu}) > \Lambda(24))$; it immediately follows that $\epsilon > 0$.

Consider the case where the number of customers in the system, q, is more than the number of servers:$q > s$. Then, $D_n \sim Binomial(s, 1-e^{-24\mu})$.

$$E[V(Q_1)|Q_0 = q] - V(q) = E[A_n] - E[D_n] = \Lambda(24) - s(1-e^{-24\mu}) \leq -\epsilon.$$

On the other hand, when there is no queue, $q \leq s$:

$$E[V(Q_1)|Q_0 = q] - V(q) \leq E[A_n] = \Lambda(24) < \infty.$$

Thus, it satisfies the Foster-Lyapunov Criterion for a positive recurrent Markov process on countable state space (Meyn and Tweedie 2009).

Instability: We next prove the instability condition. Find $\theta > 0$ small enough such that

i) $(1-e^{-24\mu})(e^{\theta}-1) < 1$

ii) $\Lambda(24)(1-e^{-\theta}) > s(1-e^{-24\mu})(e^{\theta}-1)$

It is clear that such a θ exists, since $1-e^{-24\mu} < 1$ and $\Lambda(24) > s(1-e^{-24\mu})$. Let $\tilde{V}(q) = e^{-\theta q}$. Then we have when $q > s$,

$$\begin{aligned} E[\tilde{V}(Q_1)|Q_0 = q] &= \tilde{V}(q)\exp\left(\Lambda(24)(e^{-\theta}-1)\right)\exp\left\{s\log(1+(1-e^{-24\mu})(e^{\theta}-1))\right\} \\ &\leq \tilde{V}(q)\exp\left\{-(\Lambda(24)(1-e^{-\theta}) - s(1-e^{-24\mu})(e^{\theta}-1))\right\} \\ &< \tilde{V}(q) \end{aligned}$$

and

$$\inf_{q \leq s} \tilde{V}(q) = \exp(-\theta s) < \exp(-\theta(s+1)).$$

Thus it satisfies the Foster-Lyapunov Criterion for a transient Markov process on countable state space (Meyn and Tweedie 2009). □

PROOF: [Proof of Theorem 1] We provide a proof for the case $N = 2$. The proof for general N follows exactly the same line of analysis.

$\{Q_n\}$ is a Markov Chain on a countable state space $\mathcal{S} = \mathbb{Z}^+$, which is irreducible.

Stability: We analyze the stability condition first. Let $V(q) = |q|$. As $s(1-e^{-\Delta_2^{(1)}\mu}) + s(1-e^{-\Delta_2^{(2)}\mu}) > \Lambda(24)$ by assumption, we can find $\delta > 0$ small enough, such that $s(1-e^{-\Delta_2^{(1)}\mu}) + (1-\delta)s(1-e^{-\Delta_2^{(2)}\mu}) > \Lambda(24)$. Let $\epsilon = s(1-e^{-\Delta_2^{(1)}\mu}) + (1-\delta)s(1-e^{-\Delta_2^{(2)}\mu}) - \Lambda(24) > 0$. Find $\kappa > 0$ large enough such that $P(\bar{D}_n^{(1)} > \kappa) \leq \delta$. Set $K = \{q \in \mathbb{Z}^+ : q \leq s + \kappa\}$. When $q \notin K$,

$$E[V(Q_1)|Q_0 = q] - V(q) = E[A_1^{(1)} + A_1^{(2)}] - E[D_1^{(1)}] - E[D_1^{(2)}]$$

$$\begin{aligned}
&\leq \Lambda(24) - s(1-e^{-\Delta_1\mu}) - E[D_1^{(2)}|D_1^{(1)} < \kappa]P(D_1^{(1)} < \kappa) \\
&= \Lambda(24) - s(1-e^{-\Delta_1\mu}) - (1-\delta)s(1-e^{-\Delta_2\mu}) \\
&= -\epsilon;
\end{aligned}$$

when $q \in K$,

$$E[V(Q_1)|Q_0 = q] - V(q) \leq \Lambda(24) < \infty.$$

We thus verified the Foster-Lyapunov Criterion for positive recurrent Markov process on countable state space (Meyn and Tweedie 2009).

Instability: We next analyze the instability condition. Find $\theta > 0$ small enough such that

i) $(1-e^{-\Delta_2^{(1)}\mu})(e^\theta - 1) < 1$ and $(1-e^{-\Delta_2^{(2)}\mu})(e^\theta - 1) < 1$

ii) $\Lambda(24)(1-e^{-\theta}) > (s(1-e^{-\Delta_2^{(1)}\mu}) + s(1-e^{-\Delta_2^{(2)}\mu}))(e^\theta - 1)$

Let $\tilde{V}(q) = e^{-\theta q}$. Then we have when $q > s$,

$$\begin{aligned}
E[\tilde{V}(Q_1)|Q_0 = q] &= \tilde{V}(q)\exp\left(\Lambda(24)(e^{-\theta}-1)\right) E\left[\exp\left(\theta(D_1^{(1)} + D_1^{(2)})\right)\right] \\
&\leq \tilde{V}(q)\exp\left(\Lambda(24)(e^{-\theta}-1)\right) E\left[\exp\left(\theta\bar{D}_1^{(1)}\right)\right] E\left[\exp\left(\theta\bar{D}_1^{(2)}\right)\right] \\
&= \tilde{V}(q)\exp\{\Lambda(24)(e^{-\theta}-1) + s\log(1+(1-e^{-\Delta_2^{(1)}\mu})(e^\theta-1)) \\
&\quad + s\log(1+(1-e^{-\Delta_2^{(2)}\mu})(e^\theta-1))\} \\
&\leq \tilde{V}(q)\exp\{\Lambda(24)(e^{-\theta}-1) + (s(1-e^{-\Delta_2^{(1)}\mu}) + s(1-e^{-\Delta_2^{(2)}\mu}))(e^\theta-1)\} \\
&= \tilde{V}(q)\exp\left(-(\Lambda(24)(1-e^{-\theta}) - (s(1-e^{-\Delta_2^{(1)}\mu}) + s(1-e^{-\Delta_2^{(2)}\mu}))(e^\theta-1))\right) \\
&< \tilde{V}(q)
\end{aligned}$$

and

$$\inf_{q \leq s} \tilde{V}(q) = \exp(-\theta s) < \exp(-\theta(s+1))$$

We thus verified the Foster-Lyapunov Criterion for transient Markov process on countable state space (Meyn and Tweedie 2009). □

PROOF: [Proof of Lemma 2] Let $\delta_i = 24/N - \Delta_N^{(i)}$. Then $\delta_i < 24$ and $\delta_i\mu < 1$.

$$\begin{aligned}
\sum_{i=1}^N (1-e^{-\Delta_N^{(i)}\mu}) &= N - e^{-\mu 24/N}\sum_{i=1}^N \exp(\delta_i\mu) \\
&\leq N - e^{-\mu 24/N}\sum_{i=1}^N (1+\delta_i\mu) \\
&= N - Ne^{-\mu 24/N}
\end{aligned}$$

When $\delta_i = 0$, $\sum_{i=1}^N (1-e^{-\Delta_N^{(i)}\mu}) = N - Ne^{-\mu 24/N}$. Thus, $\sum_{i=1}^N (1-e^{-\Delta_N^{(i)}\mu})$ attains its maximum when $\Delta_N^{(i)} = 24/N$ for $i = 1, 2, \cdots, N$. □

Appendix B: Proof of periodic equilibrium results

PROOF: [Proof of Theorem 2] $Q(\tau_n+t)=Q(\tau_n)+A(\tau_n,\tau_n+t)$. $A(\tau_n,\tau_n+t)$ is independent of $Q(\tau_n)$ and have the same distribution as $A(\tau_{n+1},\tau_{n+1}+t)$ for $t\in(0,24)$. When $Q(\tau_n)\sim\pi$, $Q(\tau_{n+1})\sim\pi$ by the definition of stationarity. Then for any $m\in\mathbb{Z}^+$

$$\begin{aligned}
P(Q(\tau_n+t)=m) &= \sum_{l=0}^{m} P(Q(\tau_n)+A(\tau_n,\tau_n+t)=m|Q(\tau_n)=l)P(Q(\tau_n)=l)\\
&= \sum_{l=0}^{m} P(l+A(\tau_n,\tau_n+t)=m)\pi(l)\\
&= \sum_{l=0}^{m} P(Q(\tau_{n+1})+A(\tau_n,\tau_n+t)=m|Q(\tau_{n+1})=l)P(Q(\tau_{n+1})=l)\\
&= P(Q(\tau_{n+1}+t)=m)
\end{aligned}$$

The expression for $P_\infty(W(t)>24k)$, $k\in\mathbb{Z}^+$, follows directly from the simple relationship between $Q(t)$ and $W(t)$. □

PROOF: [Proof of Proposition 1] We first observe that if a customer arrived at time τ_n+t, for some inspection time τ_n and $0<t<24$, then, he has to wait at least until the next inspection time. In other words, if $W(\tau_n+t)>0$, then $W(\tau_n+t)\geq 24-t$. Likewise, if $W(\tau_n+t)>24k$, then $W(\tau_n+t)\geq 24k+24-t$.

If we assume the system starts empty from the infinite past then it would be in periodic equilibrium at t for $t\geq 0$ (see Thorisson (1985) for theoretical support for initializing non stationary models at $t=-\infty$). We also set time 0 as the beginning of a day. Taking the queue (customers waiting to be served) as the system, applying time varying Little's law (Bertsimas and Mourtzinou 1997), we have

$$\begin{aligned}
&E_\infty[(Q(\tau_n+t)-s)^+]\\
&= \int_{-\infty}^{\tau_n+t} P(W(u)>(\tau_n+t-u))\lambda(u)du\\
&= \int_{\tau_n}^{\tau_n+t} P(W(u)>\tau_n+t-u)\lambda(u)du+\sum_{k=1}^{\infty}\int_{\tau_n-24k}^{\tau_n-24(k-1)} P(W(u)>\tau_n+t-u)\lambda(u)du\\
&= \int_{\tau_n}^{\tau_n+t} P_\infty(W(u)>0)\lambda(u)du+\sum_{k=1}^{\infty}\int_{\tau_n}^{\tau_n+24} P_\infty(W(u)>24k)\lambda(u)du
\end{aligned}$$

For $E_\infty[W(\tau_n+t)]$, we have

$$\begin{aligned}
E_\infty[W(\tau_n+t)] &= \int_0^\infty P(W(\tau_n+t)>u)du\\
&= \int_0^{24-t} P(W(\tau_n+t)>u)du+\sum_{k=1}^{\infty}\int_{24k-t}^{24(k+1)-t} P(W(\tau_n+t)>u)du\\
&= \int_0^{24-t} P(W(\tau_n+t)>0)du+\sum_{k=1}^{\infty}\int_{24k-t}^{24(k+1)-t} P(W(\tau_n+t)>24k)du\\
&= P_\infty(W(\tau_n+t)>0)(24-t)+24\sum_{k=1}^{\infty} P_\infty(W(\tau_n+t)>24k).
\end{aligned}$$

□

PROOF: [Proof or Proposition 2] As $Q(t)$ and $W(t)$ are regenerative process and $\lambda(t)$ is periodic, $X(t)\lambda(t)$ is also a regenerative process. We define the regeneration set as $\{Q(\tau_n)=0\}$ (i.e. $\{Q_n=0\}$). We also define $\kappa:=\inf\{\tau_n>0:$

$Q(\tau_n)=0\}$ and $K:=\inf\{n>0:Q_n=0\}$. Without loss of generality, we assume $\tau_0=0$ and let $E_0[\cdot]:=E[\cdot|Q_0=0]$. Then we have

$$\begin{aligned}
\frac{1}{t}\int_0^t X(u)\lambda(u)du &\to \frac{E_0\left[\int_0^\kappa X(u)\lambda(u)du\right]}{E_0\left[\kappa\right]} \text{ as } t\to\infty \\
&= \frac{E_0\left[\sum_{n=0}^{K-1}\int_{\tau_n}^{\tau_{n+1}} X(u)\lambda(u)du\right]}{E_0\left[\kappa\right]} \\
&= \frac{E_0\left[\sum_{n=0}^{K-1} E\left[\int_{\tau_n}^{\tau_{n+1}} X(u)\lambda(u)du|Q_n\right]\right]}{24E_0[K]} \\
&= \frac{1}{24}E\left[\int_{\tau_0}^{\tau_1} X(u)\lambda(u)du|Q_0\in\pi\right] \\
&= \frac{1}{24}E_\infty\left[\int_{\tau_0}^{\tau_1} X(u)\lambda(u)du\right] \\
&= \frac{1}{24}\int_0^{24} E_\infty[X(u)]\lambda(u)du
\end{aligned}$$

As $\lambda(t)$ is a periodic function with period $T=24$, we also have

$$24\frac{\Lambda(t)}{t}\to\Lambda(24) \text{ as } t\to\infty.$$

Thus,

$$\frac{1}{\Lambda(t)}\int_0^t X(u)du=\frac{\int_0^{24}E_\infty[X(u)]\lambda(u)du}{\Lambda(24)}.$$

□

Appendix C: Proof of performance approximations of the infinite server queue model

PROOF: [Proof of Lemma 3]

$$\begin{aligned}
\zeta_N(T) &= \sum_{k=1}^N\left(\frac{\bar\lambda}{\mu}-\frac{\beta}{\gamma}\cos(\gamma(T+(k-1)\Delta_N))+\frac{1}{2}\int_{T+(k-1)\Delta_N}^{T+k\Delta_N}\lambda(t)dt\right)\frac{\int_{T+(k-1)\Delta_N}^{T+k\Delta_N}\lambda(t)dt}{\Lambda(24)} \\
&= \sum_{k=1}^N\left\{\left(\frac{\bar\lambda}{\mu}+\frac{1}{2}\bar\lambda\Delta_N-\frac{1}{2}\frac{\beta}{\gamma}\cos(\gamma(T+(k-1)\Delta_N))-\frac{1}{2}\frac{\beta}{\gamma}\cos(\gamma(T+k\Delta_N))\right)\right. \\
&\quad \left.\times\frac{\bar\lambda\Delta_N+\beta/\gamma\cos(\gamma(T+(k-1)\Delta_N))-\beta/\gamma\cos(\gamma(T+k\Delta_N))}{\Lambda(24)}\right\} \\
&= \frac{\bar\lambda}{\mu}+\frac{1}{2}\bar\lambda\Delta_N-\frac{\beta}{N\gamma}\sum_{k=1}^N\cos(\gamma(T+(k-1)\Delta_N))
\end{aligned}$$

When $N=1$, $\zeta_1(T)=\bar\lambda(1/\mu+12)-\beta/\gamma\cos(\gamma T)$. When $N\geq 2$, $\zeta_N(T)=\bar\lambda(1/\mu+\Delta_N/2)$. □

PROOF: [Proof of Lemma 4] $\eta_N^{(i)}(T)=\bar\lambda/\mu+\bar\lambda\Delta_N-\beta/\gamma\cos(\gamma(T+(i-1)\Delta_N))$. As $-\cos(\gamma x)$ is increasing on $[0,12)$ and decreasing on $[12,24)$, $\eta_N^{(i)}(T)$ attains its maximum at the closest inspection time point to 12.

Let $\Theta(T)=\min\{\arg\min_i|12-(T+(i-1)\Delta_N)|\}$. Then $T+(\Theta(T)-1)\Delta_N$ is the closest inspection time point to 12. If there are two points that are of the same distance to 12, $T+(\Theta(T)-1)\Delta_N$ is the smaller one. To minimize $\eta_N^{(i)}(T)$, we want to have $T+(\Theta(T)-1)\Delta_N$ as far from 12 as possible.

When N is even, $\min_T\eta_N(T)$ is achieved when the $N/2$-th inspection time and the $(N/2+1)$-th inspection time are symmetric around 12. That is when $T+(N/2-1)\Delta_N=12-\Delta_N/2$. Then $T=12-\Delta_N/2-(N/2-1)\Delta_N$.

When N is odd, $\min_T \eta_N(T)$ us achieved when $(N+1)/2$-th inspection time and the $((N+1)/2+1)$-th inspection time are symmetric around 12. That is when $T+((N+1)/2-1)\Delta_N = 12-\Delta_N/2$. Then $T = 12-\Delta_N/2-((N+1)/2-1)\Delta_N$.

In summary, $\arg\min_T \eta_N(T) = 12-\Delta_N/2-\lceil N/2-1\rceil\Delta_N$ and $\min_T \eta_N(T) = \bar{\lambda}/\mu+\bar{\lambda}\Delta_N-\beta/\gamma\cos(\pi-\pi/N)$.

□