Queueing Models for Patient-flow Dynamics in Inpatient Wards

Jing Dong
Columbia University, New York, NY 10027

Ohad Perry
Northwestern University, Evanston, IL 60208

Hospital-related queues have unique features that are not captured by standard queueing assumptions, necessitating the development of specialized models. In this paper we propose a queueing model that takes into account the most salient features of queues associated with patient-flow dynamics in inpatient wards, including the need for a physician's approval to discharge patients, and subsequent discharge delays. In this setting, fundamental quantities, such as the (effective) mean hospitalization time and the traffic intensity, become functions of the queueing model's primitives. We therefore begin by characterizing these quantities, and quantifying the impacts that the discharge policy has on the average bed utilization and maximal throughput. We then introduce a deterministic fluid model to approximate the non-stationary patient-flow dynamics. The fluid model is shown to possess a unique periodic equilibrium, which is guaranteed to be approached as time increases, so that long-run performance analysis can be carried out by simply considering that equilibrium cycle. Consequently, evaluating the effects of policy changes on system's performance, and optimizing long-run operating costs, are facilitated considerably. The effectiveness of the fluid model is demonstrated via comparisons to data from a large hospital and simulation experiments.

Key words: Patients-Flow, Discharge Delays, Multi-Server Queue with Blocking, Deterministic Fluid Approximations, Long-Run Periodicity

1. Introduction

Healthcare spending in the United States (US) is the highest in the world, with hospital-related expenditures taking the lion's share of the National Health Expenditure (NHE). Specifically, 17.9% of the US Gross Domestic Product (GDP)–$3.3 trillion, or $10,348 per capita–were spent on healthcare in 2016, with approximately 32% of that amount (over $1 trillion) going to hospital care (Centers for Medicare and Medicaid Services 2018). The NHE is projected to keep increasing in the coming years, with hospital spending kept at approximately 32% (Keehan et al. 2008). Hence,

even moderate improvements in hospital operations can have a substantial macroeconomic impact. Furthermore, under the new Pay-for-Performance (known as P4P) policy of the Affordable Care Act ("Obamacare"), Medicare payments to hospitals are linked to the quality of care that they provide, whose measurements incorporate waiting times in the Emergency Department (ED) and overall Length of Stay (LOS) (Medicare 2015), incentivizing hospital managements to improve their queueing-related quality of service. These facts, in addition to the increasing demand experienced by many hospitals, and the fact that resources cannot be easily increased, lead hospital managements to seek policy changes to improve patients-flow among the different hospital units in an effort to reduce related congestions and delays.

To this end, queueing models are incorporated into toolkits that are dedicated to help hospitals improve related operations; see, e.g., Banner Health (2015). It is significant that such toolkits build on existing queueing models (such as the $M/M/c$ queue), despite the fact that patient-flow dynamics within the hospital possess unique features *that are not captured by those existing models*. For example, hospitalized patients do not leave their beds immediately upon their "service" completion (i.e., when they are recovered), and tend to occupy their beds for longer times than necessary – a phenomenon known as *bed blocking*. In this paper, we propose a new queueing model for Inpatient Wards (IWs), taking into account the unique features associated with patient-flow dynamics.

1.1. A "Copernican" Queueing Network

From the patient-flow perspective, a hospital can be viewed as a complex queueing network with several interconnected service stations, some of which have exogenous arrivals of patients (e.g., practically all the arrivals to the ED), while others get almost all of their patients from other units. Our focus here is on the IWs which, in a large hospital, consist of a large number of beds pooled together (Armony et al. 2015, Shi et al. 2016). We thus consider a single large service pool (referred to as "the IW"), in which beds play the role of "servers", bed requests are considered as "arrivals", and the queue consists of patients waiting, in other units of the hospital, for a bed in the IW.

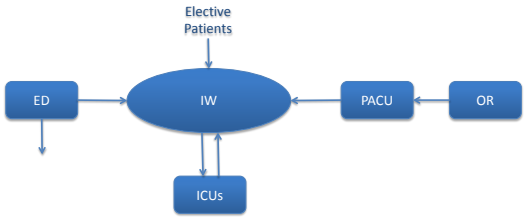


Figure 1 upstream units centered about the IW

Our focus on the IW is due to the following reasons: First, there is strong evidence that IW admission delays lead to an increase in LOS, higher mortality rates (Kc and Terwiesch 2009, Richardson 2002), and increased probabilities for readmissions (Trzeciak and Rivers 2003). Second, congestion in the IW tends to propagate to "upstream" units, such as the ED, certain Intensive-Care Units (ICU), Post-Anesthesia Care Units (PACU), and eventually, even to the Operating Rooms (OR); see Hall et al. (2006), Argo et al. (2009) and McGowan et al. (2007). Specifically, lack of available beds in the IW has been identified as a main cause for prolonged ED boarding[1], which in turn lead to ED overcrowding (Asplin and Magid 2007, Trzeciak and Rivers 2003). Thus, it is natural to focus on the IW in an effort to improve its operations, and consider it as the central unit of the hospital from the queueing perspective, which is our approach here. A schematic representation of the IW-centered network is depicted in Figure 1. See also Figures 1 and 2 in Armony et al. (2015) for a detailed depiction of the specific hospitals they study.

1.2. Main Characteristics of IW Queues

Since patient-flow dynamics are highly complex, an effective model must be relatively parsimonious, "distilling" only the most salient characteristics of the related queueing processes. We now elaborate on the three main features that are incorporated in our model.

1. Periodic Discharge Decisions. Unlike a typical queueing system, in which customers depart the server immediately when their service terminates, a recovered patient does not leave the bed until a physician examines and approves her release. In large hospitals, most patients are examined during daily inspection rounds that take place in the morning hours. Thus, patients may occupy their beds for relatively long periods of time *after their recovery*, while waiting to be examined by a physician.

2. Discharge Delays. The departures of a patients who were judged to be recovered is further delayed due to several reasons (e.g, paperwork, need for transportation arrangements, coaching by professionals, etc.), and occur in batches, several hours after the morning inspection round has ended; e.g., §4.2 in Armony et al. (2015). These delays lead to a relatively low variability of the discharge times, as most departures tend to be concentrated at a narrow time interval that is highly predictable. For example, in the hospital studied in Armony et al. (2015), most departures occur between 3pm and 4pm; a similar phenomenon is observed in the Singaporean hospital studied in Shi et al. (2016); see Figure 9 in this latter reference.

3. Non-stationarity. Like most service systems in practice, the arrival rate (of bed requests) to the IW is time-varying, implying that it is a non-stationary system (namely, it does not possess a steady state). However, unlike many other service systems for which stationary analysis can be effective despite having time-dependent arrival rates, the time-dependent queuing dynamics corresponding to patient-flow must be incorporated into a queueing model for the IW. In particular, stationary analysis is reasonable when the arrival rates are nearly fixed over sufficiently-long time intervals, so that the queue can approach a (local) steady state that is associated with those fixed rates. It is significant that such local steady-state analysis hinges on the assumption that changes in the arrival rate are relatively small during time periods that are at least several average service-times long (Gans et al. 2003, §3.2). This latter assumption cannot be made in the IW setting, since the arrival rate of bed-request changes considerably throughout the day (e.g., Figure 1 in Dai and Shi (2017)), while the average hospitalization period is several-days long.

1.3. Contribution

We make the following key contributions:

(I) We propose a simple queueing model based on the main features of the IW patient-flow dynamics described in §1.2 above. We show that the effective service-time of a patient depends on her arrival time (see Remark 2 below), so that a "service rate" does not exist. Nevertheless, we characterize the *maximum effective service rate* as an explicit function of the model primitives,

and in particular, of bed blocking that is due to the special discharge process. In turn, this allows us to quantify the maximal throughput of the IW, namely, the maximum number of arrivals per day that the IW can handle without making changes to its treatment or discharge policies. Even though our system is not stationary, we prove that it converges to a *periodic steady state* provided the average number of daily arrivals is smaller than the maximum throughput.

(II) Since the stochastic dynamics of our model are intractable, we develop an analytically-tractable fluid model to approximate those complex patient-flow dynamics. We then prove that, just like the underlying stochastic system, the fluid approximation possesses a unique periodic steady state to which it converges for all initial conditions. For the case $D = 0$, namely, when there is no discharge delay, we characterize the convergence rate of the fluid model to its periodic steady state. The accuracy of the fluid model as an approximation to the stochastic system is demonstrated via (i) simulation experiments; (ii) by proving that it holds as a many-server heavy-traffic fluid limit; and more importantly, (iii) by comparing the fluid predictions to real hospital data.

(III) The convergence of the queueing system and the corresponding fluid approximation to a periodic steady state simplifies long-run analysis considerably, because long-run computations reduce to the analysis of a single period (the stationary one); see Theorem 3 below. We demonstrate how this "stationary-type" analysis facilitates the computations of key performance measures, as well as optimization of those measures. In particular, one can easily compute the impact of changing the time of the inspection round or the discharge delay on the queue length, waiting times, or, e.g., the proportion of patients that wait more than a given time period for an IW bed.

(IV) The robustness of our proposed model is demonstrated by considering extensions to the basic model and comparing their fluid models. In the first extension, we consider a multiclass model having several patient types, each with a different service time distribution. In the second extension, we assume that the discharge occurs during a time interval according to a given density function (that would be estimated from data in practice). Numerical examples demonstrate that simple adjustments to the basic model can be applied to capture these extensions.

1.4. Related Literature

IW-related Queueing Models. The literature concerning hospital queues is quite large, with much attention given to the ED, operating rooms and ICUs, see for example Song et al. (2015), Chan et al. (2012), Green and Savin (2008); here we focus on IW-related queueing models.

Best et al. (2015) consider an optimal partitioning of the hospital beds to wings, and employ the Erlang-A model to approximate the fraction of patients that abandon when waiting for a bed. The interaction between the ED and the hospital wards is studied in Mandelbaum et al. (2012), where the hospital is modeled as an inverted V model, and the "randomized most-idle" policy is proposed so as to ensure servers fairness (measured in bed idleness ratios) in an asymptotic diffusion limit for a stationary system. The interaction between the ED and the IWs is also studied in Ramakrishnan et al. (2005), which propose a two-time-scale model (in particular, the ED is considered to operate in a faster time scale than the wards) to study the impacts of IW operations on the ED.

The most closely related works to ours are Chan et al. (2017) and the series of papers Dai and Shi (2017), Shi et al. (2016), Dai and Shi (2018). In Shi et al. (2016), the authors build a detailed simulation model to study the effects of operational policies in the IW on ED boarding times. In the subsequence paper Dai and Shi (2017), a two time scale model for the IW is proposed. In that model, patients' LOS is comprised of the number of days, measured by the number of midnights, and of a "fast" intra-day time scale, which is the additional time spent by the patient in the IW on the day of admission and the day of discharge. The assumption is that a patient is medically ready to leave the bed at the midnight hour of the day at which she is discharged. The dynamics of the two time-scale model can be analyzed exactly, but the computational complexity renders such analysis impractical. Therefore, a diffusion approximation, based on Stein's method, is developed and is shown to be effective and accurate for large systems with long service times. Based on the two time scales method, Dai and Shi (2018) employs approximate dynamic programming to develop good overflow policies between different wards in a hospital. The approximation step involves a value function obtained from a fluid control problem.

In contrast to the two time-scales approach in the latter papers, we assume that the recovery process of patients is a continuous process, e.g., a patient that is not ready to be discharged in the morning may be physically recovered in the afternoon of the same day, as is indeed the case in reality. In fact, in small units, physicians may add a second inspection round in the afternoon in order to discharge more patients[2]. On the other hand, unlike the analytical model in Dai and Shi (2017), our model cannot be analyzed exactly, and even diffusion approximations seem to be prohibitively hard to develop. Hence, we focus on approximating the *predictable variability* of the system by considering a fluid approximation for the patient-flow dynamics. Data shows that our model can accurately capture the average queueing dynamics of heavily loaded IWs (see §7 below).

In Chan et al. (2017) a time-varying queueing model with periodic inspection rounds, but *no discharge delays*, is analyzed. The main goal is to study the impacts of adding more inspection rounds on the system's performance, as well as to characterize the optimal times to perform those inspection rounds. Chan et al. (2017) consider hospital units which are relatively lightly-loaded (e.g., ICUs), and thus employs highly-tractable infinite-server approximations for their analysis. We, on the other hand, focus on heavily-loaded IWs that have little or no extra bed capacity. Furthermore, unlike Chan et al. (2017), which focus on small units for which it is feasible to add inspection rounds, more than one inspection round is unlikely to take place in large IWs due to the number of people that are involved and the time that those inspection rounds consume. Thus, our model, goals and analyses are different than those in Chan et al. (2017). In particular, we employ a fluid model to approximate the system's dynamics and key performance measures. That fluid model, which is also a fluid limit (functional weak law of large numbers) that is achieved under the many-server heavy-traffic regime, can be used to optimize operations and staffing (bed capacity) under waiting-time constraints.

Periodic Queues. Fluid and diffusion limits to approximate periodic time-dependent behavior of many-server queues were considered in Heyman and Whitt (1984), Liu and Whitt (2011a) and Puhalskii (2013). A queueing network in Perry and Whitt (2016) is shown to have a fluid limit

with two equilibria: a fixed point and a periodic equilibrium, and either one can be the limit (as time increases) for a given fluid trajectory, depending on the parameters of the problem and the initial condition. Here we show that a unique periodic equilibrium exists for our model, and that equilibrium is also the limit point of all fluid trajectories.

Related Service Models. Queueing systems with batch departures, motivated by hospitals' outpatient departments, were considered as early as the mid 1950's (Bailey 1954). There is very limited work on queues with added delays to the service process. In Maglaras et al. (2013), delay is artificially "injected" in order to differentiate between distinct classes of delay-sensitive customers; see also Afeche (2013). In §4 of Pang and Perry (2014), delay in the queue (but not in service) is forced on inbound calls to a contact center that serves both inbound and outbound call so as to guarantee that given service-level constraints are met. Unlike the settings in these latter three papers, in our work, discharge delays are a negative phenomenon that should be minimized.

1.5. Organization

The rest of the paper is organized as follows. We describe the queueing model in §2 and characterize stability condition in §3. To approximate the queueing dynamics, we introduce a fluid model in §4, and prove important qualitative results for the fluid model in §5. We then show how to use these qualitative results for long-run average performance analysis in §6. Empirical evidence for the effectiveness of the proposed fluid model is shown in §7. In §8 we demonstrate the robustness of the basic fluid model by numerical and simulation comparisons to fluid approximations for more general queueing models. We conclude in §9.

In §§A and B we provide more details on the stability of the fluid model, as well as its increased accuracy as the system's size grows, respectively. The proofs of all the results in the paper appear in §C.

2. The Model

We now specify the queueing model for the IW-centered network. In our model, the IW (which can also be two or more IWs pooled together) is considered to be a service station, its beds are

considered as the servers, and the patients waiting for an IW bed are queued in the "satellite" units, which collectively serve as the buffer of the queue. (The impacts of the queue on upstream units is not modeled directly.) More specifically, we consider the IW as a multi-server queue with c servers (beds), where c is relatively large (at least several tens).

2.1. The Primitive Processes

Arrival Process. We assume that arrivals constitute a nonhomogeneous Poisson process having a strictly positive and periodic arrival-rate function $\lambda(t)$, $t \geq 0$; see Armony et al. (2015), Kim et al. (2015) and Kim and Whitt (2014) for the validity of this assumption. We measure time in terms of the period length, namely, we have that $\lambda(t+1) = \lambda(t)$ for all $t \geq 0$, where one unit of time (the "period") represents one day in our paper, although other units of time can be considered.

Service Process. For tractability, in our basic model we assume that patients are statistically homogeneous, so that their service times are IID random variables. (We consider a multiclass generalization in §8 below.) Specifically, we assume that the service time of each patient is exponentially distributed with mean $1/\mu$, independently of all other patients. Here, "service time" of a patient can be thought of as the actual recovery time (the time it takes that patient to be medically ready to be discharged) but it is not the total time that the patient spends in the IW (her LOS). We note that, in practice, μ is a censored variable, because recovered patients are only identified when inspected. In §7 we demonstrate how our model can be employed to identify μ from data.

Discharge Process. As was mentioned above, most discharge decisions are made during the inspection round that takes place once a day, typically in the morning hours. We assume that there are further delays between the inspection round and the actual departure of the patients. Since we are not interested in the inspection process itself, but rather in its outcome, namely, the number of patients that will be discharged that day, we assume that the inspection time takes place at a specific moment of the day, which can be taken to be the typical time at which the inspection round ends. Similarly, since the actual departures of the patients are highly concentrated within a relatively short time window, we consider the discharge time to take place at a single time point

as well. This latter assumption is crucial in order for our fluid approximation to capture the fact that departures occur in large batches; see §4 below.

An important observation is that the departure mechanism just described implies that the LOS of patients are not IID, even though patients' recovery times are IID. The non-IID property of the LOS was observed in data and identified as a key component in models for patient-flow dynamics in Shi et al. (2016).

REMARK 1. The assumptions that patients are statistically homogeneous, discharges occur in one batch, and service times are exponentially distributed are relaxed in §8, where extensions to the basic model described above are considered. In particular, in §8 we demonstrate that the basic model can be easily adjusted to describe a multi-class IW (having several patient types), and that little is lost by the assumption that the discharge occurs at a single time point. Furthermore, we employ simulation experiments to verify that our models are robust to the distributional assumptions.

2.2. Process Dynamics and Notation

For $k = 0, 1, \ldots$, let T_k denote the kth inspection time, where $T_k := k + T$, i.e., the inspections occur every day at the same time T. We assume that patients that are in the IW at the beginning of the kth day, and finished their service *before* time T_k, depart at time $T_k + D$, where $D \geq 0$ and $T_k + D < k + 1$ ($T + D < 1$) for all $k \geq 0$. We refer to D as the *discharge delay*, and consider it to be a constant that is independent of k. Note that patients who finish their "service", namely, are recovered, during $(T_k, T_{k+1}]$ will stay in the IW until the next discharge occurs at time $T_{k+1} + D$.

Let $Q(t)$ denote the number of patients waiting to be admitted to the IW at time t; $Z(t)$ be the number of patients who are actively receiving service in the IW (i.e., are in a recovery process); and $B(t)$ be the number of patients who have recovered before t but have not departed yet. Alternatively, $B(t)$ is the number of "blocked beds" at time t. Given that there are c beds in the IW, the number of available (idle) beds at time t is $c - Z(t) - B(t)$, so that

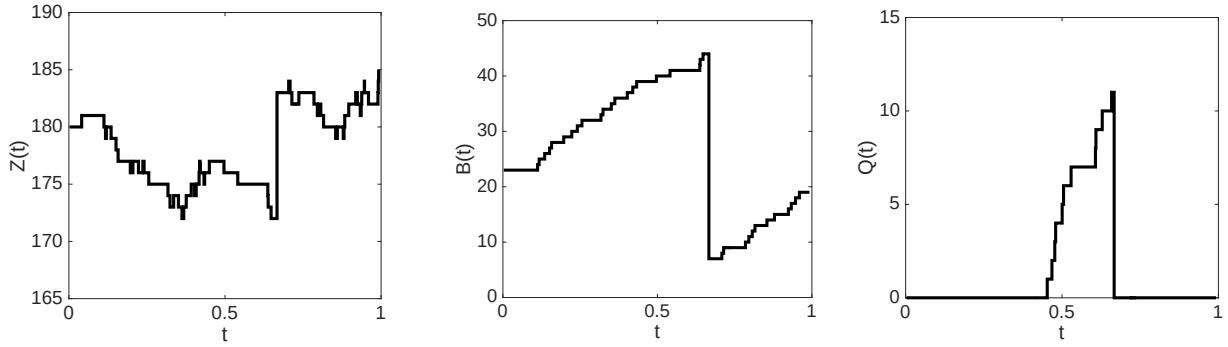
$$X(t) = (Q(t), Z(t), B(t)), \quad t \geq 0, \tag{1}$$

is a stochastic process describing the evolution of the system.

For each $k \geq 0$, assume that all the recovered patients by the inspection time T_k are identified and will be discharged at the discharge time the same day. Then at time $T_k + D$ we have that $B(T_k + D) = B((T_k + D)-) - B(T_k)$, $k \geq 0$, where $f(t-)$ denotes the left limit of a function f at time t. (As usual, all stochastic processes are assumed to be right continuous with limits from the left everywhere.) In particular, the number of blocked beds at the discharge time is equal to the number of blocked beds immediately before the discharge less the number of beds that were blocked at the inspection round. Similarly, the process $Z(t)$ will jump up at time $T_k + D$ due to the new patients entering the IW: if $B(T_k) > Q(T_k + D)$, then $Z(T_k + D) = Z((T_k + D)-) + Q((T_k + D)-)$, and otherwise $Z(T_k + D) = Z((T_k + D)-) + B(T_k)$. Finally, the size of the downward jump of the queue process $Q(t)$ is equal to the size of the upward jump of $Z(t)$ at $T_k + D$.

Notice that the evolution of $X(t)$ is completely determined by the arrival process, mean service time, number of beds, typical time of the day at which the inspection round ends, and the average delay until departures occur. In particular, the basic data set $(\lambda(\cdot), \mu, c, T, D)$ fully determines the queueing dynamics of our IW model. For illustrative purposes, we plot a sample path of $X(t)$ for an IW with 216 beds in Figure 2 over one period (day). In this example, we take $1/\mu = 4.5$ days, $T = 10$am and $D = 6$ hours. The arrival-rate function we consider is a properly-scaled version of the arrivals to all the ED's in the US during 2010 (the data is taken from Centers for Disease Control and Prevention (2010)). We fit a piecewise constant arrival rate function to the data with a constant rate per hour. We then multiply the arrival rate function by a proper constant to make the daily arrival rate $\Lambda = 38$). Observe that the three component processes of $X(t)$ have many small jumps of unit size throughout the day–corresponding to arrivals of new patients and service completion of patients–and one large jump at the discharge epoch (4pm in this example). Note that in this example, we have $\Lambda/(c\mu) = 0.8$. One may therefore expect to have 80% average bed utilization, with short queues and waiting time. However, queues form on a daily basis, and waiting times for those patients that wait can be several-hours long. Our analysis in the next section shows that the *effective traffic intensity* to the IW in this example is much larger than 0.8, and is approximately equal to 0.93; see ρ_e in (3) below.

Figure 2 A simulated sample path of $X(t)$



(a) $Z(t)$ (b) $B(t)$ (c) $Q(t)$

3. Stability and Maximum Throughput

We begin by characterizing a necessary and sufficient condition for the stability of $X(t)$ in (1). In turn, this allows us to characterize the maximum throughput of the IW, namely, the maximum daily arrival rate that the IW can handle without making changes to its policies. To this end, we first observe that the continuous-time process $X(t)$ in (1) is not Markovian when $D > 0$ because the size of the jump at time $T_k + D$ is determined at time T_k, for all $k \geq 0$. Nevertheless, we can embed a Discrete-Time Markov Chain (DTMC) in this continuous-time process by considering $X(t)$ immediately after inspections epochs. In particular, it is easy to see that $X_k := X(T_k)$, $k \geq 0$, is a DTMC.

Let Λ denote the total input rate into the system during a period,

$$\Lambda := \int_0^1 \lambda(s)ds = \int_{T_k}^{T_{k+1}} \lambda(s)ds, \quad k \geq 0, \tag{2}$$

where the second equality in (2) follows from the periodicity of the function $\lambda(t)$. Define

$$\rho_e := \frac{\Lambda}{c\beta}, \quad \text{for} \quad \beta := \frac{(1 - e^{-\mu})}{1 - e^{-\mu} + e^{-\mu(1-D)}}. \tag{3}$$

Due to Theorem 1 below, we interpret ρ_e as the *effective traffic intensity* to the system, so that $c\beta$ is the *maximum throughput* of the IW. It follows that β can be thought of as the *maximal effective service rate* per bed, taking into account the long-run proportion of time that it is blocked. See also Remark 2 below for an explanation regarding the values of β and ρ_e.

THEOREM 1. *The DTMC* $\{X_k : k \geq 0\}$ *is ergodic, so that* $X(t)$ *is stable, if and only if* $\rho_e < 1$.

We note that $X(t)$ is null recurrent if $\rho_e = 1$ and transient if $\rho_e > 1$; see the proof of Theorem 1.

Discussion. Theorem 1 demonstrates the dramatic effect that the special structure of the service and departure processes has on the IW's throughput. We first make the observation that β is decreasing in D and that ρ_e is increasing in D, so that decreasing (increasing) D has the effect of increasing (decreasing) the effective service rate. We also observe that

$$c\beta := c\frac{1-e^{-\mu}}{1-e^{-\mu}+e^{-\mu(1-D)}} < c(1-e^{-\mu}) < c\mu.$$

To interpret these inequalities, note that when there is no delay, i.e., when $D = 0$, the maximum throughput is $cP(S < 1) = c(1 - e^{-\mu})$, where S is a generic random variable representing the *nominal* service time of a patient. Thus, the first inequality quantifies how much service capacity is lost due to the discharge delay D. The second inequality measures how much service capacity is lost due to the fact that patients depart only after their discharge is approved by a physician during the inspection round. Indeed, $c\mu$ would have been the maximum throughput if patients were to leave immediately upon finishing their service (i.e., if there was no inspection round and no discharge delays); see Heyman and Whitt (1984).

For a concrete numerical example, consider an IW with an expected service time of 5 days and discharge delay of 5 hours. Then $\beta = 0.175$ and $1 - e^{-\mu} = 0.18$, whereas $\mu = 0.2$. In particular, there is 10% loss of service capacity when $D = 0$ and 12.5% loss of service capacity when $D = 5$ hours, relative to a "standard" model (in which patients leave immediately upon their service completion) having the same arrival process and service-rate μ. Note that incorporating the delays into the "standard" model, namely, taking the expected service time to be $1/\mu + D$, gives service rate $(1/\mu + D)^{-1} \approx \mu$, since the delay is small relative to the average service time. In this example, $(1/\mu + D)^{-1} = 0.192$. Thus, simply adding D to the service time of a "standard model" does not capture the impact of D on the effective service rate.

REMARK 2 (EXPLAINING β). The value of β can be understood once we observe that an **overloaded system** (with $\rho_e > 1$) reduces to an overloaded $M_p(t)/G/c$ model with a discrete service-time distribution $\lceil S + D \rceil$, where S is exponentially distributed with mean $1/\mu$, and $\lceil x \rceil$ is the smallest integer larger than x. To see this, observe that if the system is overloaded, so that its queue never empties, patients enter the IW only at the discharge time epochs $T_k + D$. We can shift the time axis such that $T_k = k$, $k \geq 0$, so that a patient that enters service at a time $k + D$ is identified as recovered at an inspection time $\lceil (k+D) + S \rceil$ and thus leaves the IW at time $\lceil (k+D) + S \rceil + D$. Overall, such a patient occupies an IW bed for a duration $\lceil k + S + D \rceil + D - (k + D) = \lceil S + D \rceil$. Therefore, the expected service time of a patient *in an overloaded system* is

$$E\left[\lceil S + D \rceil\right] = (1 - e^{-\mu}) + \sum_{n=1}^{\infty} (n+1)\left(e^{-\mu(n-D)} - e^{-\mu(n+1-D)}\right) = 1/\beta.$$

Indeed, the stability condition of our system (stated in Theorem 1) and that $M_p(t)/G/c$ model is the same. Note, however, that in a non-overloaded system, a nonnegligible proportion of the patients will enter the IW upon arrival, and that the time that a patient entering at time $t \neq T_k + D$ spends in the IW is distributed as $\lceil t + S \rceil + D - t$ (which is almost surely not equal to $\lceil S \rceil + D$ for all $t > 0$). Thus, even if patients are statistically homogeneous, the actual times they spend occupying the beds depend on their arrival times, and are therefore not identically distributed.

4. The Fluid Approximation

It follows from Theorem 1 that there exists a unique stationary distribution for the DTMC $X_k := \{X_k : k \geq 0\}$ which is also the limiting one as $k \to \infty$. In turn, this implies that $X(t)$ possesses a unique *periodic steady state* (PSS), as in Liu and Whitt (2011a,b), Dai and Shi (2017)) ,with period 1. That is, if X_0 is distributed according to the stationary distribution of the DTMC X_k, then $X(t) \stackrel{\mathrm{d}}{=} X(t+1)$ for all $t \geq 0$, where $\stackrel{\mathrm{d}}{=}$ denotes equality in distribution. Unfortunately, computing the stationary distribution of X_k numerically, and evaluating the PSS of $X(t)$, is impractical due to the complexity of the transition matrix of X_k. Therefore, we develop a fluid approximation for the dynamics of $X(t)$, and use its periodic equilibrium to approximate the PSS of $X(t)$.

To this end, we regard the patients and servers as continuous quantities, and assume that arrivals and service completions (recoveries) occur continuously and deterministically at rates $\lambda(t)$ and μ, respectively. We "replace" the discrete stochastic dynamics of the process $X(t)$ with a deterministic dynamical system described by a function $x(t) = \{x(t) : t \geq 0\}$, which is assumed to be almost-everywhere (a.e.) differentiable. We use lower-case letters to denote the fluid counterparts of each of the stochastic components of $X(t)$, e.g., $q(t)$ denotes the quantity of patient-fluid "waiting" in queue to be processed, so that $x(t) := (q(t), z(t), b(t))$, $t \geq 0$. The fluid process $x(t)$ takes values in $[0, \infty) \times [0, c]^2$, where c is a positive real number.

We assume that $\lambda(t)$ is piecewise continuous, namely, that it has at most a finite number of discontinuities over any finite time interval. This assumption, which is clearly not restrictive in practice, supports the fluid analysis by ensuring that the differential equation describing $x(t)$ given in (4)-(5) below, has a unique solution; see the proof of Proposition 1.

We divide the analysis into cycles that are determined by inspection times, i.e., the kth cycle is the time interval $[T_k, T_{k+1})$. Then for any $x_0 \in [0, \infty) \times [0, c]^2$ and all *regular* points $t \geq 0$, where a time point t is regular if $x(t)$ is differentiable at t (in particular, $t \neq T_k + D$, $k \geq 0$), $x(t)$ is the (unique) solution to the following Initial-Value Problem (IVP): $x(0) = x_0$, and

$$\begin{aligned} dq(t) &= 1_{\{z(t)+b(t)=c\}} \lambda(t) dt; \\ dz(t) &= 1_{\{z(t)+b(t)<c\}} \lambda(t) dt - z(t)\mu dt; \\ db(t) &= z(t)\mu dt, \end{aligned} \tag{4}$$

and at the jump epoch $T_k + D$, $k \geq 0$,

$$\begin{aligned} q(T_k + D) &= (q((T_k + D)-) - b(T_k))^+; \\ z(T_k + D) &= z((T_k + D)-) + q((T_k + D)-) \wedge b(T_k); \\ b(T_k + D) &= b((T_k + D)-) - b(T_k), \end{aligned} \tag{5}$$

where $a^+ := \max\{a, 0\}$, $a \wedge b := \min\{a, b\}$, for $a, b \in \mathbb{R}$, and 1_A denotes the indicator function that is equal to 1 on the set A, and to 0 otherwise.

To understand how the differential equations in (4) for the fluid model is obtained, note that $z(t)$ increases at time t at an instantaneous rate $\lambda(t)$ only if beds are available at that time t (hence the indicator function for the event $\{z(t)+b(t)<c\}$), due to new arrivals that receive idle beds, and is decreasing at an instantaneous rate $\mu z(t)$ at any time $t \geq 0$, due to service completions (which do not lead to departures from the IW). At the jump epoch $T_k + D$, $z(T_k + D)$ increases by the queue content immediately before that time, if sufficient capacity becomes available due to the discharge, and otherwise, $z(T_k + D)$ increases by $b(T_k)$, which is the amount of fluid that departs the system at $T_k + D$. Similar reasonings lead to the equations of $q(t)$ and $b(t)$.

For $t \in [T_k + D, T_{k+1} + D)$, let $\Psi(t, x, b_k)$ denote the Right-Hand Side (RHS) of (4), so that the differential equation (4) can be compactly presented via $\dot{x}(t) = \Psi(t, x(t), b_k)$ over the kth cycle. Denote the jump size at the kth cycle by J_k, i.e., $J_k := x(T_k + D) - x((T_k + D)-)$. The following proposition shows that the dynamics described in (4)-(5) determine a well-defined process. Let $x_k := x(T_k)$, $k \geq 0$.

PROPOSITION 1. *For all $x_0 \in [0,\infty) \times [0,c]^2$ and $0 < u < \infty$ there exists a unique solution to the IVP*

$$\dot{x}(t) = \Psi(t, x(t), b_k) \quad \textit{with } x(0) = x_0, \quad \textit{and} \quad J_k = (-(q_k \wedge b_k), q_k \wedge b_k, -b_k), \quad t \in [0, u]. \tag{6}$$

We remark that the dynamics of $x(t)$ over $[T_k + D, T_{k+1})$ depend on $b_k = b(T_k)$, rendering the IVP a non-Markov deterministic process; in particular, it is not an Ordinary Differential Equation (ODE).

We next show that the fluid model should approximate large IW's well by proving that it is also the fluid limit of the stochastic model in the many-server limiting regime. In particular, the fluid approximation $x(t)$ is achieved via a Functional Weak Law of Large Numbers (FWLLN) for a sequence of properly-scaled stochastic systems when the number of beds and the arrival volume increase without bound. To this end, consider a sequence of stochastic models described in §2,

indexed by n, $X^n = (Z^n, B^n, Q^n)$, $n \geq 1$. We assume that the arrival rate function $\lambda^n(t)$ and the number of servers c^n in the nth system satisfy

$$\lim_{n\to\infty} \lambda^n(t)/n = \lambda(t) \quad \text{and} \quad \lim_{n\to\infty} c^n/n = c, \tag{7}$$

where the limit for $\lambda^n(t)/n$ holds uniformly in t over bounded intervals. We use $\bar{X}^n := X^n/n$, $n \geq 1$, to denote the *fluid-scaled* sequence. We note that, with a slight abuse of notation, we use the same notation c, which we have used earlier to denote the (integer) number of beds in an IW, for the limit in (7) (which can take any positive real value), since in practice, the fluid must be applied with an integer c.

PROPOSITION 2 **(FWLLN)**. *Assume that* (7) *holds such that* $\lambda(t)$ *is almost everywhere continuous. If* $\bar{X}^n(0) \Rightarrow x(0)$ *in* $\mathbb{R}^3$ *as* $n \to \infty$, *for some deterministic element* $x(0) \in \mathbb{R}^3$, *then* $\bar{X}^n \Rightarrow x$ *uniformly over compact intervals as* $n \to \infty$, *where* x *is the unique solution to the IVP* (6) *with initial condition* $x(0)$.

It follows from Proposition 2 that the fluid model $x(t)$ approximates the mean dynamics of $X(t)$, provided the IW is sufficiently large. In particular, the dynamics of $X^n(t)$ are approximated by $nx(t)$, $t \geq 0$, plus random noise which is negligible with respect to n.

5. Periodic Stationarity

A clear difficulty of working with the fluid model is that it can only be solved numerically for finite time intervals, making it difficult to optimize. Fortunately, long-run analysis is simplified because, just as the underlying stochastic system $X(t)$, the fluid approximation $x(t)$ possesses a unique periodic steady state, or a "periodic equilibrium", to which $x(t)$ converges as $t \to \infty$, regardless of its initial condition. Formally,

DEFINITION 1. A solution $u^* := \{u^*(t) : t \geq 0\}$ to the IVP (6) is a periodic equilibrium if there exists $p > 0$, such that $u^*(p+t) = u^*(t)$ for all $t \geq 0$. The smallest such p is called the period.

We first show that the fluid approximation for a stable stochastic system possesses a unique periodic equilibrium. Henceforth, we append all the parameters associated with the periodic equilibrium with a superscript $*$.

THEOREM 2. (*periodic stationarity*) *If* $\rho_e < 1$, *then there exists a unique periodic equilibrium* $x^*(t) := (q^*(t), z^*(t), b^*(t))$, $t \geq 0$. *Furthermore,* $z^*(T+D) + b^*(T+D) < c$, *so that* $q^*(T+D) = 0$.

Unlike the PSS of the stochastic model for the IW, the unique periodic equilibrium of its fluid approximation can be easily computed; see the proof of Theorem 2 and Remark 5 in §C below.

We will say that the fluid model is *asymptotically periodic* if it converges to the (unique) periodic equilibrium as time increases, regardless of the initial condition. That is, the fluid model is *asymptotically periodic* if, for any initial condition $x(0)$,

$$\|x(t) - x^*(t)\| \to 0 \quad \text{as} \quad t \to \infty, \tag{8}$$

where $\|\cdot\|$ denotes the usual Euclidean norm in $\mathbb{R}^3$.

We note that, in general, asymptotically-periodic dynamical systems with equilibrium u^* do not converge to the equilibrium as in (8) since there is typically a time shift between the trajectory of any solution $u \neq u^*$ and the trajectory of u^*. Hence, convergence to a periodic equilibrium is usually defined to hold in the image space of the solutions (with time suppressed). Nevertheless, in our case the jump epochs are time- (as opposed to state-) dependent, ensuring that, if a solution $x(t)$ converges to a periodic equilibrium $x^*(t)$ as $t \to \infty$, then the convergence is of the form in (8).

THEOREM 3. *If* $\rho_e < 1$, *then the fluid model is asymptotically periodic. As a result,*

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t f(x(s))ds = \int_0^1 f(x^*(s))ds, \tag{9}$$

for any function $f : \mathbb{R}_3 \to \mathbb{R}$ *that is a.e. continuous.*

Our next result shows that, if the system is large enough, the queue in the stochastic system is likely to empty together with the fluid queue, so that $Q^n(T_k + D) = 0$ with a high probability in large systems, once the fluid model is sufficiently close to its equilibrium. To state this result formally, consider the sequence of stochastic systems as in Proposition 2, and let

$$\tau_0^n := \inf\{k \geq 0 : Q^n(T_k + D) = 0\} \quad \text{and for } i \geq 1, \quad \tau_i^n := \inf\{k \geq \tau_{i-1}^n : Q^n(T_k + D) = 0\}.$$

COROLLARY 1. *If $\rho_e < 1$, then $P(\tau^n_{k+1} - \tau^n_k = 1) \to 1$ as $n \to \infty$, for all $k \geq K$, for some $K \geq 0$.*

We also note that when $\rho_e = 1$, it can be shown (see Theorem 5 below) that there exist infinitely-many periodic equilibria, and that the behavior of the fluid model depends on the initial condition. Furthermore, the fluid model may never reach a state with zero queue. Since by Proposition 2 $Q^n(t) \approx nq(t)$ for large n, this implies that the queue in the underlying stochastic system can be arbitrarily large.

5.1. Rate of Convergence to Equilibrium: Systems with No Delays

In practice, the arrival-rate pattern necessarily changes over time, and it is therefore important to evaluate how long it takes the process to approach its equilibrium. Even though we are unable to quantify the rate of convergence in the general case, we can establish results for the special case $D = 0$, in which $X(t)$ is a Markov process and the dynamics of its fluid model $x(t)$ are governed by an ODE. The case $D = 0$ can be thought of as a perturbation of the original system (with $D > 0$) for the following reason.

Let $x(\cdot; D)$ denote the solution to the IVP (6) when the delay is D and let d_{J_1} denote the Skorohod J_1 metric; see Whitt (2002). Fix D_0 satisfying $0 < D_0 < 1 - T$ (so that $T_k + D_0 < k + 1$, $k \geq 1$). Then it is easy to show that $x(t; D)$ is continuous in D for all $D < D_0$ over all the intervals of the form $[0, T_k + D_0) \cup (T_k + D_0, k + 1]$, $k \geq 0$, and converges uniformly in t over those intervals. It then follows that both the locations of the jumps and their magnitudes converge to the respective values when $D = 0$, so that $d_{J_1}(x(t; D), x(t; 0)) \to 0$ as $D \downarrow 0$. See (Whitt 2002, p.79).

Since the case $D = 0$ is considerably simpler than the general case, we obtain an explicit representation for the periodic equilibrium. To derive it, consider a *stable* system having no discharge delays ($D = 0$), and let $w(t)$ denote the total amount of fluid in the system at time t. Then $w(t)$ is monotonically increasing on (T_k, T_{k+1}) and has a jump of size $b(T_k-)$ at the inspection-and-discharge time T_k. Let $w^* := w^*(T_k)$ denote the total fluid in the system immediately after discharge *in equilibrium*, and define Δ_0 to be the unique solution to the following equation.

$$w^* + \int_0^{\Delta_0} \lambda(s) ds = c. \tag{10}$$

By Corollary 1, $w^* \le c$, so that Δ_0 is well-defined. If $\Delta_0 > 1$, then the queue is empty throughout the equilibrium cycle, whereas if $\Delta_0 < 1$, the system experiences both an underload period (in which there is no queue) and an overload period (in which the queue is positive) during the equilibrium cycle. Equating the inflow to the outflow during a full cycle gives the balance equation

$$\Lambda = w^*(1-e^{-\mu}) + \int_0^{\Delta_0 \wedge 1} \lambda(s)(1-e^{-\mu(1-s)})ds. \tag{11}$$

Indeed, Λ is the total fluid input to the system during any cycle, while the output in a given period in equilibrium consists of all the fluid that recovered during the previous period (with all periods being equivalent in equilibrium), giving us the first term in the RHS of (11), and all the arriving fluid during the current period which recover by the end of this period. That arriving fluid is captured by the second term in the RHS of (11), because a proportion $1-e^{-\mu(1-s)}$ of the fluid that enters service at time $s \in [0,1)$ is served (recovers) by time 1. Note that, if $\Delta_0 < 1$, then newly arriving fluid enters service only up to time Δ_0, which is why we integrate up to $\Delta_0 \wedge 1$.

We can similarly equate inflow to outflow at any time t in equilibrium, to derive the equations

$$\begin{aligned} z(T_k+t) &= w^* e^{-\mu t} + \int_0^{\Delta_0 \wedge t} \lambda(s) e^{-\mu(1-s)} ds; \\ b(T_k+t) &= w^*(1-e^{-\mu t}) + \int_0^{\Delta_0 \wedge t} \lambda(s)\left(1-e^{-\mu(1-s)}\right) ds; \\ q(T_k+t) &= \int_{\Delta_0}^{t} \lambda(s) ds, \quad 0 \le t < 1, \end{aligned} \tag{12}$$

where $z_k^* := z^*(T_k) = w^*$, $q_k^* := q^*(T_k) = 0$, and the values of w^* and Δ_0 are computed from (10) and (11).

Let $w_k := z_k + q_k$ (since there is no delay, $b_k = 0$). Note that $q_k = (w_k - c)^+$ and $z_k = w_k \wedge c$, so that w_k determines z_k and q_k uniquely. Hence, it is sufficient to prove the asymptotic periodicity and rate of convergence for the process $\{w_k : k \ge 0\}$.

THEOREM 4. *Assume that $D = 0$ and $\rho_e < 1$ (equivalently, $\Lambda < c(1-e^{-\mu})$). Then the unique periodic equilibrium u_0^* is the solution to* (10) - (12)*. Moreover,*

(i) If $w_0 \le c$, then

$$|w_k - w^*| < |w_0 - w^*|(1-\alpha(w_0))^k, \tag{13}$$

for some constant $\alpha(w_0) \in (0,1)$*, which depends only on the initial condition* w_0*.*

(ii) If $w_0 > c$*, then for* $K_c := \min\{k \geq 1 : w_k \leq c\}$*, it holds that*

$$|w_k - w_0| = |w_k - w^*| + \left(\Lambda - c(1 - e^{-\mu})\right) k, \quad \textit{for all } k \leq K_c,$$

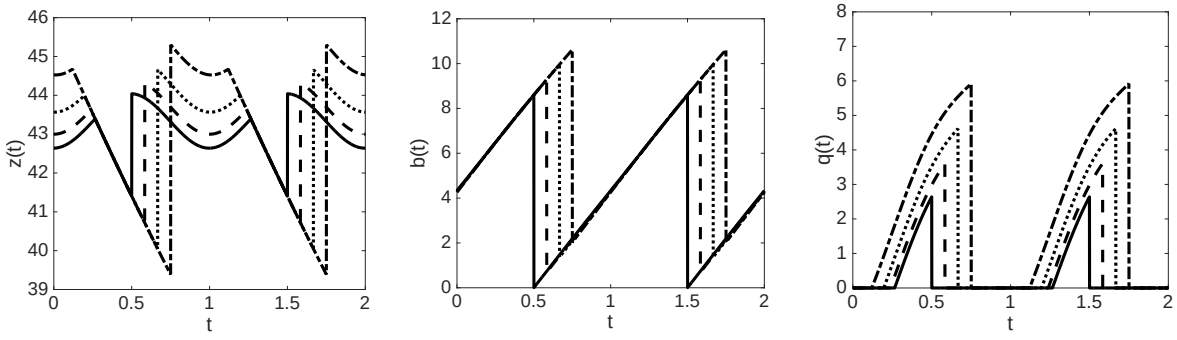
and (13) *holds for all* $k > K_c$*, with* w_{K_c} *replacing* w_0 *in* (13)*.*

Theorem 4 demonstrates that, similarly to the fluid approximation for the $M/M/c$ queue, $x(t)$ converges to x^* geometrically fast if there is no queue initially, while if the initial queue is positive, the convergence rate is linear until the first time the queue empties, and then switches to geometric.

5.2. Numerical Example

We now present a numerical example, comparing the fluid equilibrium for different values of D. We consider a sinusoidal arrival-rate function $\lambda(t) = 8.6 + 4.3\sin(2\pi t)$, and set the service rate to $\mu = 1/5$ (i.e., the average service time is 5 days). The inspections take place at time $12/24 = 0.5$ daily, and $D \in \{0, 2, 4, 6\}$ (measured in hours). The equilibrium trajectories, over two periods, are depicted in Figure 3. Observe that the queues of the fluid are increasing with D; e.g., the maximum

Figure 3 Equilibrium behavior (Solid: $D = 0$**;** $--: D = 2/24$**,** $\cdots: D = 4/24$**,** $-\cdot: D = 6/24$**)**



(a) active beds (b) blocked-beds (c) queue

queue when $D = 6$ hours is *more than twice* the size of the maximum queue when $D = 0$. The reason for this is due to the fact that $b(t)$ is increasing in D. Moreover, the process $b(t)$ drops to zero immediately after discharge when $D = 0$, and is always strictly positive otherwise, so that some service capacity is wasted at all times when $D > 0$. The convergence of the trajectories of $x(\cdot; D)$

as D decreases is also apparent, because both the jump times and jump magnitudes converge to the respective values when $D=0$ as $D\downarrow 0$.

6. Performance as a Function of the Discharge Policy and Capacity

In this section we demonstrate how long-run averages of key performance measures can be estimated from the fluid model, and provide important insights regarding how changes to the discharge policy and/or capacity impact the patient-flow in the IW. Furthermore, we can employ (9) to compute long-run costs whenever a cost function is given (taking $f(\cdot)$ in (9) to be the cost function). For example, a particular suggestion by the American College of Emergency Physicians (ACEP) to decrease discharge delays is to establish a discharge lounge (ACEP 2008) to which patients are transferred after it is decided that they should be discharged. The costs of establishing and running a discharge lounge should then be weighted against the savings corresponding to the increased throughput.

Here, we do not make any assumption on the cost function. Our focus is on applying Theorem 3 to compute key performance measures as functions of the controlled variables in the model, namely, the inspection time T, the discharge delay D, and the number of beds c. (We do not think of $\lambda(\cdot)$ and μ as being controlled.) In particular, the performance measures we consider are the long-run (i) proportion of fluid that is delayed in queue; (ii) average waiting time; and (iii) proportion of fluid that waits more than a predetermined time period.

6.1. Impacts of the Discharge Policy on Performance

Assume that $\rho_e<1$ and let

$$\Delta^* := \inf\{t\geq D : z^*(T+D+t)+b^*(T+D+t)\geq c\}$$

denote the first time during the equilibrium period that all the beds are occupied, where $\inf(\phi):=\infty$. The proof of Theorem 2 (see §C) shows that, either $\Delta^*\in(D,1+D)$, or $\Delta^*=\infty$. We explain how to compute Δ^* in Remark 5, following the proof of Theorem 2 in §C. Note that $\Delta^*=\infty$ implies that the fluid queue is null throughout the equilibrium period, whereas $D<\Delta^*<1+D$

implies that the IW is overloaded during the time interval $[T+\Delta^*, T+1+D)$, namely, the fluid queue is positive over that interval. We consider the latter case.

Let p_w denote the long-run proportion of fluid that is delayed in queue before entering the IW,

$$p_w := \lim_{t\to\infty} \frac{\int_0^t \lambda(s)1\{z(s)+b(s)=c\}ds}{\int_0^t \lambda(s)ds} = \frac{1}{\Lambda}\int_{T_k+\Delta^*}^{T_{k+1}+D} \lambda(s)ds,$$

and let $\bar{w}$ denote the long-run average waiting time of that fluid, averaged over all fluid that is delayed,

$$\bar{w} := \lim_{t\to\infty} \frac{\int_0^t \lambda(s)w(s)ds}{\int_0^t \lambda(s)ds} = \frac{1}{\Lambda}\int_{T_k+\Delta^*}^{T_{k+1}+D} \lambda(s)(T_{k+1}+D-s)ds.$$

Finally, let $\bar{p}_{w_\tau}$ denote the long-run proportion of fluid that waits at least τ units of time;

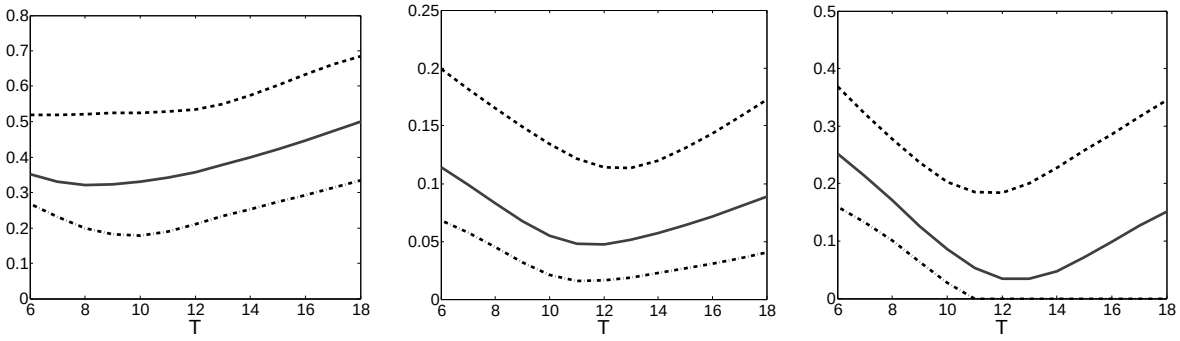
$$p_{w_\tau} := \lim_{t\to\infty} \frac{\int_0^t \lambda(s)1\{w(s)>\tau\}ds}{\int_0^t \lambda(s)ds} = \frac{1}{\Lambda}\int_{T_k+\Delta^*}^{T_{k+1}+D-\tau} \lambda(s)ds.$$

For the numerical example, we take the average service time be 4.5 days, i.e. $\mu = 1/4.5$, and the number of beds be $c = 50$. We use the same piecewise constant arrival rate function as the one used in the example in Figure 2, with a proper scaling to make the daily arrival rate $\Lambda = 9.53$. Note that the *nominal traffic intensity* in this case is $\Lambda/(c\mu) = 0.86$, whereas the effective traffic intensity ρ_e is higher, and depends on the actual value of the discharge delay D.

Figure 4 plots $\bar{p}_w$, $\bar{w}$ and $\bar{p}_{w_6}$ for different values of the inspection time T (in hours) and the discharge delay D. We restrict T to be in the range $6-18$. We consider three values for D: $0, 2$ and 4 hours, for which the effective traffic intensities are, respectively, $0.96, 0.97$ and 0.99. Thus, if the nominal traffic intensity is taken as the measure of arriving workload per day, as is often the case, then the system is (erroneously) considered as suitably loaded, although the system is actually severely congested in all three cases. We observe that, for a fixed value of D, the performance varies with different inspection times, as expected, but that there is no consistent trend across all performance measures. For example, for $D = 4/24$, when we move the inspection from $T = 10$ to $T = 12$, the value of p_w increases but $\bar{w}$ and p_{w_6} decreases, i.e. even though more fluid is waiting, the average waiting times are decreasing. This trade-off is to be expected since simply shifting

the inspection times has little impact on the effective total service rate in the system; indeed, $\rho_e = \Lambda/(c\beta)$ is independent of T. On the other hand, ρ_e is decreasing in D, so that decreasing the discharge delay does increase the maximal effective service rate. Specifically, we observe that as D decreases, all performance measures improve, and those improvements are quite significant. See also Figure 13 in §B.1 for additional simulation experiments of the stochastic systems that support these insights.

Figure 4 **Equilibrium performance measures based on the fluid model. (upper: $D = 4/24$, middle: $D = 2/24$, lower: $D = 0$)**



(a) p_w (b) $\bar{w}$ (c) p_{w_6}

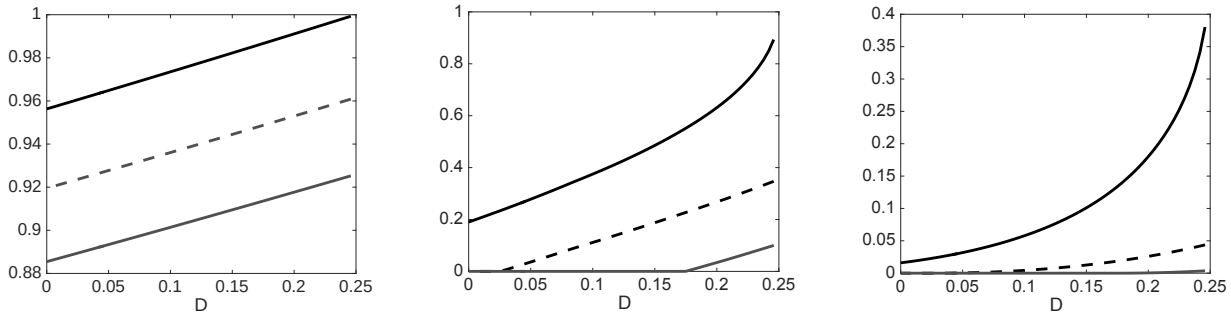
6.1.1. A Closer Look at the Impact of D We next investigate how D impacts ρ_e when all other parameters are kept fixed, treating $\rho_e := \rho_e(D)$ as a function of D. Employing Taylor's expansion to $e^{\mu D}$ around 0 gives $e^{\mu D} \approx 1 + \mu D$, so that

$$\rho_e(D) \approx \frac{\Lambda}{c(1-e^{-\mu})} + \frac{\Lambda \mu e^{-\mu}}{c(1-e^{-\mu})} D,$$

namely, $\rho_e(D)$ is approximately linear in D; see Figure 5 (a)[3]. (The inspection time in Figure 5 is $T = 11$, and the arrival-rate function and service rate are the same as in the example in Figure 4.) This near-linearity implies that changes to D of similar magnitudes have nearly the same relative impact on $\rho_e(D)$, e.g., reducing D from 4 to 3 hours has approximately the same relative impact on ρ_e as reducing D from 3 to 2 hours. However, from basic queueing theory we know that systems' performance deteriorates rapidly as the traffic intensity approaches 1. Since $\rho_e(D)$ is decreasing in

D, we expect improvements to the performance to diminish as D decreases. This is indeed observed in the numerical example in Figure 5 (b) and (c).

Figure 5 Equilibrium performance measures based on the fluid model. (upper: $c=50$, middle: $c=52$, lower: $c=54$)



(a) ρ_e (b) p_w (c) $\bar{w}$

REMARK 3. There are many causes for the discharge delays, some of which are due to the patients (e.g., the need to have a ride home or to a nursery home, etc.), whereas some are due to the hospital (e.g. waiting for lab results, cleaning beds etc.). Often times, the hospital can only decrease the delays that are due to its own policies, and so it may seem that any achievable reduction in the discharge delay is inconsequential. Specifically, when the average hospitalization time is several days long, and the delay is several hours long, one intuitively expects that reducing the delay by, say, one hour, will have a negligible impact on the IW's performance. However, our analysis here suggests that in a highly-congested IW (when ρ_e is close to 1), even a small reduction in the delay can lead to large improvements in the performance, so that investing resources in order to reduce the delay may be worthwhile. See Figure 8 (b) for an example based on real data.

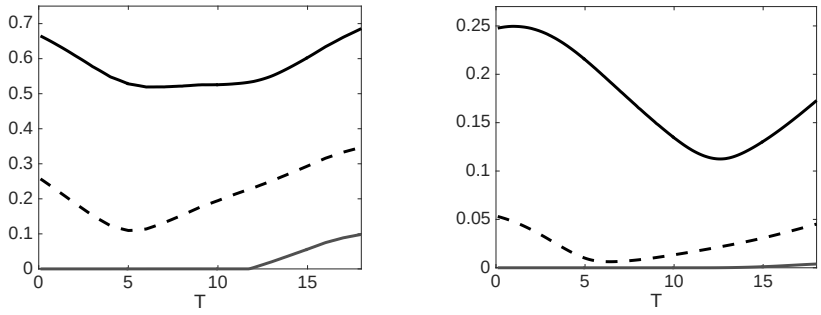
6.2. Impacts of Capacity on Performance

As we wrote above, capacity decisions, namely, changes to the number of beds in the IW, is a complex and costly project that can only occur once in several years. The goal is then to have the minimal number of beds such that certain Quality-of-Service (QoS) constraints are met for a projected arrival rates. We now demonstrate the impact of increasing the number of beds on

different performance measures by considering an example with the same arrival rate function and service rate as the one depicted in Figure 4. We set $D = 4$ hours.

In Figure 6 we plot $\bar{p}_w$ and $\bar{w}$ for different values of the number for beds c and the inspection time T (in hours). We plot a large range of T's to observe the full shape of the performance function. We observe that the optimal T can be very different for different performance measures and for different staffing levels. However, as expected, if c is increasing, then all performance measures improve. We note that $\rho_e = 0.985$ when $c = 50$; $\rho_e = 0.947$ when $c = 52$; and $\rho_e = 0.912$ when $c = 54$.

Figure 6 Equilibrium performance measures based on the fluid model. (upper: $c = 50$, middle: $c = 52$, lower: $c = 54$)



(a) p_w (b) $\bar{w}$

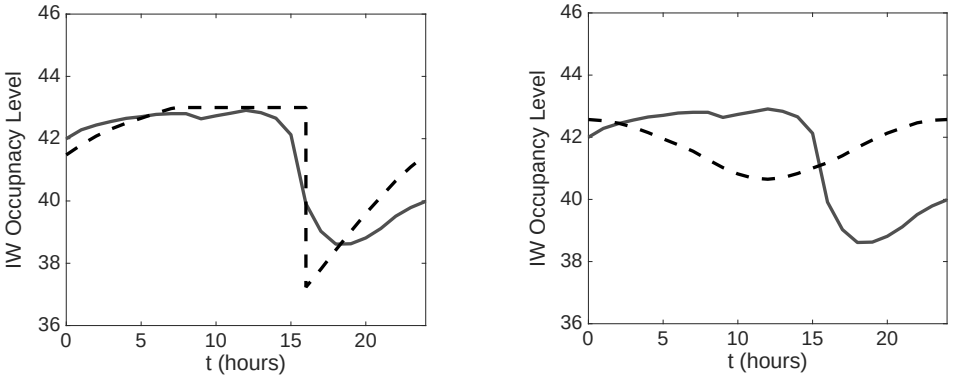
7. Empirical Evidence

We now provide empirical evidence of the effectiveness of the fluid approximation by comparing the long-run average behavior of our fluid model, namely, its time-dependent equilibrium behavior, to data from a large Israeli hospital, which was made available by the Service Engerprise Engineering Center (2012). To demonstrate that the assumptions we make regarding the discharge process are necessary, we also compare the data to the (time-dependent) equilibrium behavior of the fluid model of a "standard" time-varying multi-server queue, which we denote by $M_p(t)/M/c$. (We use $M_p(t)$ to denote a nonhomogeneous Poisson arrival process with a *periodic* intensity function; the second M in the notation represents IID exponentially-distributed service times.)

The hospital in the example has five IWs which are not pooled together, and we therefore focus on only one of those IWs. That IW has 43 beds; the inspection time ends at around 10:00 am; and most departures are concentrated around 4:00 pm, so that the discharge delay is taken to be $D = 6$ hours, i.e., $D = 6/24$ day. We fit the arrival rate function as a piecewise constant function based the ED-to-IW transfer counts over each hour of the day. Note that the *nominal service rate*, namely the *recovery time* of a patient, is a censored parameter, because recovered patients are only identified during the inspection round. Nevertheless, we can derive the average recovery time from the data regarding the bed-occupancy process via (3); in this example, the average recovery time is computed to be 4.16 days. For the $M_p(t)/M/c$ queue, we use the average length of stay fitted from data, 4.91 days, as the average service time. Note that the two models have the same traffic intensity.

The comparisons between the data and (a) our fluid approximation and (b) the fluid approximation for the $M_p(t)/M/c$ queue for the bed-occupancy level, are depicted in Figure 7. Observe that our fluid model captures the average dynamics of the bed-occupancy process, whereas the fluid model of the $M_p(t)/M/c$ and the average dynamics of the real IW move in opposite directions. In fact, the fluid model of the $M_p(t)/M/c$ queue attains its minimum at about the same time in which the actual process attains its maximum.

Figure 7 long-run average bed occupancy: comparison of (a) our fluid approximation and (b) fluid approximation for $M_p(t)/M/c$ to data (solid: real data, dashed: fluid)



(a) our fluid approximation

(b) fluid approximation for $M_P(t)/M/c$

In addition to the mismatch seen in Figure 7(b), the fluid approximation for the $M_p(t)/M/c$ model *exhibits no queue*, so that waiting times for IW beds cannot be estimated. (This is despite the fact that we increased the mean service time in this model.) On the other hand, our fluid model is experiencing a period of congestion (with a positive queue) every day in equilibrium which, as is seen in Figure 8, fits well with the data. We note that there is a baseline delay of 2 hours in the actual data at all times of the day. These baseline delays (which should not be confused with our discharge delays at the IW) are not due to the unavailability of IW beds (these delays exist even when beds are available at the IW), rather, they are caused by the need to prepare (clean) an assigned IW bed, have the transfer team available to physically move the patient to the IW, finish required paperwork, etc. Since the waiting time in our model is considered to be the time that passes between the bed request is made for a patient and until a bed is assigned to that patient, and not the time until the patient is physically moved to the IW, these baseline delays are not captured by our model. Nevertheless, the baseline delays can be added to the time-dependent waiting time in the fluid model, which we denote by $w(t)$, so that in Figure 8 (a) we compare $w(t)+2$ to the average waiting times obtained from the data.

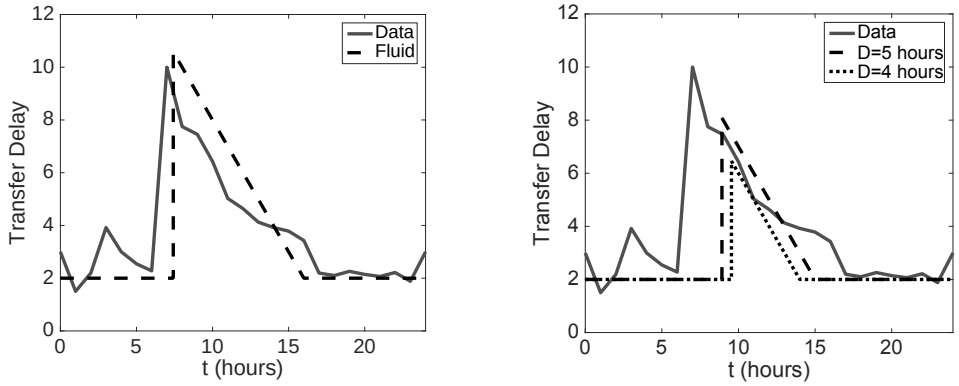
Similar to the performance analysis we conducted in Section 6, once we fit the model, we can check the impact of different policies changes. In Figure 8 (b), we compare $w(t)+2$ for different values of D, to demonstrate the benefit of decreasing discharge delay in a real IW. Observe that reducing the delay by one hour (to $D=5$ hours) leads to significant reduction of the predicted waiting times, and a smaller relative reduction of the predicted waiting times is achieved when the discharge delay is reduced by an additional hour (to $D=4$ hours); see §6.1.1.

8. Robustness of the Fluid Model and Extensions

We now consider two extensions to our basic IW model, and demonstrate the robustness of the fluid approximation to our basic modeling assumptions.

The first extension is multi patient-type queueing model in which the patient's type is determined by her service distribution. It is shown that the dynamics of the different patient-type processes

Figure 8 Average ED-to-IW transfer delays: comparison of the data to our fluid prediction for the virtual waiting-time process $w(t)+2$



(a) fluid approximation for the average waiting times compared to data

(b) fluid predictions for the average waiting times for reduced values of D

can be easily obtained from the basic homogeneous model by a simple transformation of the service rate μ. In the second extension, we assume that discharges occur during a time window according to some density function. We develop the fluid model for this model and compare it to the fluid approximation of the basic model, in which all patients depart at the same instant, to demonstrate that little is lost by making this simplifying assumption. Finally, we use simulations to show that the fluid model is quite insensitive to the assumption that service times are exponentially distributed.

8.1. Multiclass Model

For the multiclass model, we assume that there are m classes (types) of patients, where the service times of class-i patients are exponentially distributed with rate μ_i, $i = 1, \ldots, m$. Bed requests for each patient type arrive to the IW in accordance with a nonhomogeneous Poisson process having a rate function $\lambda_i(t)$. We let $\lambda(t) := \sum_{i=1}^{m} \lambda_i(t)$ and define $p_i := \lambda_i(t)/\lambda(t)$, $t \geq 0$. Due to basic properties of the nonhomogeneous Poisson process, we can equivalently assume that bed requests arrive according to a nonhomogeneous Poisson process with rate $\lambda(t)$, and each request is for a type-i patient with probability p_i, independently of all other requests. As in the basic model, there is a single inspection round each day at time T, and the delay between the inspection and the actual discharge is $D \geq 0$. If patients are admitted on a first-come-first-served basis regardless of

their type, it is sufficient to consider the total number of patients in queue, without keeping track of their types.

Since the fluid model is also a FWLLN by Proposition 2, it approximates the mean dynamics of the stochastic model. It is therefore intuitively clear that the dynamics of the multi-class fluid model can be inferred from the fluid approximation for the homogeneous-patient model by taking the service rate in the homogeneous model to be $\mu_s := \left(\sum_{i=1}^m p_i/\mu_i\right)^{-1}$. To see this, we now explicitly present the fluid model for the multi-class case, and compare it to the basic homogeneous model.

Let $q(t)$ denote the total amount of fluid (of all patient types) waiting for a bed in the IW, $b(t)$ be the amount of blocked beds, and $z(t) := (z_1(t), \ldots, z_m(t))$, where $z_i(t)$ is the number of class-i patients that are actively being served at time t. Also, let $z_s(t) := \sum_{i=1}^m z_i(t)$. Then the fluid model $x(t) := (q(t), b(t), z(t))$ for this system (note that $x(t)$ is now an $(m+2)$-dimensional process; we do not change the notation) is the unique solution to the following differential equation, given an initial condition $x(0)$.

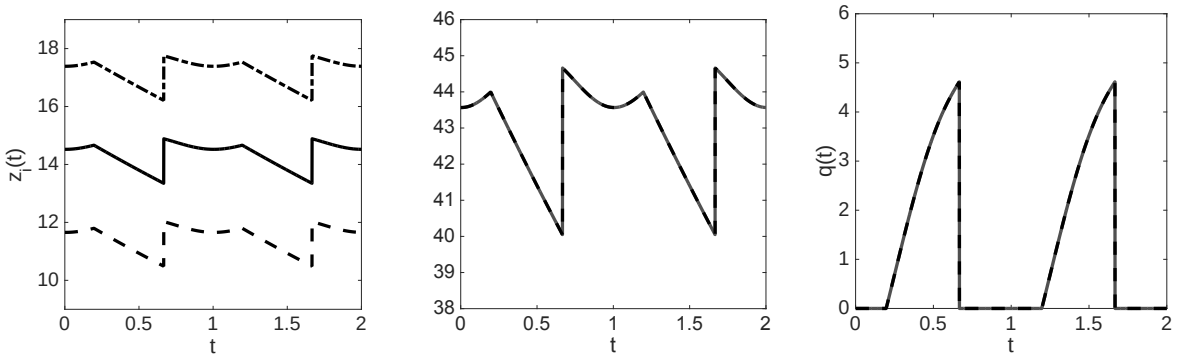
$$\begin{aligned}
dq(t) &= 1_{\{z_s(t)+b(t)=c\}}\lambda(t)dt;\\
dz_i(t) &= 1_{\{z_s(t)+b(t)<c\}}\lambda(t)p_i dt - z_i(t)\mu_i dt, \quad i=1,\ldots,m;\\
db(t) &= \sum_{i=1}^m z_i(t)\mu_i dt \quad \text{for } t \neq T_k + D,
\end{aligned}$$

and at the jump epoch $T_k + D$, $k \geq 0$,

$$\begin{aligned}
q(T_k+D) &= \left(q((T_k+D)-) - b(T_k)\right)^+\\
z_i(T_k+D) &= z_i((T_k+D)-) + \left(q((T_k+D)-) \wedge b(T_k)\right)p_i, \quad i=1,\ldots,m;\\
b(T_k+D) &= b((T_k+D)-) - b(T_k).
\end{aligned}$$

8.1.1. Numerical Examples Qualitative analysis of the fluid approximation of the multiclass model is more tedious than that of the single-class fluid approximation, but numerically solving this model is not computationally harder. Figure 9 plots $z(t)$, $z_s(t)$ and $q(t)$ in equilibrium for an example with three patient types having service rates $(\mu_1, \mu_2, \mu_3) = (1/4, 1/5, 1/6)$. We take

Figure 9 **Equilibrium behavior of the multiclass model ($c=50$, $\lambda(t)=8.6+4.3\sin(2\pi t)$, $\mu=(1/4,1/5,1/6)$, $p=(1/3,1/3,1/3)$, $T=12/24$, $D=4/24$, $\mu_s=1/5$)**



(a) active beds (upper: $z_3(t)$, middle: $z_2(t)$, lower: $z_1(t)$)

(b) Total amount of active beds (solid: multiclass, dash: single-class)

(c) queue (solid: multiclass, dash: single-class)

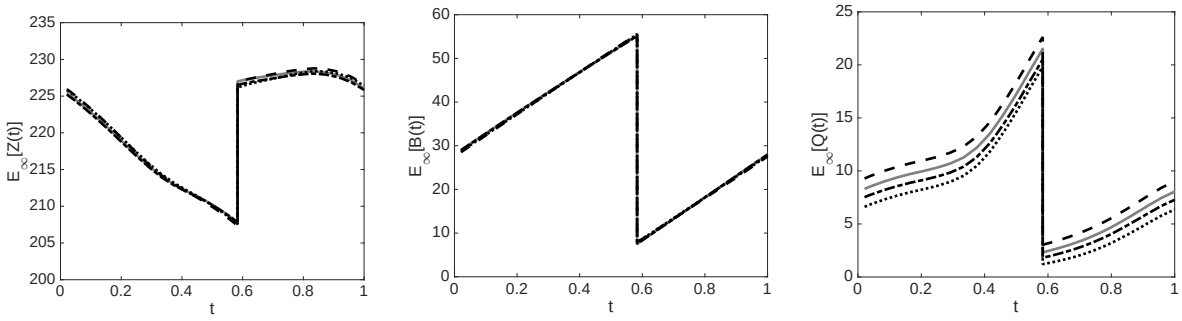
$p_i=1/3$, so that $\mu_s=1/5$. We compare the equilibrium dynamics of $z_s(t)$ and $q(t)$ to the periodic equilibrium of a single-class model with service rate $\mu=\mu_s=1/5$.

To further demonstrate the robustness of the fluid model we conduct simulation experiments for the multiclass model with non-exponential service times. In Figure 10 we compare three multiclass systems, each having a different type of service time distribution. In particular, we compare the following three distribution in which we match the means of the service times: 1) exponential distributions with rate $\mu_1=1/\exp(4+0.5)$, $\mu_2=1/\exp(4.2+0.5)$ and $\mu_3=1/\exp(4.4+0.5)$; 2) lognormal distribution with $(\mu_1,\sigma_1)=(4,1)$, $(\mu_2,\sigma_2)=(4.2,1)$ and $(\mu_3,\sigma_3)=(4.4,1)$; 3) Erlang-2 distribution with rate $\mu_1=2/\exp(4+0.5)$, $\mu_2=2/\exp(4.2+0.5)$ and $\mu_3=2/\exp(4.4+0.5)$. We set the proportions of each patient type to be $p_1=p_2=p_3=1/3$. Note that the lognormal distribution has a heavier tail and is more variable, whereas the Erlang distribution has a lighter tail, and is less variable, than the exponential distribution. The arrival-rate function $\lambda(t)$ is fitted from the same data we used to produce Figure 4. The equilibrium of the stochastic processes are calculated using batch means with 10 batches based on a sample path of 10^4 periods.

We note that, even though the examples shown here demonstrate the robustness of the fluid model to the service-time distribution, caution is needed, because not all distributions provide the

similar results. For example, if the service-time distribution is discrete, and in particular deterministic, or it it possesses a very heavy tail, such as the Weibull distribution, then the effectiveness of the fluid approximation deteriorates.

Figure 10 Simulation comparison of the equilibrium behavior of the multiclass with different service time distributions and a homogeneous model with exponential service times. (gray: single-class exponential with the same mean, $\cdots$: exponential, $--$: lognormal, $-\cdot$: Erlang-2) ($c = 265$, $\mu = (4, 4.2, 4.4)$, $\sigma = (1, 1, 1)$ $p = (1/3, 1/3, 1/3)$, $T = 10/24$, $D = 4/24$, $\mu_s = \left(\frac{1}{3}\left(\exp(4+0.5) + \exp(4.2+0.5) + \exp(4.4+0.5)\right)\right)^{-1}$)



(a) total number of active beds (b) blocked beds (c) queue

8.2. Model with a Discharge Window

We now consider a model in which discharges are distributed over a time window, as is observed in data, and compare it to our basic model, which has a single discharge epoch. As before, the daily inspection still takes place at a time T_k on day k, and all the blocking patients $B(T_k)$ are identified and discharged on the same day. However, instead of leaving at a single time point, the discharged patients in this new model leave the IW over a time window $[T_k + D_1, T_k + D_2]$, $0 \le D_1 < D_2 \le 1$ according to a density function $\eta(t)$ with support over the discharge window.

In the fluid model for this system, the instantaneous discharge rate at time $t \in [T_k + D_1, T_k + D_2]$ is $b(T_k)\eta(t)$, and is equal to 0 outside that interval. We assume that the discharge pattern is similar each day, so that $\eta(t+1) = \eta(t)$, $t \ge 0$. Then for $t \in [T_{k-1} + D_2, T_k + D_1)$, $k \ge 1$, the fluid model satisfies the following differential equation:

$$dq(t) = 1_{\{z(t)+b(t)=c\}}\lambda(t)dt;$$

$$dz(t) = 1_{\{z(t)+b(t)<c\}}\lambda(t)dt - z(t)\mu dt; \tag{14}$$
$$db(t) = z(t)\mu dt;$$

and during the discharge window, $t \in [T_k + D_1, T_k + D_2)$, $k \geq 1$,

$$\begin{aligned} dq(t) &= 1_{\{z(t)+b(t)=c\}}\lambda(t)dt - 1_{\{q(t)>0\}}b(T_k)\eta(t)dt; \\ dz(t) &= 1_{\{z(t)+b(t)<c\}}\lambda(t)dt + 1_{\{q(t)>0\}}b(T_k)\eta(t)dt; \\ db(t) &= z(t)\mu dt - b(T_k)\eta(t)dt. \end{aligned} \tag{15}$$

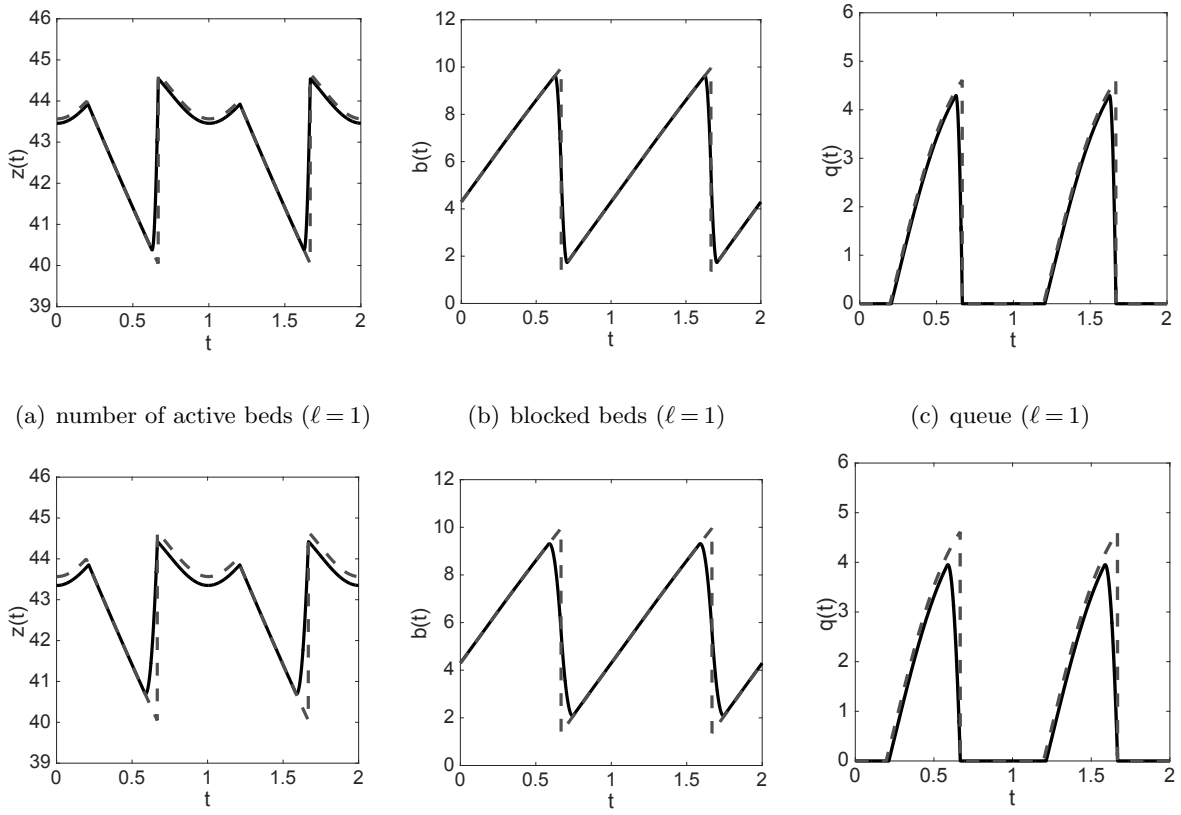
Note that, in order to apply this model in practice, the data required is $(\lambda(\cdot), \mu, c, T, \eta(\cdot))$. In particular, one must estimate the function $\eta(\cdot)$ instead of the single parameter D that is required for the basic model.

8.2.1. Numerical Examples We consider two systems, differing from each other only by the length of their discharge window, which are equal to 2 hours and 4 hours. In both examples, the discharge density functions take a *symmetric triangular shape*, both having their maximum at the center of their support. We let $D := (D_1 + D_2)/2$ denote the center of these supports (which is equal in both examples). Thus, the discharge windows are the intervals $[D - \ell, D + \ell]$, for $\ell = 1, 2$.

Figure 11 plots the periodic equilibria of the two systems, and compares them to the periodic equilibrium of our basic model having a single jump at time D (the center of the supports of the discharge windows). The three component processes of $x(t)$ for the system with $\ell = 1$ are depicted in sub-figures (a)–(c), and for the system with $\ell = 2$ in sub-figures (d)–(f). We observe that, in either example, the periodic equilibrium of the system with a single jump at time D approximates the model with a discharge window well, although, as the window size gets larger, the discrepancy between the two models increases.

When the discharge window is too large, or when the empirical discharge distribution is highly skewed, the basic model with a single departure epoch may not approximate the "real system" well. However, the basic model can still be useful because the density function $\eta(\cdot)$ may not be known

Figure 11 Comparisons between the basic fluid model and the fluid model of an IW with a discharge window. (solid: model with discharge window, dash: model with jump) ($c = 50$, $\lambda(t) = 8.6 + 4.3\sin(2\pi t)$, $\mu = 1/5$, $D = 4/24$, $D_1 = D - \ell$, $D_2 = D + \ell$)



(a) number of active beds ($\ell = 1$) (b) blocked beds ($\ell = 1$) (c) queue ($\ell = 1$)

(d) number of active beds ($\ell = 2$) (e) blocked beds ($\ell = 2$) (f) queue ($\ell = 2$)

exactly, either because there is no sufficient data, or because management is considering changes to the discharge policy, and the discharge pattern under a future policy is unknown. In such cases, the basic model can be employed to obtain best- and worst-case scenarios by considering the delay to be D_{min} and D_{max}, respectively. When the discharge distribution is known, the fluid model in (14)-(15) can be employed. In this case, one would numerically solve the fluid equations until equilibrium behavior is observed (assuming that the fluid model is asymptotically periodic), and the observed equilibrium can then be employed to evaluate desired performance measures.

9. Summary and Future Work

In this paper we proposed a queueing model for patient-flow dynamics associated with large IWs. Those dynamics differ from classical non-stationary queueing models by the special discharge mech-

anism under which patients can leave the IW only if they are approved to leave by a physician. Based on current practices in most large hospitals, as well as empirical data, we assumed that most discharge decisions are made once per day during a morning inspection round, and that a large majority of the patients leave the IW during a relatively short time window several hours after the inspection round has ended.

Stochastic Model. We started by identifying the *maximum effective service rate per bed* β and *effective traffic intensity* ρ_e in (3), and proving (in Theorem 1) that $\rho_e < 1$ is a necessary and sufficient condition for stability of the queue.

Fluid Approximation. Due to the complexity of the stochastic system, we employed a deterministic fluid model to analyze the queueing dynamics. The fluid model, which is a bona-fide limiting approximation for a large IW by Proposition 2, was shown in Theorem 3 to be asymptotically periodic, namely, to converge to a unique periodic equilibrium regardless of its initial condition if $\rho_e < 1$. Theorem 5 in §A below shows that the fluid model is stable if and only if $\rho_e < 1$.

Robustness of Fluid Model. To facilitate qualitative analysis, our basic model had one type of patients, exponential service times, and a single discharge epoch. In §8 we have shown that these assumptions are not very restrictive for the fluid analysis. In particular, this fluid model can be employed to analyze a multi-patient type model by a simple transformation of the service rate, and serves as a good approximation for a fluid approximation for a model with a discharge window. Finally, simulation experiments show that the fluid model is an effective approximation for stochastic systems with non-exponential service times.

Future Work. It is significant that the model analyzed here does not describe all the units in the hospital; it is designed according to data and observations in Medicine Units (general wards), and was also found to fit the dynamics in Observation and Cardiology Units in a large teaching hospital. The patient-flows in other units, such as ICUs, Obstetrics, Oncology and Surgery, have different characteristics than those described in our model. Nevertheless we have found that the "usual assumptions" taken in the literature of queueing models for service systems (such as IID service times) do not fit any of these other units, so that specialized models are needed.

Acknowledgements

We thank Service Enterprise Engineering (SEE) Lab in the Technion for providing us with the patient-flow data. We also thank an AE and three referees for many useful comments and suggestions which helped improve the paper. Jing Dong received support from the National Science Foundation (NSF), grant CMMI 1762544. Ohad Perry received support from NSF grants CMMI 1436518 and CMMI 1763100.

Appendix A: More on Stability of the Fluid Model

Under any form of stability, $q(t)$ should be bounded, namely, it must hold that, for any initial condition $q(0)$, there exists $M_0 < \infty$ (that depends on $q(0)$), such that

$$q(t) < M_0, \quad \text{for all } t \geq 0. \tag{16}$$

However, taking (16) as the condition for stability is clearly too weak. Indeed, as discussed below Theorem 3, the fluid queue may be bounded on $[0, \infty)$, and may even converge to a periodic equilibrium, without ever reaching state 0. In such a case, Proposition 2 implies that the queue in the stochastic system the fluid approximates remain very large throughout (since $Q^n(t) \approx nq(t)$ for large n). We therefore must require that $q(t)$ visits 0 indefinitely as $t \to \infty$.

Consider the hitting times

$$\tau_0 := \inf\{k \geq 0 : q(T_k + D) = 0\} \quad \text{and for } i \geq 1, \quad \tau_i := \inf\{k \geq \tau_{i-1} : q(T_k + D) = 0\}, \tag{17}$$

where $\inf(\phi) := \infty$. Then τ_i is the ith time at which the queue empties, $i \geq 0$.

Definition 2 (Stability). We say that the system $x(t)$ is stable if for any initial condition x_0, there exists a constant $M := M(x_0) < \infty$, such that $\tau_i - \tau_{i-1} < M$ for all $i \geq 1$, for τ_i in (17), $i \geq 0$. Otherwise, the system is said to be unstable.

Remark 4. It follows from Definition 2 that a system is unstable if there exists an initial condition $x(0)$ for which one of the following holds. (Note that (16) might hold in either case.)

(i) $\tau_N = \infty$, for some $N \geq 0$, so that the queue empties a finite number of times on $[0, \infty)$.

(ii) $\tau_i < \infty$ for all $i \geq 0$, but $\tau_i - \tau_{i-1} \to \infty$ as $i \to \infty$. Namely, the queue experiences increasing oscillations.

The following theorem shows that the condition for stability of the stochastic system is necessary and sufficient for stability of its fluid approximation.

THEOREM 5. *The fluid model is stable if and only if* $\rho_e < 1$.

Proof. We consider the embedded discrete-time system $x_k := x(T_k)$, $k \geq 0$. In the kth cycle, the input to the system is Λ, and the output from the system, in the form of a jump at time $T_k + D$, is b_k.

When $\rho_e < 1$**:** We first prove that, for any initial condition $x(0)$, there exists $\tau_0 < \infty$, such that $q(\tau_0) = 0$. Take the contradictory assumption that, for some $x(0)$ with $q(0) > 0$, it holds that $q(T_k + D) > 0$ for all $k \geq 0$. This assumption implies that at discharge times $\{T_k + D : k \geq K\}$ there will be b_k new fluid entering service, so that $z(T_k + D) = z((T_k + D)-) + b(T_k)$, $k \geq 0$. Then

$$b_{k+1} = z_k(1 - e^{-\mu}) + b_k(1 - e^{-\mu(1-D)}) = (c - b_k)(1 - e^{-\mu}) + b_k(1 - e^{-\mu(1-D)}) = c(1 - e^{-\mu}) + b_k(e^{-\mu} - e^{-\mu(1-D)}).$$

This forms a recursive formula for b_k, and we have

$$b_k = \frac{c(1 - e^{-\mu})}{1 + e^{-\mu(1-D)} - e^{-\mu}} + \left(e^{-\mu} - e^{-\mu(1-D)}\right)^k \left(b_0 - \frac{c(1 - e^{-\mu})}{1 + e^{-\mu(1-D)} - e^{-\mu}}\right) = c\beta + \alpha^k(b_0 - c\beta), \quad k \geq 0, \quad (18)$$

for β in (3) and for

$$\alpha := e^{-\mu} - e^{-\mu(1-D)}. \quad (19)$$

Since $0 \leq b_0 \leq c$ and $-1 < \alpha < 0$, it follows that $\lim_{k\to\infty} b_k = c\beta$. Now, this limit of b_k implies that, for any $\epsilon \in (0, c\beta - \Lambda)$, there exists $K_\epsilon \geq 0$, such that $|b_k - c\beta| < \epsilon$ for all $k \geq K_\epsilon$. Let $\delta = c\beta - \Lambda - \epsilon > 0$. Then $b_k > c\beta - \epsilon = \Lambda + \delta$ for $k \geq K_\epsilon$. Since $q_{k+1} = (q_k + \Lambda - b_k)^+$, so that $q_{k+1} = q_k + \Lambda - b_k$ for all $k \geq 0$ due to the contradictory assumption, it holds that $q_{k+1} - q_k < -\delta$ for all $k \geq K_\epsilon$. This implies that q_k will reach zero for k large enough, contradicting our assumption that $q(t)$ is always positive. Hence, for any initial condition $x(0)$, there exists $\tau_0 < \infty$, such that $q(\tau_0) = 0$. The time homogeneity of the fluid model implies that $\tau_n < \infty$ for all $n \geq 0$.

We next develop a uniform upper bound for $\tau_{n+1} - \tau_n$, for all $n \geq 0$. Let $L_n := \tau_{n+1} - \tau_n$. Observe that

$$b(T_k + D) + z(T_k + D) + q(T_k + D) = b(T_{k-1} + D) + z(T_{k-1} + D) + q(T_{k-1} + D) + \Lambda - b_{T_k}, \quad k \geq 1,$$

because the net flow of fluid into the system between $T_{k-1} + D$ and $T_k + D$ is $\Lambda - b_{T_k}$. It then follows that $q(T_k + D) \leq \left(q(T_{k-1} + D) + \Lambda - b(T_k)\right)^+$. In particular, taking $k := T_{\tau_n + d}$, and recalling that $q(T_{\tau_n} + D) = 0$, we see that $L_n \leq \inf\left\{d : \sum_{i=1}^d b_{\tau_n + i} \geq d\Lambda\right\}$. If $L_n > 1$, then by (18), we have that, for any $d < L_n$,

$$\sum_{i=1}^{d} b_{\tau_n + i} = \sum_{i=1}^{d} \left(c\beta + \alpha^{i-1}(b_{k_n+1} - c\beta)\right) = dc\beta + \frac{1 - \alpha^d}{1 - \alpha}(b_{k_n+1} - c\beta) > dc\beta - c\beta, \quad (20)$$

where α is defined in (19). Let

$$M := \lceil c\beta/(c\beta - \Lambda) \rceil. \tag{21}$$

Plugging $d = M$ in (20) gives

$$\sum_{i=1}^{M} b_{\tau_n+i} > Mc\beta - c\beta \geq M\Lambda,$$

so that $\tau_{n+1} - \tau_n = L_n \leq M$, for all $n \geq 0$. Thus, if $\rho < 1$, the system is stable.

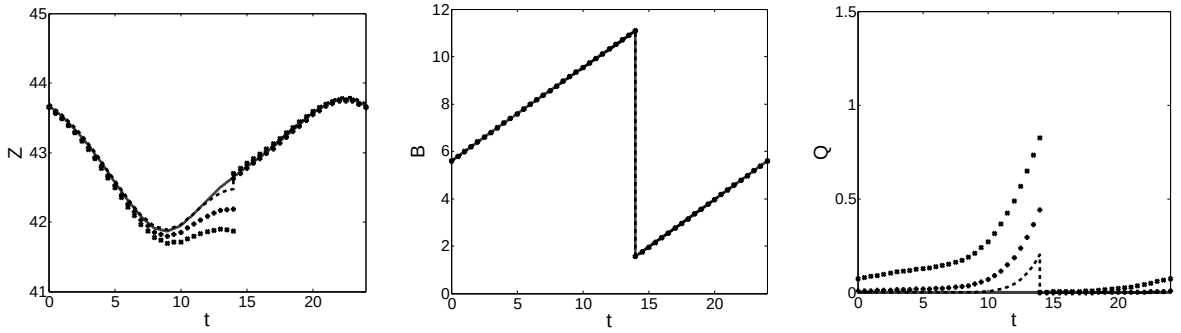
When $\rho_e \geq 1$: Take $x(0) = x_0 = (z_0, b_0, q_0)$ such that $b_0 = c\beta = c - z_0$, and $q_0 > \Lambda$. Then $q(T_1 + D) > 0$, and (18) shows that $b_1 = c\beta$. Hence, since $q_1 = q_0 + \Lambda - b_0 = q_0 + \Lambda - c\beta$, we see that, if $\rho_e = 1$ ($\Lambda = c\beta$), then $q_1 = q_0$, whereas, if $\rho_e > 1$, then $q_1 > q_0 > \Lambda$. It follows immediately that $q_n = q_0$ for all $n \geq 1$ when $\rho_e = 1$, and that $q_n \to \infty$ as $n \to \infty$, if $\rho_e > 1$. $\square$

Appendix B: The Fluid Approximation as a Limit

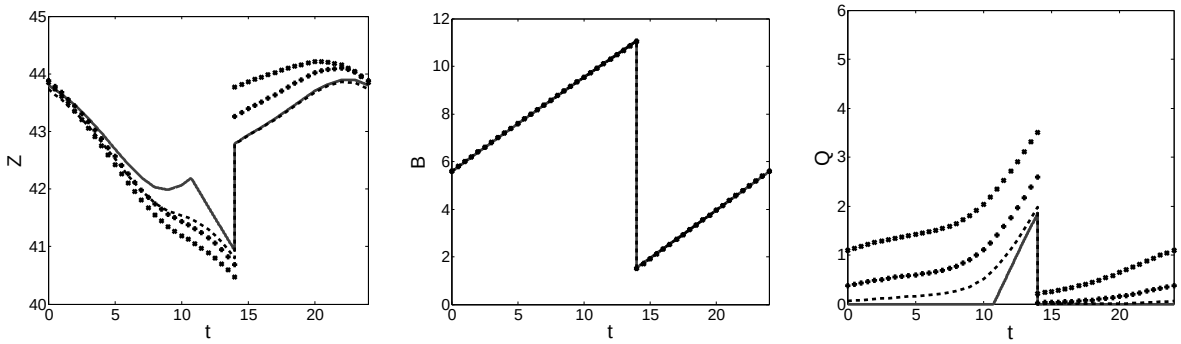
By Proposition 2, the fluid model is achieved as a FWLLN under many-server scaling, assuming the service times are exponential random variables. We now present numerical results to demonstrate that the fluid approximation is effective for finite systems.

Figure 12 compares the periodic steady-state average of scaled stochastic processes to the equilibrium of their fluid approximation. We take the service rate to be $\mu = 4.5$, and fitted a $\lambda(t)$ from real data as in Figure 4. Specifically, we set $\lambda_n(t) = n\lambda(t)$ and $c_n = n \cdot c$, where $c \in \{50, 52, 55\}$ and $n \in \{5, 10, 20\}$. The inspection takes place at T =10am and the discharge takes place at 2pm, so that $D = 4/24$ (a four-hour delay). The equilibrium of the stochastic processes are long-run average calculated using batch means with 10 batches based on a sample path of 10^3 days. We observe that as n increase from 5 to 20, the scaled stochastic processes get closer the fluid model. However, the convergence becomes slower as the *effective traffic intensity* ρ_e increases to 1, since stochastic fluctuations become more dominant in the critical case. (This is consistent with other queueing models, such as the $M/M/c$.) We also note that the nominal traffic intensity is much lower than the effective one. For example, when $c = 52$, the nominal traffic intensity is $\Lambda/(c\mu) = 0.82$, while the effective one is $\rho_e = 0.95$, for ρ_e in (3); effectively, the IW is critically loaded. In particular, while the fluid queue is null immediately after the jump at 2pm, the stochastic process has a positive queue right after the jump for $\rho_e \geq 0.95$. (Naturally, the fluid queue is smaller than in the stochastic system, since it does not capture lower-order stochastic fluctuations.) Hence, the stationary fluid queue serves as a lower bound for the stochastic queue, but captures its dynamics, as is clear from Figure 12.

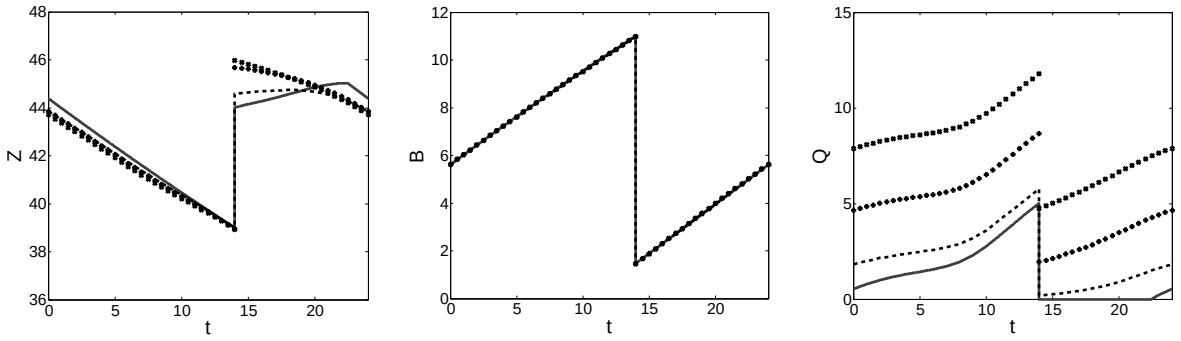
Figure 12 Comparison between the periodic equilibrium of the scaled stochastic processes and the fluid model (×: $n=5$, +: $n=10$, dotted: $n=20$, solid: fluid)



(a) $\rho_e = 0.9 (c = 55)$: $Z_n(t)/n$ (b) $\rho_e = 0.9 (c = 55)$: $B_n(t)/n$ (c) $\rho_e = 0.9 (c = 55)$: $Q_n(t)/n$



(d) $\rho_e = 0.95 (c = 52)$: $Z_n(t)/n$ (e) $\rho_e = 0.95 (c = 52)$: $B_n(t)/n$ (f) $\rho_e = 0.95 (c = 52)$: $Q_n(t)/n$



(g) $\rho_e = 0.985 (c = 50)$: $Z_n(t)/n$ (h) $\rho_e = 0.985 (c = 50)$: $B_n(t)/n$ (i) $\rho_e = 0.985 (c = 50)$: $Q_n(t)/n$

B.1. The fluid performance measures

We now present a simulation experiment to support the insights regarding the impact of the inspection time T and discharge delay D on performance, as in the numerical example depicted in Figure 4. In Figure 13 below, we compare simulated long-run averages of the stochastic systems to their corresponding fluid equilibrium performance measures. The system we simulate is the same as in Figure 12 with $c = 50$ and

$n = 20$ to minimize the stochastic noise. We also provide the 95% confidence bound for the simulated results. As expected, and as can be seen in Figure 12, stochastic fluctuations (that are not captured by the fluid limit) have a non-negligible impact on the stochastic queue process. In turn, this stochastic noise causes the simulated systems to exhibit worse performance than their fluid estimates. Nevertheless, it is evident that (with 95% confidence), the shapes of the simulated curves are similar to those generated by the fluid model, and so the main insights pointed out in §6.1 still hold. In particular, there is no universally optimal inspection time which uniformly improves all performance measures, whereas reducing the discharge delay achieves universal improvements across all performance measures.

Appendix C: Proofs of the Results in the Paper

Before proving the results stated in the paper, we prove the following auxiliary lemma. Recall $\{\tau_i : i \geq 0\}$ in (17) and that τ_i^* denotes the ith time at which the queue empties *in equilibrium*, i.e., it is τ_i when the initial condition $x(0)$ is on the equilibrium trajectory.

LEMMA 1. *If a periodic equilibrium u^* exists, then $\tau_{i+1}^* - \tau_i^* = 1$, $i \geq 0$.*

Proof. We prove the lemma by assuming that there exists a periodic equilibrium with

$$\tau_{k+1}^* - \tau_k^* = n, \quad \text{for some } n > 1, \tag{22}$$

implying that $q^*(T_i + D) > 0$ for $i = 1, \ldots, n-1$ and $q^*(T_n + D) = 0$. We will show that this assumption leads to a contradiction. Assuming (22) holds, we have the following system of equations for b_k^*. For α in (19) and $k = 2, \ldots, n$,

$$b_k^* = c\beta + \alpha^{k-1}(b_1 - c\beta) \quad \text{and} \quad \sum_{i=1}^{n} b_i^* = n\Lambda,$$

implying that

$$b_1^* = c\beta - \frac{1-\alpha}{1-\alpha^n} n(c\beta - \Lambda).$$

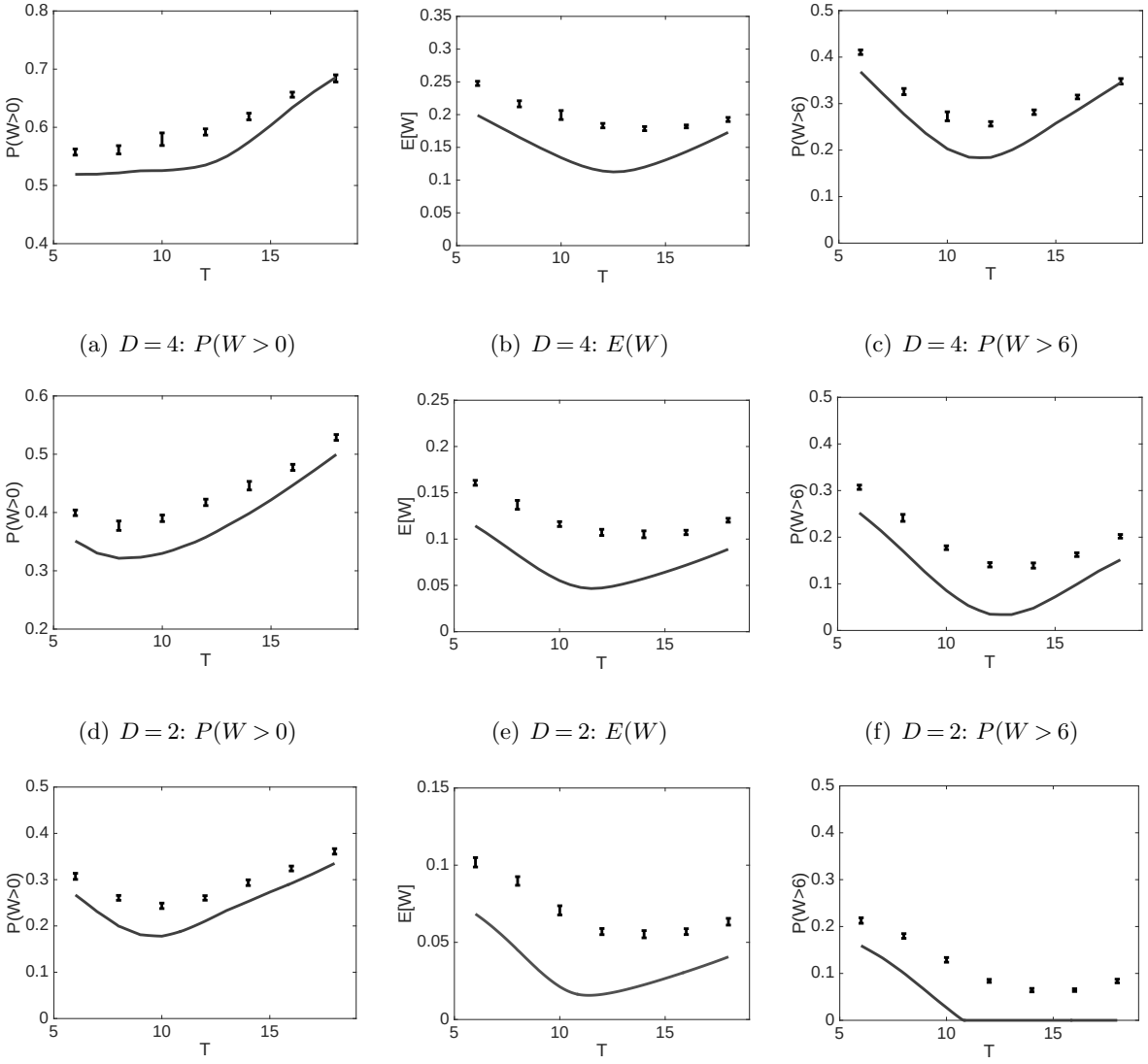
Furthermore, as $b_{n+1}^* = b_1^*$ and $q^*(T_n + D) = 0$, we have that $b_1^* = b_{n+1}^* > (c - b_n^*)(1 - e^{-\mu})$. Plugging in the formula for b_n^* gives $b_1^* > (c - c\beta + \alpha^{n-1}(b_1^* - c\beta))\,(1 - e^{-\mu})$. It follows that

$$b_1^* > \frac{c(1-e^{-\mu}) - \alpha^{n-1} c\beta (1-e^{-\mu})}{1 - \alpha^{n-1}(1-e^{-\mu})},$$

which further implies that

$$c\beta - \frac{1-\alpha}{1-\alpha^n} n(c\beta - \Lambda) > \frac{c(1-e^{-\mu}) - \alpha^{n-1} c\beta (1-e^{-\mu})}{1 - \alpha^{n-1}(1-e^{-\mu})}. \tag{23}$$

Figure 13 Comparison between the long-run average performance measures of the stochastic system and the fluid model (95% error bar, solid: fluid)



(a) $D=4$: $P(W>0)$ (b) $D=4$: $E(W)$ (c) $D=4$: $P(W>6)$

(d) $D=2$: $P(W>0)$ (e) $D=2$: $E(W)$ (f) $D=2$: $P(W>6)$

(g) $D=0$: $P(W>0)$ (h) $D=0$: $E(W)$ (i) $D=0$: $P(W>6)$

For the left hand side of (23), we have

$$c\beta - \frac{1-\alpha}{1-\alpha^n} n(c\beta - \Lambda) < c\beta.$$

For the right hand side of (23), we have (since $c\beta = c(1-e^{-\mu})/(1-\alpha)$),

$$\frac{c(1-e^{-\mu}) - \alpha^{n-1} c\beta(1-e^{-\mu})}{1-\alpha^{n-1}(1-e^{-\mu})} = \frac{c(1-e^{-\mu})(1-\alpha-\alpha^{n-1}(1-e^{-\mu}))/(1-\alpha)}{1-\alpha^{n-1}(1-e^{-\mu})} > \frac{c(1-e^{-\mu})}{1-\alpha} = c\beta.$$

We have thus reached a contradiction, implying that (22) cannot hold for a stationary equilibrium. □

We next prove Theorem 1. We note that the stability of the embedded DTMC when $D = 0$ follows from Lemma 1 in Chan et al. (2017), but a new proof is needed for the general case (with $D > 0$). To prove the theorem, we employ a more involved Lyapunov argument than the one in Chan et al. (2017), building on the theory developed in Foss and Konstantopoulos (2006). In particular, unlike the standard Lyapunov argument, in which one establishes that the Lyapunov function applied to the corresponding DTMC has a negative drift over one step, in our case the drift is shown to be negative over a random number of steps. The instability of the embedded DTMC X_k when $\rho_e \geq 1$ is proved via a coupling argument, based on the observation made in Remark 2.

Proof of Theorem 1. Consider the case $\rho_e < 1$. Since irreducibility and aperiodicity of the DTMC X_k are both immediate, we need only show that the DTMC is also positive recurrent. To this end, we define the function $V : \mathbb{R}_3 \to \mathbb{R}$ via $V(X_k) = Q_k$, and note that, since $V(x) \geq 0$ and $\sup_{x\in\mathbb{N}^3} V(x) = \infty$, V is a candidate Lyapunov function. Next, for any $N \in \mathbb{N}$, let $\kappa_N := \inf\{k \geq 1 : V(X_k) \leq N\}$, $K := \lceil c\beta/(c\beta - \Lambda)\rceil$ and $N_0 := c(K + 1)$. Finally, for $X_0 = x = (z, b, q)$, we define the functions $g(x) = 1 + K1_{\{q>N_0\}}$ and $h(x) = (c\beta - \Lambda)1_{\{q\geq N_0\}} - \Lambda 1_{\{q<N_0\}}$. Then for $x \in \{r : V(r) \leq N_0\}$, it holds that

$$E[V(X_{g(x)})|X_0 = x] - V(x) = E[Q_1|X_0 = x] - q = (q + \Lambda - b)^+ - q \leq \Lambda = -h(x).$$

We next consider $x \in \{r : V(r) > N_0\}$. First, observe that there can be at most c departures in each cycle, if $X_0 = x$, $Q(t) > 0$ for $0 \leq t \leq T_{K+1}$. We also observe that $B_k|B_{k-1} \stackrel{\mathrm{d}}{=} D_1 + D_2$ where $D_1 \sim Binomial(c - B_{k-1}, 1 - e^{-\mu})$ and $D_2 \sim Binomial(B_{k-1}, 1 - e^{-\mu(1-D)})$, for $1 \leq k \leq K + 1$. In particular, for $1 \leq k \leq K + 1$,

$E[B_k|B_{k-1}] = (c - B_{k-1})(1 - e^{-\mu}) + B_{k-1}\left(1 - e^{-\mu(1-D)}\right) = c(1 - e^{-\mu}) + \alpha B_{k-1}$ and

$E[B_k|X_0 = x] = c\beta + \alpha^k(b - c\beta)$.

Let A_k denote the total number of arrivals in the kth day. Then $(A_k : k \in \mathbb{N})$ are independent Poisson random variables, each with mean Λ, so that

$$\begin{aligned}
E[V(X_{g(x)})|X_0 = x] - V(x) &= E[Q_{K+1}|X_0 = x] - q \\
&= E\left[q + \sum_{k=1}^{K+1} A_k - \sum_{k=1}^{K+1} B_k\right] - q \\
&= (K + 1)\Lambda - \sum_{k=1}^{K+1} \left(c\beta + \alpha^k(b - c\beta)\right) \\
&\leq -(c\beta - \Lambda) = -h(x).
\end{aligned}$$

For our choice of $h(x)$ and $g(x)$, we also have $\inf_{x\in\mathbb{N}^3} h(x) \geq -\Lambda$; and for $x \in \{x : V(x) > N_0\}$; $h(x) > 0$ and $g(x)/h(x) < (K+1)/(c\beta - \Lambda)$. The positive reccurence then follows from Theorem 2 in Foss and Konstantopoulos (2006).

Next assume that $\rho_e \geq 1$. The proof that X_k is null recurrent (when $\rho_e = 1$) or transient (when $\rho_e > 1$) follows by coupling our IW model with the $M_p(t)/G/c$ model in Remark 2, in which the service times are distributed as $\lceil S + D \rceil$. Specifically, the event $A :=$\{the queue does not empty for at least one period\} clearly has a positive probability, and so we can initialize the system with an appropriate initial condition in A. We couple the IW system with the aforementioned $M_p(t)/G/c$ by considering both systems on the same probability space, giving both the exact same initial condition, the same arrival process, and by letting each service completion and departure occur at the same time at both systems. We can keep the two systems coupled as long as the queue does not empty. (Once the queues empty, the service-time in the IW system is no longer distributed as $\lceil S + D \rceil$, as was explained in Remark 2, and so the coupling cannot be kept.) However, we know that if $\rho_e = 1$, the queue in the $M_p(t)/G/c$ will empty with probability 1, but the expected time to empty is infinite. Thus, the same holds for the IW. Similarly, when $\rho_e > 1$ there is a positive probability that the $M_p(t)/G/c$ queue never empties, which again implies the same for the IW. $\square$

Proof of Proposition 1. Note that we set $x(T_0) = x_0$. Assume that $z_0 + b_0 \neq c$ and let ν_0 denote the first discontinuity point of Ψ in its argument x. There can be two causes for a discontinuity: either due to a jump at time $T_0 + D$, or due the *discontinuity subspace* $\mathbf{C} := \{x \in \mathbb{R}^3 : z + b = c\}$. Let $\nu_0 := \inf\{t \geq T_0 : z(t) + b(t) = c\} \wedge (T_0 + D)$; by assumption, $\nu_0 > 0$. The function Ψ is clearly Lipschitz continuous in its argument x over the interval $[T_0, \nu_0)$. Since Ψ is also piecewise continuous in t, by our assumption on $\lambda(t)$, there exists a unique solution $x(t)$ to (6) over $[0, \nu_0)$ by, e.g., Theorem 3.2 in Khalil (2002).

Suppose that $\nu_0 < T_0 + D$, so that ν_0 is the hitting time of $\mathbf{C}$. Observing that $dz(t) + db(t) = 0$ and $dq(t) = \lambda(t)dt > 0$ over the interval $[\nu_0, T_0 + D)$, we see that $z(t) + b(t) = c$, implying in turn that Ψ is locally-Lipschitz continuous on $[\nu_0, T_0 + D)$. Hence, there exists a unique solution over this time interval as well. Since the size of the jump at time $T_0 + D$ is b_0, the value of $x(T_0 + D)$ is uniquely determined by $x((T_0 + D)-)$ and $b(0)$. Together with the Lipschitz continuity of Ψ over $[T_0 + D, T_1]$, the uniqueness of the solutions over $[T_0, T_1]$ follows.

If $\nu_0 = D$, then $x(D)$ is determined uniquely by b_0 and $x(D-)$. Letting

$$\nu_0' := \inf\{t > T_0 + D : z(t) + b(t) = c\} \wedge T_1,$$

we see that the same arguments given above give that a unique solution holds on $(\nu_0', T_1]$. Due to the periodicity of $\lambda(t)$, we can treat $x_1 := x(T_1)$ as a new initial condition and repeat the arguments above for the second cycle. The proof follows by induction on the number of cycles by time T. $\square$

We next prove Proposition 2. We note that the dynamics of X^n switch in a discontinuous manner when the service pool fills up and at the discharge epochs. Hence, there is no continuous-mapping representation for X^n over arbitrary finite intervals. Nevertheless, we can exploit the continuity of the dynamics and the monotonicity of the component processes of the vector X^n over the (random) intervals between these switching times to prove the stated convergence.

For $t > 0$, let $\mathcal{D}^3[0,t]$ denote the space of right-continuous functions with left limits over the time interval $[0,t]$ with values in $\mathbb{R}^3$, endowed with the usual Skorohod's J_1 topology; e.g., Whitt (2002).

Proof of Proposition 2. Let all the random variables and processes be defined on a common probability space $(\Omega, \mathcal{F}, P)$. To simplify the arguments, we employ the Skorohod representation theorem, e.g., Theorem 3.2.2 in Whitt (2002), and prove that the convergence to the fluid limit in the statement of the proposition holds w.p.1 in a new probability space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P})$, under the assumption that $\bar{X}^n(0) \to x(0)$ w.p.1 as $n \to \infty$. Our proof will then imply the stated weak convergence result for the original sequence of stochastic processes. However, we do not change the notation of any of the random variables and processes.

Let

$$\mathcal{N}_0^n := \inf\{0 \le t \le 1 : Z^n(t) + B^n(t) = c^n\} \wedge (T_0 + D) \quad \text{and} \quad \nu_0 := \inf\{0 \le t \le 1 : z(t) + b(t) = c\} \wedge (T_0 + D),$$

where, $\inf \phi := \infty$ (ϕ denoting the empty set). We use time-changed independent Poisson processes to represent the sample-paths of X^n over the interval $[0, T_0 + D)$ as in Pang et al. (2007),

$$\begin{aligned}
Z^n(t) &= Z^n(0) + N_a\left(\int_0^t 1_{\{Z^n(s)+B^n(s)<c^n\}}\lambda(s)ds\right) - N_z\left(\mu\int_0^t Z^n(s)ds\right);\\
B^n(t) &= B^n(0) + N_z\left(\mu\int_0^t Z^n(s)ds\right);\\
Q^n(t) &= Q^n(0) + N_a\left(\int_0^t 1_{\{Z^n(s)+B^n(s)=c^n\}}\lambda(s)ds\right); \quad 0 \le t < T_0 + D,
\end{aligned} \tag{24}$$

where N_a and N_z are independent unit-rate Poisson processes.

Let $L^n(t) := Z^n(t) + B^n(t) = (Z^n(0) + B^n(0)) + N_a\left(\int_0^t 1_{\{Z^n(s)+B^n(s)\le c^n\}}\lambda(s)ds\right)$, $t \ge 0$, and observe that L^n is nondecreasing over the interval $[0, \mathcal{N}_0^n)$. If $z(0) + b(0) < c$, then, by the assumed convergence of the initial condition, there exists N_c, such that $L^n(0) < c^n$ for all $n > N_c$. Due to the Functional Strong Law of

Large Numbers (FSLLN) for Poisson processes and the fact that $\mathcal{N}_0^n \to \nu_0$ in $\mathbb{R}$, it holds that $\bar{L}^n := L^n/n \to z(0) + b(0) + \int_0^t \lambda(s)ds$ in $D[0, \nu_0)$, as $n \to \infty$. Hence, for any $\epsilon > 0$ there exits N_ϵ, such that the indicator function in the representation for Z^n in (24) is equal to 1 over $[0, \nu_0 - \epsilon]$ for all $n > N_\epsilon$, so that

$$Z^n(t) = Z^n(0) + N_a\left(\int_0^t \lambda(s)ds\right) - N_z\left(\mu \int_0^t Z^n(s)ds\right), \quad 0 \le t < \nu_0 \text{ for all } n \text{ sufficiently large.}$$

We can therefore apply the version of the Continuous-Mapping Theorem (CMT) in Theorem 4.1 of Pang et al. (2007) over compact subintervals of $[0, \nu_0)$ to obtain the fluid limit for $\bar{Z}^n$ over all such subintervals, from which the fluid limit for $\bar{B}^n$ follows immediately. Since Q^n is trivially null over $[0, \mathcal{N}_0^n)$, we have $\bar{X}^n \to x$ in $D^3[0, \nu_0)$.

To establish the convergence of $\{\bar{X}^n : n \ge 1\}$ beyond ν_0, let $\nu_c := \inf\{0 \le t \le 1 : z(t) + b(t) = c\}$ and assume that $\nu_c < T_0 + D$, so that $\nu_0 = \nu_c$. In this case, the discontinuity in the dynamics of X^n are due to the service process hitting full capacity. Since $\nu_c < (T_0 + D)$, and since $L^n(t)$ is nondecreasing over the interval $I_1 := [\nu_c, T_0 + D)$ for all $n \ge 1$, it holds that $\bar{L}^n \to c$ w.p.1 over I_1 as $n \to \infty$, and so the indicator function in the representation for Z^n is fixed at 0, whereas the indicator function in the representation for Q^n is fixed at 1 for all n large enough over the open interval $(\nu_c, T_0 + D)$. Once again, the CMT gives the uniform convergence of $\bar{X}^n$ to x over $[\nu_0 + \epsilon, T_0 + D - \epsilon]$ for any $\epsilon > 0$. The continuity of x at the point ν_0 implies that we have $\bar{X}^n \to x$ as $n \to \infty$ over the interval $[0, T_0 + D - \epsilon)$. The jump at time $T_0 + D$ is determined by $B^n(T_0)$ which converges to $b(T_0)$ in $\mathbb{R}$ due to the convergence of the process B^n over $[0, T_0 + D - \epsilon)$. Since the jumps in all the elements of the sequence $\{X^n : n \ge 1\}$ occur at exactly the same time $T_0 + D$, and since the sizes of those jumps converge, we have $\bar{X}^n(T_0 + D) \to x(T_0 + D)$ in $\mathbb{R}^3$ as $n \to \infty$, from which it follows that $\bar{X}^n \to x$ in $D^3[0, T_0 + D)$ as $n \to \infty$. Now, if $z(T_0 + D) + b(T_0 + D) < c$, we can apply the same arguments above to conclude that the convergence holds over $[T_0 + D, 1]$, namely, that $\bar{X}^n \to x$ in $\mathcal{D}^3[T_0 + D, 1]$. (The representation of X^n over $[T_0 + D, 1]$ is similar to that in (24), where we replace $X^n(0)$ with $X^n(T_0 + D)$.) Moreover, the same arguments can be applied when $\nu_c \ge T_0 + D$, in which case the jump due to discharge occurs before a queue builds up in $\bar{X}^n$ for all n large enough. (The only difference between the former and the latter case is that, if $z(T_0 + D) + b(T_0 + D) < c$, then there will be no queue at the jump epoch in system n for all n large enough.)

So far, the stated convergence to the fluid limit was proved under the assumption that $z(0) + b(0) < c$ and $z(T_0 + D) + b(T_0 + D) < c$. If $z(0) + b(0) = c$, then it may hold that $Z^n(0) + B^n(0) < c^n$ for infinitely-many

n's, in which case **it is not true** that $1_{\{Z^n(0)+B^n(0)<c^n\}} = 0$ and $1_{\{Z^n(0)+B^n(0)=c^n\}} = 1$ for all n large enough. However, the facts that $Z^n(0) + B^n(0) - c^n = o(n)$ and that the service process L^n is increasing at a rate $\lambda^n(0) = \Theta(n)$ at time 0, imply that, for any $\delta > 0$, there exists an n_δ, such that $Z^n(s) + B^n(s) = c^n$ for all $s \in [\delta, T_0 + D)$ and for all $n \geq n_\delta$ [4]. Hence, the convergence of $\bar{X}^n$ to x now holds in $D^3[\delta, T_0 + D)$ for all sufficiently small $\delta > 0$. The right continuity of the limit x, together with the fact that the fluid service process $z + b$ is strictly increasing over $[0, \delta]$ implies that the convergence to the fluid limit can be extended to time 0, i.e., we have $\bar{X}^n \to x$ in $D^3[0, T_0 + D)$ as $n \to \infty$. We similarly deal with case $z(T_0 + D) + b(T_0 + D) = c$ to conclude the convergence to the fluid limit over $[T_0 + D, 1]$.

In particular, we have established that $\bar{X}^n \to x$ in $D^3[0, 1]$ as $n \to \infty$. We can then take $\bar{X}^n(1)$ as the new "initial condition" (shifting the time axis by one unit of time to the left), and apply the same proof of convergence over the time interval $[1, 2]$. Continuing inductively, we conclude that $\bar{X}^n$ converges uniformly to the fluid limit x over any interval $[0, t]$, $0 < t < \infty$. Finally, the strong convergence just established on the probability space $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P})$ implies the weak convergence in the statement of the proposition over the original probability space $(\Omega, \mathcal{F}, P)$. $\square$

For the proof of Theorem 2 it will be convenient to rewrite the stability condition $\rho_e < 1$, for ρ_e in (3), as follows.

$$\Lambda + \frac{\Lambda e^{-\mu(1-D)}}{1 - e^{-\mu}} < c.$$

Proof of Theorem 2. We start by showing that a periodic equilibrium exists. To simplify the notation, we let $x^* = x^*(T_k)$ denote the value of the periodic equilibrium we consider at the inspection epochs immediately after the jump. When the system is in equilibrium, the rate into the system during the period must equal the rate out of the system; this, together with Lemma 1, implies that $b^* = \Lambda$. Let

$$\Delta^* := \inf \left\{ t \geq 0 : z^* + q^* + \int_0^t \lambda(s) ds \geq c \right\}. \tag{25}$$

Note that, provided a periodic equilibrium exists, Δ^* is well defined (with $\Delta^* = 0$ if $z^* + q^* \geq c$) since $\lambda(s)$ is positive and periodic (so that its support is infinite).

We next show that a solution to the balance equations must exists when $\rho_e < 1$, implying that a periodic equilibrium indeed exists. We divide the analysis into three cases, based on the value of Δ^*.

Case (I): $\Delta^* > 1$. In this case (recall that $b^* = \Lambda$), $z^* + q^* + b^* = z^* + q^* + \Lambda < c$, so that $q^* = 0$. To compute the value of z^*, we consider the two following possible subcases.

Case (Ia): $\Delta^* \geq 1+D$ In this case there is no queue at all during the whole cycle in equilibrium. Thus,

$$z^* = z^* e^{-\mu} + \int_0^1 \lambda(s) e^{-\mu(1-s)} ds, \quad \text{so that} \quad z^* = \frac{\int_0^1 \lambda(s) e^{-\mu(1-s)} ds}{1-e^{-\mu}}.$$

Observe that, indeed, $b^* = z^*(1-e^{-\mu}) + \int_0^1 \lambda(s)(1-e^{-\mu(1-s)})ds = \Lambda$.

Case (Ib): $\Delta^* < 1+D$ In this case, a queue forms some time between T_k and $T_k + D$ during the kth cycle in equilibrium. Then

$$z^* = z^* e^{-\mu} + \int_0^{\Delta^*-1} \lambda(s) e^{-\mu(1-s)} ds + \int_{\Delta^*-1}^{D} \lambda(s) e^{-\mu(1-D)} ds + \int_D^1 \lambda(s) e^{-\mu(1-s)} ds,$$

so that

$$z^* = \frac{\int_0^1 \lambda(s) e^{-\mu(1-s)} ds + \int_{\Delta^*-1}^{D} \lambda(s) \left(e^{-\mu(1-D)} - e^{-\mu(1-s)} \right) ds}{1-e^{-\mu}}.$$

Note further, that

$$b^* = z^*(1-e^{-\mu}) + \int_0^{\Delta^*-1} \lambda(s)(1-e^{-\mu(1-s)})ds + \int_{\Delta^*-1}^{D} \lambda(s)ds(1-e^{-\mu(1-D)}) + \int_D^1 \lambda(s)(1-e^{-\mu(1-s)})ds = \Lambda.$$

Case (II): $D < \Delta^* \leq 1$. In this case, the system becomes full at some time between $T_k + D$ and T_{k+1} in the kth cycle in equilibrium, and necessarily remains full until $T_{k+1} + D$. In this case, $z^* = c - \Lambda$ and $q^* = \int_{\Delta^*}^1 \lambda(s)ds$. We also notice that

$$b^* = z^*(1-e^{-\mu}) + (q^* + \int_0^D \lambda(s)ds)(1-e^{-\mu(1-D)}) + \int_D^{\Delta^*} \lambda(s)(1-e^{-\mu(1-s)})ds.$$

Plugging $b^* = \Lambda$, we obtain

$$\Lambda = (c-\Lambda)(1-e^{-\mu}) + (\int_{\Delta^*}^1 \lambda(s)ds + \int_0^D \lambda(s)ds)(1-e^{-\mu(1-D)}) + \int_D^{\Delta^*} \lambda(s)(1-e^{-\mu(1-s)})ds.$$

By rearranging the last equation, we have

$$\Lambda + \frac{\Lambda e^{-\mu(1-D)}}{1-e^{-\mu}} + \frac{\int_D^{\Delta^*} \lambda(s) \left(e^{-\mu(1-s)} - e^{-\mu(1-D)} \right)}{1-e^{-\mu}} = c.$$

Case (III): $0 \leq \Delta^* \leq D$. We now show that this is in fact impossible. Indeed, assuming that $\Delta^* \leq D$ gives

$$b^* = z^*(1-e^{-\mu}) + b^*(1-e^{-\mu(1-D)}),$$

because the IW fills up immediately after the discharge. Plugging $b^* = \Lambda$ and $z^* = c - \Lambda$ gives

$$\Lambda = (c-\Lambda)(1-e^{-\mu}) + \Lambda(1-e^{-\mu(1-D)}),$$

and, after rearrangement of the terms, we obtain $\Lambda = c(1-e^{-\mu})/\left(1+e^{-\mu(1-D)}-e^{-\mu}\right)$, contradicting the assumed stability condition $\rho_e < 1$. Thus, Case (III) *cannot hold* under the stability condition.

Since x^* is defined uniquely in terms of Δ^*, there exists a unique periodic equilibrium that empties daily. Since by Lemma 1 any periodic equilibrium must empty daily, the uniqueness of the periodic equilibrium follows. The fact that $z^*(T_i+D)+b^*(T_i+D) < c$ follows from the fact that $\Delta^* > D$, since only Cases (I) and (II) can hold by the above. □

REMARK 5. A clear computational difficulty with Δ^* is that it is defined in terms of the periodic equilibrium. If Δ^* is greater than 1, then there is no need to compute it in order to to solve for the periodic equilibrium. For the case $D < \Delta^* < 1$ (which is the only other case to consider, according to the proof of Theorem 2, we now show how Δ^* can be computed from the basic model parameters.

Define Δ^{**} to be the unique solution to the following equation.

$$\Lambda + \frac{\Lambda e^{-\mu(1-D)}}{1-e^{-\mu}} + \frac{\int_D^{\Delta^{**}} \lambda(s)\left(e^{-\mu(1-s)} - e^{-\mu(1-D)}\right) ds}{1-e^{-\mu}} = c.$$

Observe that under the stability condition $\rho_e < 1$, $\Lambda + \frac{\Lambda e^{-\mu(1-D)}}{1-e^{-\mu}} < c$, so that Δ^{**} is well defined. It is easy to check that if $\Delta^{**} > 1$, then $\Delta^* > 1$ as well, whereas, if $D < \Delta^{**} \le 1$, then $\Delta^{**} = \Delta^*$.

Proof of Theorem 3. Let $\{x(t) : t \ge 0\}$ be the trajectory of the fluid model when the initial condition is $x(0)$, and consider the sequence $x_k := x(T_k)$, $k \ge 0$. The arguments in the proof of Theorem 5 (for the case $\rho_e < 1$) show that $\|x_k\| \le cM$, for all $k \ge 0$, where M is defined in (21) and $\|x\| := \sum_{i=1}^3 |x_i|$. Hence, the sequence $x_k := \{x_k : k \ge 1\}$ is precompact in $\mathbb{R}^3$. Take a converging subsequence $\{x_{k_r} : r \ge 0\}$ of x_k and let $\bar{x}_\ell$ be its limit. Now, if $\bar{x}(t)$ is a trajectory with $\bar{x}(T_0) = \bar{x}_\ell$, then since $\bar{x}(T_k) = \bar{x}_\ell$ for all $k \ge 0$, $\bar{x}(t)$ must be a periodic equilibrium. The uniqueness of the periodic equilibrium u^*, shown in Theorem 2, implies that $\bar{x}(T_0) = u^*(T_0)$, so that $\bar{x}(t) = u^*(t)$, $t \ge 0$.

To prove (9), consider an almost-everywhere continuous function $f : \mathbb{R}^3 \to \mathbb{R}$. Then

$$\lim_{t\to\infty} \frac{1}{t}\int_0^t f(x(s))ds = \lim_{t\to\infty} \frac{1}{t}\left(\sum_{i=0}^{\lfloor t\rfloor}\int_i^{i+1} f(x(s))ds + \int_{\lfloor t\rfloor}^t f(x(s))ds\right), \tag{26}$$

where $\lfloor t\rfloor$ denotes the integer part of t, i.e., the largest integer that is smaller than t. Observe that if $1/t\int_{t-1}^t f(x(s))ds \to 0$ as $t\to\infty$, then the remainder term in the RHS of (26) converges to 0 as $t\to\infty$. Furthermore, (8) implies that for any $\epsilon > 0$, we can find a positive integer K_ϵ such that $|f(x(s)) - f(u^*(s))| < \epsilon$ for almost all $s \ge K_\epsilon$. Hence, for all $t > K_\epsilon$,

$$\frac{1}{t}\sum_{i=0}^{\lfloor t\rfloor}\int_i^{i+1} f(x(s))ds = \frac{1}{t}\sum_{i=0}^{K_\epsilon}\int_i^{i+1} f(x(s))ds + \frac{1}{t}\sum_{i=K_\epsilon}^{\lfloor t\rfloor}\int_i^{i+1} f(x(s))ds. \tag{27}$$

Since the value of K_ϵ does not depend on t, the first argument in the RHS of the equality (27) converges to 0 as $t \to \infty$, whereas

$$\int_0^1 f(u^*(s))ds - \epsilon \leq \frac{1}{t} \sum_{i=K_\epsilon}^{\lfloor t \rfloor} \int_i^{i+1} f(x(s))ds \leq \int_0^1 f(u^*(s))ds + \epsilon,$$

and (9) follows by taking $\epsilon \downarrow 0$. $\square$

Proof of Corollary 1. The statement of the corollary follows from Proposition 2 and Theorems 2 and 3. In particular, by Theorem 3 the fluid model converges to its unique periodic equilibrium for any initial condition. Since the queue empties daily in equilibrium, it holds that, for any $x(0)$, there exists a K such that $\tau_{i+1} - \tau_i = 1$ for all $i \geq K$. Now, the proof of Theorem 2 shows that $q^*((T_k + D)-) = 0$ in Case (Ia); $q^*((T_k + D)-) = \int_{\delta^*}^D \lambda(t)dt$, where $\delta^* \in (0, D]$, in Case (Ib); or $q^*((T_k + D)-) = \int_{\Delta^*}^{1+D} \lambda(t)dt$, where $\Delta^* \in (D, 1]$, in Case (II). In all cases, $q^*((T_k + D)-) < b^*(T_k) = \Lambda$, so that there is idleness in fluid scale for some time period after the discharge. The convergence of the fluid-scaled sequence of stochastic systems $\bar{X}^n(t)$ is equivalent to convergence in probability because the fluid limit is deterministic, and the result follows. $\square$

Proof of Theorem 4 Consider the discrete time process w_k. For $k \geq 0$, let $\Delta_k := \inf\left\{t \geq 0 : w_k + \int_0^t \lambda(s)ds \geq c\right\}$. Then $w_{k+1} = w_0 + \Lambda - (w_0 \wedge c)(1 - e^{-\mu}) - \int_0^{\Delta_0 \wedge 1} \lambda(s)\left(1 - e^{-\mu(1-s)}\right) ds$. We first show that, since the period of the equilibrium is equal to one (day), it must hold that $w^* < c$. We prove this by contradiction. Suppose in equilibrium $w^* \geq c$, so that $w_k = w^* \geq c$ whenever $w_0 = w^*$, $k \geq 1$. Then

$$w_k = w_{k+1} = w_k + \Lambda - (w_0 \wedge c)(1 - e^{-\mu}) - \int_0^{\Delta_k \wedge 1} \lambda(s)\left(1 - e^{-\mu(1-s)}\right) ds = w_k + \Lambda - c(1 - e^{-\mu}) < w_k,$$

leading to a contradiction. Hence, in equilibrium, $w_k = w^* < c$.

We next establish the rates of convergence to equilibrium, depending on the initial condition w_0. We divide the analysis into three cases

i) When $w_0 \geq c$,

$$w_1 = w_0 + \Lambda - c(1 - e^{-\mu}) \quad \text{so that} \quad w_1 - w^* = w_0 - w^* + \Lambda - c(1 - e^{-\mu})$$

As $w_0 \geq c$ and $w^* = w^* e^{-\mu} + \Lambda - \int_0^{\Delta^* \wedge 1} \lambda(s)\left(1 - e^{-\mu(1-s)}\right) ds$ from (11),

$$w_0 - w^* + \Lambda - c(1 - e^{-\mu}) \geq c + \int_0^{\Delta^* \wedge 1} \lambda(s)\left(1 - e^{-\mu(1-s)}\right) ds - w^* e^{-\mu} - c(1 - e^{-\mu}) \geq (c - w^*)e^{-\mu} \geq 0.$$

Then we have $w_1 - w^* > 0$ and $|w_1 - w^*| = |w_0 - w^*| - (c(1 - e^{-\mu}) - \Lambda)$. Repeating this argument as long as $w_k > c$, we obtain that w_k decreases at rate $(c(1 - e^{-\mu}) - \Lambda) < 0$ to the set $[0, c)$ on the (discrete) time interval $\{0, 1, \ldots, K_c\}$, where $K_c := \min\{k \geq 1 : w_k \leq c\}$.

ii) When $w^* < w_0 < c$,

$$w_1 = w_0 + \Lambda - w_0(1-e^{-\mu}) - \int_0^{\Delta_0 \wedge 1} \lambda(s)\left(1-e^{-\mu(1-s)}\right) ds = w_0 e^{-\mu} + \Lambda - \int_0^{\Delta_0 \wedge 1} \lambda(s)\left(1-e^{-\mu(1-s)}\right) ds$$

As $w^* = w^* e^{-\mu} + \Lambda - \int_{\Delta^* \wedge 1}^1 \lambda(s)\left(1-e^{-\mu(1-s)}\right) ds$ and $w_0 - w^* = \int_{\Delta_0}^{\Delta^*} \lambda(s) ds \geq \int_{\Delta_0 \wedge 1}^{\Delta^* \wedge 1} \lambda(s) ds$,

$$\begin{aligned} w_1 - w^* &= (w_0 - w^*)e^{-\mu} + \int_{\Delta_0 \wedge 1}^{\Delta^* \wedge 1} \lambda(s)\left(1-e^{-\mu(1-s)}\right) ds \\ &\leq (w_0 - w^*)e^{-\mu} + \left(1-e^{-\mu(1-\Delta_0)}\right) \int_{\Delta_0 \wedge 1}^{\Delta^* \wedge 1} \lambda(s) ds \\ &= (w_0 - w^*)\left(1 + e^{-\mu} - e^{-\mu(1-\Delta_0)}\right). \end{aligned}$$

As $0 < 1 + e^{-\mu} - e^{-\mu(1-\Delta_0)} < 1$, then $|w_1 - w^*| \leq |w_0 - w^*|\left(1 + e^{-\mu} - e^{-\mu(1-\Delta_0)}\right)$. Indeed, we have for $k \geq 1$,

$$|w_k - w^*| \leq |w_{k-1} - w^*|\left(1 + e^{-\mu} - e^{-\mu(1-\Delta_k)}\right) < |w_0 - w^*|\left(1 + e^{-\mu} - e^{-\mu(1-\Delta_0)}\right)^k$$

iii) When $w_0 < w^*$, following the same line of analysis as in case ii), we have

$$\begin{aligned} w^* - w_1 &= (w^* - w_0)e^{-\mu} + \int_{\Delta^* \wedge 1}^{\Delta_0 \wedge 1} \lambda(s)\left(1-e^{-\mu(1-s)}\right) ds \\ &\leq (w^* - w_0)e^{-\mu} + \left(1-e^{-\mu(1-\Delta^*)}\right) \int_{\Delta^* \wedge 1}^{\Delta_0 \wedge 1} \lambda(s) ds \\ &= (w^* - w_0)\left(1 + e^{-\mu} - e^{-\mu(1-\Delta^*)}\right), \end{aligned}$$

so that $|w_k - w^*| < |w_0 - w^*|\left(1 + e^{-\mu} - e^{-\mu(1-\Delta^*)}\right)^k$.

To summarize, when $w_0 < w^*$ (case iii), w_k increases to w^* at exponential rate, i.e. $|w_k - w^*| < |w_0 - w^*|\left(1 + e^{-\mu} - e^{-\mu(1-\Delta^*)}\right)^k$; when $w^* < w_0 < c$ (case ii), w_k decreases to w^* at exponential rate, i.e. $|w_k - w^*| < |w_0 - w^*|\left(1 + e^{-\mu} - e^{-\mu(1-\Delta_0)}\right)^k$; when $w_0 > c$ (case i), w_k first decreases to some level below c at a linear rate, i.e. $|w_1 - w^*| = |w_0 - w^*| - (c(1-e^{-\mu}) - \Lambda)$, and then follows the trajectory of case ii. □

References

ACEP (2008) Task force report on boarding: Emergency department crowding: High-impact solutions.

Afeche P (2013) Incentive-compatible revenue management in queueing systems: optimal strategic delay. *Manufacturing & Service Operations Management* 15(3):423–443.

Argo JL, Vick CC, Graham LA, Itani KM, Bishop MJ, Hawn MT (2009) Elective surgical case cancellation in the veterans health administration system: identifying areas for improvement. *The American Journal of Surgery* 198(5):600–606.

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

Asplin BR, Magid DJ (2007) If you want to fix crowding, start by fixing your hospital. *Annals of emergency medicine* 49(3):273–274.

Bailey N (1954) On queueing processes with bulk service. *Journal of the Royal Statistical Society. Series B (Methodological)* 80–87.

Banner Health (2015) Door-to-doc patient safety toolkit. `https://www.bannerhealth.com/About+Us/Innovations/DoortoDoc/About+D2D.htm`.

Best TJ, Sandikci B, Eisenstein DD, Meltzer D (2015) Managing hospital inpatient bed capacity through partitioning care into focused wings. *Manufacturing & Service Operations Management* 17(2):157–176.

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/`.

Centers for Medicare and Medicaid Services (2018) National heath expenditure data. `http://www.cms.gov`.

Chan C, Dong J, Green L (2017) Queues with time-varying arrivals and inspections with applications to hospital discharge policies. *Operations Research* 65(2):469–495.

Chan C, Farias VF, Bambos N, Escobar G (2012) Optimizing intensive care unit discharge decisions with patient readmissions. *Operations Research* 60(6):1323–1341.

Dai J, Shi P (2017) A two-time-scale approach to time-varying queues in hospital inpatient flow management. *Operations Research* 65(2):514–536.

Dai J, Shi P (2018) Inpatient bed overflow: An approximate dynamic programming approach. *Manufacturing and Service Operations Managemen* Forthcoming.

Foss S, Konstantopoulos T (2006) Lyapunov function methods. Lecture Notes.

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

Green L, Savin S (2008) Reducing delays for medical appointments: a queueing approach. *Operations Research* 56(6):1526–1538.

Hall R, Belson D, Murali P, Dessouky M (2006) Modeling patient flows through the healthcare system. Hall R, ed., *Patient Flows: Reducing Delay in Healthcare Delivery* (Springer).

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

Kc DS, Terwiesch C (2009) Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Science* 55(9):1486–1498.

Keehan S, Sisko A, Truffer C, Smith S, Cowan C, Poisal J, Clemens MK, the National Health Expenditure Accounts Projections Team (2008) Health spending projections through 2017: the baby-boom generation is coming to medicare. *Health Affairs* 27(2):w145–w155.

Khalil HK (2002) *Nonlinear Systems* (Prentice hall New Jersey), 3 edition.

Kim S, Vel P, Whitt W, Cha WC (2015) Poisson and non-poisson properties in appointment-generated arrival processes: The case of an endocrinology clinic. *Operations Research Letters* 43(3):247–253.

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

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

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

Maglaras C, Yao J, Zeevi A (2013) Optimal price and delay differentiation in queueing systems Available at SSRN 2297042.

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

McGowan JE, Truwit JD, Cipriano P, Howell RE, VanBree M, Garson A, Hanks JB (2007) Operating room efficiency and hospital capacity: factors affecting operating room use during maximum hospital census. *Journal of the American College of Surgeons* 204(5):865–871.

Medicare (2015) Linking quality to payment. `https://www.medicare.gov/hospitalcompare/linking-quality-to-payment.html?AspxAutoDetectCookieSupport=1`.

Pang G, Perry O (2014) A logarithmic safety staffing rule for contact centers with call blending. *Management Science* 61(1):73–91.

Pang G, Talreja R, Whitt W, et al. (2007) Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Probability Surveys* 4:193–267.

Perry O, Whitt W (2016) Chattering and congestion collapse in an overload switching control. *Stochastic Systesms* 6(1):132–210.

Puhalskii AA (2013) On the $M_t/M_t/K_t + M_t$ queue in heavy traffic. *Mathematical Methods of Operations Research* 78(1):119–148.

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

Richardson D (2002) The access-block effect: relationship between delay to reaching an inpatient bed and inpatient length of stay. *The Medical Journal of Australia* 177(9):492–495.

Service Engerprise Engineering Center (2012) SEEStat Online. http://ie.technion.ac.il/Labs/Serveng/.

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

Song H, Tucker A, Murrell K (2015) The diseconomies of queue pooling: An empirical investigation of emergency department length of stay. *Management Science* 61(12):3032–3053.

Trzeciak S, Rivers EP (2003) Emergency department overcrowding in the united states: an emerging threat to patient safety and public health. *Emergency medicine journal* 20(5):402–405.

Whitt W (2002) *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues* (Springer).