Determining the optimal configuration of hospital inpatient rooms in the presence of isolation patients

Edieal Pinker
Yale University, School of Management
edieal.pinker@yale.edu

Tolga Tezcan
University of Rochester, Simon Graduate School of Business
tolga.tezcan@simon.rochester.edu

We study the optimal configuration of hospital in-patient rooms with private and semi-private rooms when some of the patients have infectious diseases and need to be isolated. We assume that the demand is random and seasonal. We propose a computationally efficient solution procedure that is based on a stochastic program that uses asymptotic approximations for the system performance under different admission policies and show its accuracy for large systems. Using our model we study the appropriateness of the recent trends in hospital design calling for 100% private rooms. We show that even with isolation patients such an extreme approach could result in a significant degradation in the access of patients to hospital beds.

Key words: Hospital bed configuration, queueing, admission control, hospital operations

1. Introduction

When building or renovating a hospital, the number of beds per room is an important choice faced by designers. Currently it is common to find both private and semi-private rooms (two beds) but the trend is toward private rooms (Dowdeswell et al. (2004); Jones and Thomas (2004); Detsky and Etchells (2008)). The relative merits of both types of room are thoroughly discussed in the review papers (Chaudhury et al. (2005a,b), van de Glind et al. (2007), and Boardman and Forbes (2011)) and we summarize them here. There is clinical evidence that private rooms reduce infection rates. It is also believed that patients get less noise and disturbances in private rooms and therefore get better rest. The privacy of private rooms facilitates communication between the care providers and the patient and their family. Private rooms accommodate family visits better which may also aid the recovery. Finally patient satisfaction is reported to be higher in private rooms. On the other hand, semi-private rooms take less space and are thus cheaper to construct. It is also believed that because of the way semi-private and private rooms are laid out semi-private rooms require nursing staff to walk shorter distances than in a floor plan with primarily private rooms. Semi-private
rooms are also believed to support greater supervision of patients. The analyses in the literature generally conclude that private rooms are superior although it is also agreed that much of the supporting evidence lacks scientific rigor.

In this paper we focus our attention on one important difference between private and semi-private rooms that has not been well studied in the literature; how they are affected by known infectious patients. When a patient with an infectious disease is admitted to a hospital that patient cannot share a room with another patient because of transmittal risk. As a result, in a setting with semi-private rooms an infectious patient in effect uses two beds whereas if they could be placed in a private room only one bed is used. It has been claimed in the literature (Bobrow and Thomas 2000, p. 150) that the occupancy of multi-bed rooms reaches a maximum of 80 to 85 percent, whereas single-bed rooms can reach 100 percent occupancy. Thus even if semi-private rooms are cheaper to construct per bed they are not able to generate the same revenue per bed because of infectious patients. While several other papers have cited these figures we have not found any substantiation of them in either Bobrow and Thomas (2000) or elsewhere. Yet, the logic behind the claim is valid and must be taken into account when evaluating the relative merits of private and semi-private rooms. Vayalumkal et al. (2007) found that 17.1% of all patients in a children hospital were isolated with the isolation rate varying from 5.7% to 39.2% across wards. We expect that these rates are lower for adult patients but we can see that the need for isolation can be significant and thus it makes sense to consider it in capacity configuration decisions. We also note that there are noninfectious patients who prefer private rooms. These patients may pay more for private rooms and the inability to serve them will lead to lost revenue.

The debate about the room types is often framed in terms of beds. I.e. it is argued that the same number of beds is cheaper to construct as semi-private but the revenue generated per bed is higher if they are private. Proponents of private rooms claim that over time the extra revenue generated per bed far exceeds the additional cost of private room beds. We think this framing of the comparison of bed configurations is problematic because it ignores access issues. The number of beds is the wrong quantity to hold constant in the analysis. An alternative perspective is to assume that there is a fixed construction budget or fixed space for the facility and we must decide how best to configure it. In the same space one can fit more semi-private beds than private beds and as a result the theoretical capacity of a facility configured with semi-private rooms is higher than that of one configured with only private rooms. The need to isolate infectious patients reduces the capacity advantage of semi-private rooms.
Our approach here is to model a department of a hospital with two patient types, type 1 patients, who need to be isolated and type 2 patients that can share a room with another patient and two room types; private and semi-private. We consider non-infectious patients who demand private rooms as part of type 1. These two types of patients will each have different arrival rates with seasonal changes. We allow each patient type to have its own service rate, and revenue per visit. We study the problem of maximizing expected revenue when the hospital has a fixed space to configure using private and semi-private rooms. Maximizing the expected number of patients treated (access) is a special case. The health policy literature on the subject has not taken into account the existence of these two types of patients or the seasonality of arrivals, in their analyses. It is widely acknowledged that there is seasonality in infectious diseases (Fisman (2007), Pascual and Dobson (2005)) a well known example is influenza but it exists for non-infectious hospitalizations as well, including acute myocardial infarctions (Fischer et al. (2004), Frederick et al. (1998)). Seasonality in patient arrival rates can cause the bed configuration to work better during some parts of the year than others. Therefore we explicitly include seasonal arrival patterns in the analysis.

In order to solve the optimal configuration problem, one needs to determine the patient admission procedure. Admission decisions may have a significant impact on the revenue and admission control problems are well studied in the queueing literature, see Ulukus et al. (2011), Stidham (2002) and the references therein. For example, it may be beneficial to deny admission to an isolation patient when there are only a few semi-private rooms available, anticipating the arrival of noninfectious patients. We focus on two admission policies, FCFS-T and REVMAX$_P$. Under the FCFS-T policy, an arriving patient is admitted as long as there is sufficient number of beds to accommodate that patient, hence “FCFS”, possibly after transferring previously admitted patients to another room to best utilize the rooms hence the “-T”. The REVMAX$_P$ policy’s goal is to maximize revenue.

A crucial step in our analysis is approximating the performance of the FCFS-T admission policy in steady state. Our approximation procedure is based on a many-server asymptotic analysis of loss systems. Loss systems with multiple customer classes and a homogeneous server pool are analyzed in Hunt and Kurtz (1994), Kelly (1986) and Reiman (1991) under FCFS. (A similar model is used in the literature to study the bin packing service systems, see Gamarnik (2004), Coffman and Stolyar (2001), where the resources are assumed to be homogeneous as well.) In a many-server regime Kelly (1986) and Reiman (1991) establish the limits of the admission probabilities for these systems. Their analyses are based on establishing the limits of the steady state probabilities of loss systems, which are available in closed form. However, to the best of our knowledge, such analytical formulas are not available for loss systems with two or more resource pools, e.g. two
types of rooms. Hunt and Kurtz (1994) establish a many-server fluid limit for these systems. In this paper, we devise an approximation method by using the many-server fluid limits in Hunt and Kurtz (1994) and the properties of the limits of the systems we consider. We refine the fluid limit approximations by using a one dimensional Markov chain to approximate the expected number of idle beds in steady state as well the admission probabilities for each patient type. The transitions rates of this approximating Markov chain are based on the steady state quantities obtained from the fluid limits. A similar approach is used in Perry and Whitt (2009) and Perry and Whitt (2011) to build approximations for a different system.

We proceed as follows:
1. We formulate an approximation of the capacity configuration problem as a stochastic optimization problem.
2. We propose an admission policy REVMAX$P$ for revenue maximization and construct approximations for the expected revenue given a configuration for REVMAX$P$ and FCFS-T.
3. We show that with the capacity found using the associated stochastic optimization problem, REVMAX$P$ is asymptotically optimal for large systems.
4. We verify using numerical experiments that the solutions found using our proposed stochastic optimization method generate revenues that are very close to the actual optimal revenue (found by exhaustive simulation experiments) under REVMAX$P$ as well as under FCFS-T.
5. We carry out an extensive numerical experiment to study the effect of using FCFS-T vs. REVMAX$P$ and the effect of using only one type of rooms on the revenue. We find that an optimal mix of private and semi-private rooms can be significantly superior to using all of one type. We also find that a FCFS-T admission policy performs well relative to REVMAX$P$ when the optimal bed configurations are used.
6. Finally we conclude with our assessment of the arguments for private rooms.

2. Model
To put our model in context it is useful to describe the typical arrival or admission process at a hospital. Admissions are broadly categorized as emergency or elective. Emergency patients come from the Emergency Department (ED) and when hospital wards are full the ED will go on diversion which means that ambulances are diverted to other hospitals that have the ability to admit patients to inpatient beds. Elective admissions are more of a mix of sub-categories. These include patients who are having scheduled procedures which can be of different levels of urgency. These also include transfers from other hospitals or nursing homes. Direct admissions to the hospital by physicians are
classified as elective but can be viewed as emergency cases that bypassed the ED. When hospital beds are scarce, more urgent procedures will be scheduled at other hospitals, transfers will be sent to other hospitals etc. Some of these elective admissions will be delayed and arrive later.

We model a single department or ward of a hospital (also known as a nursing unit) as a loss system with two patient types and two room types. Because it is a loss system we ignore waiting and delayed arrivals that might occur as a result of congestion. Type 1 patients, referred to as isolation patients, are infectious and need to be isolated and thus can not share a room with another patient. Type 2 patients, referred to as non-isolation patients, are those who can share a room with another non-isolation patient. The first type of room is private with only one bed per room, the second type is semi-private with two beds per room and can accommodate one isolation patient or two non-isolation patients at a time. We assume that length of stay (LOS) is exponentially distributed. To account for the possibility that isolation patients may differ from non-isolation patients, in their LOS, and given the claims that patients recuperate better in private rooms we assume that a type $i$ patient in a type $j$ room has care completion rate $\mu_{ij}$, $i=1,2$. Naturally we set $\mu_{11} = \mu_{12}$ because a type 1 patient in a type 2 room is effectively in a private room. (We consider additional LOS distributions in our numerical experiments.)

We also assume that different types of patients generate different revenue for the hospital and this may also differ by room type. It is challenging to model revenue streams in a hospital because the payment mechanisms in the U.S. healthcare system are quite complex. The revenue generated by a patient varies not only by the care they receive but also by the type of insurance coverage they have. Insurers may differ by the rates they pay but also by whether they pay a fixed fee that is a certain degree independent of how long the patient ends up spending in the hospital, i.e. the hospital bears the risk of uncertainty in the recovery time. If the insurer is paying based upon fee for service then they will be paying for every day it is deemed necessary for the patient to be in the hospital. Some insurers may pay more for stays in private rooms even when not medically necessary but many, if not most, won’t. If a patient desires a private room even if they do not require isolation they may be charged more and pay the difference between the charge and the insurance coverage out of pocket. It is not possible to represent every patient diagnosis and insurance pair as a distinct patient type and expect to derive any insights so we aggregate patients into the two types described above. In terms of revenue generation this requires us to average across many types of patients. For isolation patients (Type 1) we know that they require private rooms or are the sole occupants of a semi-private room. In either case we can assume that such a patient generates an average revenue of $r_{11} = r_{12}$ across all patient diagnoses and lengths of stay and insurance policies. For non-isolation
(Type 2) patients we must take into account the fact that some will prefer and seek out private rooms even though they are not medically necessary. Therefore Type 2 patients who are boarded in private rooms will on average yield an expected revenue $r_{21}$ that could be higher than that generated in a semi-private room $r_{22}$. As we did for the Type 1 patients we assume that the $r_{2j}$ are independent of LOS because we are averaging revenues across all diagnoses, insurance policies, and LOS. This model of revenues is less accurate if patients are transferred between private and semi-private rooms, which we assume happens in the policy FCFS-T that we describe and analyze in more detail later. If one is interested only in total throughput, this can be accommodated by setting $r_{1j} = r_{2j}$, $j = 1, 2$. Throughout the paper we set $i' = 1$ if $i = 2$ and $i' = 2$ if $i = 1$ for notational simplicity.

When an isolation patient arrives and there is no completely empty room available, private or semi-private, that patient cannot be admitted. We assume that the non-isolation patients may be admitted as long as there is at least one bed available, in a private or a semi-private room. We comment on other cases in Remark 2. Even if there are beds available, the hospital may choose to turn away the patient. In such cases, we assume that patient is diverted to another hospital and so there is no queuing.

We analyze the capacity and the configuration problem over a finite time horizon under time dependent (or seasonal) demand. Let time $t = 0$ represent the start of the planning period and time $t = T$ denote its end. Similar to Harrison and Zeevi (2005), we assume that arrival rates follow a stochastic process $\Lambda = (\Lambda(t) : 0 \leq t \leq T)$ taking values in $\mathbb{R}_+^2$, and given that $\Lambda(t) = \lambda = (\lambda_1, \lambda_2)$, the conditional distribution of arrivals in patient types 1 and 2, respectively, immediately after time $t$ is that of independent Poisson processes with average arrival rates $(\lambda_1, \lambda_2)$ respectively $0 \leq t < T$.

The number of arrivals $A_i(t)$ to class $i$ by time $t$ is given by

$$A_i(t) = L_i \left( \int_0^t \Lambda_i(s) ds \right),$$

where $L_i$’s are independent Poisson processes with rates 1, for $i = 1, 2$.

To formulate the configuration problem, let $c_i$ denote the amount of space required (or cost) per bed of type $i$. For example in Boardman and Forbes (2011) it is demonstrated based on Canadian practices that $c_1/c_2$ is around 1.6 if only room space is considered and it is around 1.74 if in addition the difference in required hall space is also take into account. Based on the floor area per bed of each room type in ten different hospital floors, Chaudhury et al. (2005b) reports that $c_1/c_2$ is around 1.5.
Our objective is to maximize revenue and the decisions are the capacity configuration \( N = (N_1, N_2) \), where \( N_i \) is the number of beds in type \( i \) rooms, \( i = 1, 2 \), and the admission policy \( \xi \). Let \( D_{ij}(t) \) denote the number of type \( i \) patients whose treatment is completed in a type \( j \) room by time \( t \), \( i, j = 1, 2 \). Obviously, \( D \) depends on the admission policy \( \xi \) being used. We restrict ourselves to Markovian admission policies denoted by \( \mathcal{P} \) (specifically we assume that admission and transfer decisions are made only when a patient arrives or departs from the system). The expected revenue under the policy \( \xi \) for a fixed capacity configuration \( N \) by time \( T \) is given by

\[
R_\xi(T, \Lambda, N) = \sum_{i,j=1}^{2} r_{ij} E[ D_{ij}(T) ] .
\]

Let \( B \) denote the total space (or budget) available. Under a fixed admissible policy \( \xi \) the configuration problem can be formulated as follows.

\[
\max_{N} R_\xi(T, \Lambda, N) \quad \text{s.t.} \quad \sum_{j=1}^{2} c_j N_j \leq B, \quad N_i \geq 0, \quad i = 1, 2 .
\]

(2)

Because the admission policy \( \xi \) is assumed to be fixed, the only decision variable is the configuration of the rooms \( N \). Let \( \Psi_\xi(T, \Lambda, B) \) denote the optimal value of (2) for policy \( \xi \). The revenue maximization problem, which we refer to as REVMAX, can then be written as follows.

\[
\Psi(T, \Lambda, B) = \sup_{\xi \in \mathcal{P}} \Psi_\xi(T, \Lambda, B) .
\]

While parsimonious, our model is flexible enough to capture many characteristics of the differences between private and semi-private rooms including different space requirements, revenue rates, and recovery rates. The following remarks describe the additional phenomena the model can capture as well as some limitations in our approach.

**Remark 1.** Although our main focus is on the revenue maximization problem the profit maximization problem can be formulated similarly and the solution approach explained below can also be used in that case. Let \( c_i \) denote the financial cost of building a type \( i \) room, \( i = 1, 2 \). The profit maximization problem can be formulated as follows.

\[
\max_{\xi \in \mathcal{P}, N} \sum_{i,j=1}^{2} r_{ij} E[ D_{ij}(T) ] - \sum_{j=1}^{2} c_j N_j .
\]

**Remark 2.** In general, it may not be possible to board any two non-isolation patients in the same semi-private room mainly because of gender differences. Hence, when all but one of the semi-private beds are full, the next non-isolation may not be admitted if that patient is not compatible
with the patient boarding in the room with the only available bed. For pediatric hospitals that is less of a concern because it may be possible to board some of the younger non-isolation patients with different genders in the same room. It is possible to extend our results to this case by assuming that when only one semi-private bed is left and arriving non-isolation patient can be admitted with a fixed probability $p_l \in (0,1)$. We discuss the details of this case in Appendix F.

**Remark 3.** Some of the non-isolation patients may prefer to be boarded in a private room. In practice, hospitals charge these patients more to board them in a private room. Usually insurance companies do not cover such additional charges, hence patients may need to pay the additional charges, which could be substantial, out of their pocket. Although this is not considered in our original model, it is possible to extend our results (see Appendix F) to accommodate this feature by increasing the arrival rate to isolation patients and using a weighted average LOS.

**Remark 4.** Not only do patients in private rooms recover faster that is $\mu_{i1} > \mu_{i2}$ but their outcomes are better. By better outcomes we mean for example that their probability of dying during this visit to the hospital is lower. But more generally it could mean that their general health is better at the time of discharge than if they had been in a semi-private room. We could use our model to account for outcomes by redefining $r_{ij}$ as an expected medical outcome quality measure rather than revenue.

**Remark 5.** We are assuming that we know which patient is infectious and which is not at the time of admission and that this status does not change during the stay. In practice this is not the case and will lead to more infections and thus longer stays for other patients and worse outcomes because we may unintentionally place infectious patients in semi-private rooms with non-infectious patients facilitating the spread of illnesses. In our model this can be represented by the difference in recovery rates in semi-private versus private rooms but this does not fully capture the phenomenon.

**Remark 6.** With semi-private rooms it will be necessary to transfer patients between rooms as the mix of patients changes. With private rooms there is no need to transfer patients to free up space. When a patient is transferred their care is disrupted and work is created for hospital staff. If a patient is infectious the transfer may expose more people to that infection. We do not keep track of the number of transfers or the cost of these transfers in our model. For analytical tractability the FCFS-T policy we use leads to many patient transfers. In the Appendix we show using simulation that a modified version of FCFS-T leads to relatively few transfers.

**Remark 7.** In this paper we model a single hospital in isolation that is a monopolist in its market. In practice hospitals may be competing for patients with other hospitals. Given that most patients prefer private rooms a hospital may be driven to use more private rooms because
of competition. Analyzing such competition is beyond the scope of this paper but its existence strengthens the case for bed configurations with all private rooms.

3. Solution procedure

Even when an admission policy and a configuration are fixed, it is not possible to find a closed form expression for its performance, $R_\xi$, excluding very trivial cases. To simplify the problem we use a point-wise stationary approximation (PSA) for the demand process, see for example Green and Kolesar (1991), Harrison and Zeevi (2005), Whitt (1991), Bassamboo et al. (2006). In PSA, it is assumed that the system acts as if the system at time $t$ were in steady state with the arrival rate occurring at that instant. PSA’s are especially accurate when average LOS for patients is small relative to the time scale on which arrival rate changes occur, see Whitt (1991), Bassamboo et al. (2006). In the majority of specialties, average LOS is less than a few weeks and significant changes in arrival rates occur seasonally (see Langley et al. (1994) and Vayalumkal et al. (2007)).

To use the PSA procedure, we define the distribution of the demand instead of working with the stochastic process $\Lambda$. Let

$$F_T(\lambda) = \frac{1}{T} \int_0^T P\{\Lambda(t) < \lambda\} \, dt$$

denote the “distribution” of the demand. Therefore $F(\lambda)$ is the fraction of time the demand is less than $\lambda$. Given the arrival rate $\lambda$, the configuration $N = (N_1, N_2)$ and an admission policy $\xi$, let $\psi_\xi(\lambda, N)$ denote the total expected revenue obtained in steady state under policy $\xi$. For a fixed configuration $N$, using the PSA we estimate the total expected revenue by

$$\hat{R}_\xi(T, \Lambda, N) = T \int_{\mathbb{R}_+^2} \psi_\xi(\lambda, N) dF_T(\lambda).$$

Our proposed solution for (2) is the optimal solution to the following stochastic program

$$\max_{N} \hat{R}_\xi(T, \Lambda, N)$$

$$\text{s.t.} \quad \sum_{i=1}^2 c_i N_i \leq B$$

(3)

for a fixed policy $\xi$ and space constraint $B$. If we have a way to calculate $\psi_\xi(\lambda, N)$ and can discretize $\lambda$ appropriately it is possible to solve (3) using a linear search over $N_1$. The challenge is to derive $\psi_\xi(\lambda, N)$ for a candidate policy $\xi$. Let $N_\xi^*(T, \Lambda, B) = (N_1^*, N_2^*)$ denote the maximizer and $\phi_\xi(T, \Lambda, B)$ the optimal objective value under the policy $\xi$. 
3.1. REVMAX Approximation

Now we find the optimal configuration when we have the flexibility to choose an admission policy that maximizes the expected revenue. Fix \( N = (N_1, N_2) \) and \( \lambda = (\lambda_1, \lambda_2) \). The following LP is a continuous deterministic approximation that provides an upper bound on the achievable revenue.

\[
\begin{align*}
\max_{s_{ij}, i,j=1,2} & \sum_{i,j=1}^{2} r_{ij} s_{ij} \\
\text{s.t.} & \quad \frac{s_{11}}{\mu_{11}} + \frac{s_{21}}{\mu_{21}} \leq N_1, \\
& \quad \frac{2s_{12}}{\mu_{12}} + \frac{s_{22}}{\mu_{22}} \leq N_2, \\
& \quad s_{i1} + s_{i2} \leq \lambda_i, \text{ for } i = 1, 2, \\
& \quad s_{ij} \geq 0, \text{ for } i,j = 1, 2.
\end{align*}
\]

(4)

In this formulation, \( s_{ij} \) can be interpreted as the rate type \( i \) patients are admitted to type \( j \) rooms. The constraints (5) and (6) are the capacity constraints, based on the Little’s law, for type 1 and 2 rooms, respectively. Note that if an isolation patient is boarded in a type 2 room, only one patient can be boarded in that room and therefore that patient would effectively be using two beds. The constraint (7) is used determine how arriving type \( i \) patients are allocated to different types of rooms. Let \( \psi^*(\lambda, N) \) denote the optimal objective function value and \( s^*(\lambda, N) \) denote the optimal solution of (4). In the following we propose an admission policy \( \xi \) that asymptotically achieves \( \psi_\xi = \psi^* \). For notational simplicity we set

\[
\hat{R}_\omega(T, \Lambda, N) = T \int_{\mathbb{R}_+^2} \psi^*(\lambda, N) dF_T(\lambda)
\]

(9)

and denote the optimal solution of (3) with \( \hat{R}_\xi(T, \Lambda, N) = \hat{R}_\omega(T, \Lambda, N) \) by \( N^*(T, \Lambda, B) \) and the optimal objective function value by \( \phi^*(T, \Lambda, B) \).

Although our solution procedure is general, for most of the paper we focus on a special case of particular practical interest. We assume that it is more profitable (or efficient) to use private rooms to treat isolation patients and semi-private rooms to treat non-isolation patients. That is,

\[
r_{11}\mu_{11} \geq r_{21}\mu_{21}
\]

and

\[
0.5r_{12}\mu_{12} \leq r_{22}\mu_{22}.
\]

(10)

We also assume boarding patients of both types in a private room is more profitable, that is,

\[
r_{11}\mu_{11} \geq r_{12}\mu_{12} \text{ and } r_{21}\mu_{21} \geq r_{22}\mu_{22}.
\]

(12)
We need the following result in several places in the paper. When (10)--(12) hold, the optimal solution \( s^*(\lambda, N) \) of the LP (4) is given by

\[
\begin{align*}
    s_{11}^*(\lambda, N) &= \lambda_1 \wedge (\mu_{11}N_1), \\
    s_{21}^*(\lambda, N) &= \left( \mu_{21} \left( N_1 - \frac{s_{11}^*(\lambda, N)}{\mu_{11}} \right) \right) \wedge \lambda_2, \\
    s_{22}^*(\lambda, N) &= (\lambda_2 - s_{21}^*(\lambda, N)) \wedge (\mu_{22}N_2), \\
    s_{12}^*(\lambda, N) &= \left( 0.5\mu_{12} \left( N_2 - \frac{s_{22}^*(\lambda, N)}{\mu_{22}} \right) \right) \wedge (\lambda_1 - s_{11}^*(\lambda, N)).
\end{align*}
\] (13) (14)

In words, (13) and (14) imply that private rooms are used first and isolation patients are given priority over non-isolation patients to be boarded in private rooms. If the private rooms are full, the semi-private rooms are allocated to non-isolation patients and any remaining capacity in semi-private rooms are then allocated to any unmet demand from isolation patients. The following result identifies a set of conditions that guarantees to have all private rooms optimal.

**Lemma 1.** Assume that in addition to (10)--(12), the following conditions hold

\[ r_{11}\mu_{11} \geq \frac{c}{2} r_{12}\mu_{12}, \quad r_{21}\mu_{21} \geq c r_{22}\mu_{22}, \] (15)

where \( c = c_2/c_1 \). Then the optimal solution of (9) is given by \( N_1^* = B/c_1 \) and \( N_2^* = 0 \). Specifically it is optimal to have private rooms only.

The proof is in Appendix A.

### 3.2. Performance approximations for FCFS-T

In this section we present the FCFS-T policy and an approximation for its performance with the following two goals in mind. First we want to present a policy, FCFS-T, that is fair and easily implementable. (We later show that it performs well.) Also we would like to demonstrate how the solution procedure described in §3 can be used with this policy by establishing approximations for its performance. The approximations are obtained by considering an asymptotic regime and the details are quite technical. Therefore, we present the details (which are not essential to understand the other results in the paper) in a separate section, §5, after we describe the asymptotic regime in §4.

Under FCFS-T the goal is to avoid turning away patients if possible so when a new patient arrives, a patient who was admitted earlier may have to be transferred to another room to open up enough space for the incoming patient. For concreteness we assume that the following procedure is followed for patient transfers when (10) and (12) hold. The same transfer procedures should be used whether (11) holds or not.
We assume that an arriving patient is admitted to a private room if one is available. When a patient is discharged from a private room, a patient from one of the semi-private rooms is transferred to that private room with preference given to isolation patients. If an isolation patient arrives to find all private rooms taken, a non-isolation patient boarding in a private room (if there are any) is transferred to a semi-private room with an empty bed (if there are any available) and the arriving isolation patient is admitted to the private room that became available. Recall that for an isolation patient to be admitted to a semi-private room, both beds in that room must be empty. We assume that if there are two non-isolation patients boarding in two different rooms alone and an isolation patient needs to be admitted to a semi-private room, and if all the semi-private rooms are occupied, one of the non-isolation patients is transferred to the same room with the other non-isolation patient boarding alone. Approximations we propose below can be modified to accommodate other patient transfer policies when (10) and (12) do not hold as well. We discuss additional cases in Appendix G.

Next we propose an approximation for the expected revenue in steady state, \( F_{\text{under the FCFS-T admission policy given}} \), under the FCFS-T admission policy given \((\lambda, N)\). Our approach is based on an approximating procedure for the expected number of idle beds. In this procedure we use the steady state estimates from a many-server asymptotic analysis. More details are provided in §5 below. Fix the arrival rates \( \lambda = (\lambda_1, \lambda_2) \) and the capacity \( N = (N_1, N_2) \). Let \( P_i \) denote the probability that a type \( i \) patient is admitted in steady state and \( \rho_{ij} = \lambda_i / \mu_{ij}, \) for \( i,j = 1, 2 \). Our approximations for the admission probabilities depend on whether the system is overloaded or not. If

\[
2 \frac{\mu_{11}}{\mu_{12}} (\rho_{11} - N_1)^+ \leq (N_2 - \rho_{22})^+ \quad \text{and} \quad \frac{\mu_{22}}{\mu_{21}} (\rho_{22} - N_2)^+ \leq (N_1 - \rho_{11})^+,
\]

then the system has sufficient nominal capacity, hence we set \( P_1 = P_2 = 1 \). To see why (16) implies that the system has enough capacity, assume that \( \lambda_1 > \mu_{11}N_1 \). If the first condition in (16) holds then it implies (by simple algebra) that \( \mu_{12} (N_2 - \rho_{22}) \geq 2(\lambda_1 - \mu_{11}N_1) \). Observe that \( 0.5\mu_{12} (N_2 - \rho_{22})^+ \) is the amount of capacity in type 2 rooms that can be allocated to type 1 patients after serving all type 2 patients. Hence, if \( \lambda_1 > \mu_{11}N_1 \), then the first condition in (16) ensures that the system has enough nominal capacity to serve all the patients. The second condition in (16) works in a similar way when \( \lambda_2 > \mu_{22}N_2 \). If \( \lambda_1 \leq \mu_{11}N_1 \) and \( \lambda_2 \leq \mu_{22}N_2 \), then it is easy to show that both conditions in (16) hold.

Now assume that at least one of the inequalities in (16) does not hold, i.e. the system is overloaded. Then either \( N_1/\rho_{11} < 1 \) or \( N_2/\rho_{22} < 1 \), or both hold. When the system is overloaded there are three possible cases we consider separately that lead to different approximations. First, type \( i \)
patients are boarded in type $i$ rooms only, second some type 1 patients overflow to type 2 rooms but not the other way around and third some type 2 patients overflow to type 1 but not the other way around. If

$$\frac{N_1}{\rho_{11}} > \frac{N_2}{\rho_{22}}$$

then some of the type 2 patients overflow to type 1 rooms. This follows from the fact that if this were not the case then the admission probability for type 2 patients (approximately equal to $N_2/\rho_{22}$) will be lower than the admission probability for type 1 patients (approximately equal to $N_1/\rho_{11}$). Under FCFS-T policy this is not possible in general. Similarly if

$$\frac{N_1}{\rho_{11}} < \frac{N_2}{\rho_{22}}$$

then some of the type 1 patients overflow to type 2 rooms.

If (16) does not hold and $\frac{N_1}{\rho_{11}} = \frac{N_2}{\rho_{22}}$ then there is no overflow and so we set $P_i = N_i/\rho_{ii}$, $i = 1, 2$. If (16) does not hold and (17) holds then we set $P_1 = P_2 = P$, where $P$ is the unique root of

$$-\left(\frac{\mu_{11}}{\mu_{22}} + \rho_{22}\right)P^2 + \left(\frac{\mu_{11}}{\mu_{22}} + \rho_{22} + 1 + \frac{\mu_{21}}{\mu_{22}}N_1 + N_2\right)P - \left(\frac{\mu_{21}}{\mu_{22}}N_1 + N_2\right) = 0$$

between 0 and 1. We describe the details of how (19) is obtained in §5. Denote the second order polynomial on the left hand side of (19), by $f_1(P)$, it satisfies $f_1(0) < 0$, $f_1(1) > 0$, $f_1(P) \rightarrow 1$ as $P \rightarrow 1$. Hence, there is a unique solution $P$ on $(0,1)$.

If (18) holds there are two situations we need to consider. First, let $\hat{P}$ denote the unique root of

$$-2\rho_{12}P^3 + (2\rho_{12} - \rho_{22})P^2 + \left(\rho_{22} + 1 + 2\frac{\mu_{11}}{\mu_{12}}N_1 + N_2\right)P - \left(2\frac{\mu_{11}}{\mu_{12}}N_1 + N_2\right) = 0. \quad (20)$$

between 0 and 1. (See §5 for details.) We note that the polynomial on the left hand side of (20), denote it by $f_2(P)$, satisfies $f_2(P) \rightarrow \infty$ as $P \rightarrow -\infty$, $f_2(0) < 0$, $f_2(1) > 0$, and $f_2(P) \rightarrow -\infty$ as $P \rightarrow \infty$. Hence there exists a unique root on $(0,1)$. If $\hat{P}^2\lambda_1 > N_1/\rho_{11}$, then we set $P_1 = \hat{P}^2$ and $P_2 = \hat{P}$, otherwise we set $P_1 = \mu_{11}N_1 + 0.5\mu_{12}\tilde{N}_{12}$ and $P_2 = (N_2/\rho_{22}) \wedge 1$, where $\tilde{N}_{12} = (N_2 - \rho_{22})^+$. If (16) holds then $\lambda_{ij} = s_{ij}^*$, for $s_{ij}^*$ defined as in (13) and (14).

Using (19) and (20), we obtain the following approximations depending on system load for the total expected revenue under FCFS-T in steady state to be used in solving (9). We have

$$\psi_F(\lambda, N) = \sum_{i,j=1}^{2} r_{ij}\lambda_{ij}(\lambda, N),$$

where $\lambda_{ij}$ are determined as follows. If (16) holds then $\lambda_{ij} = s_{ij}^*$, for $s_{ij}^*$ defined as in (13) and (14). If (16) does not hold and $\frac{N_1}{\rho_{11}} = \frac{N_2}{\rho_{22}}$ then

$$\lambda_{11} = \mu_{11}N_1, \lambda_{22} = \mu_{22}N_2 \text{ and } \lambda_{ij} = 0 \text{ if } i \neq j, i, j = 1, 2.$$
If (16) does not hold but (17) holds, then
\[ \lambda_{11} = \lambda_1 P_1, \lambda_{21} = \mu_2 (N_1 - \lambda_{11}), \lambda_{22} = P_2 \lambda_2 - \lambda_{21} \text{ and } \lambda_{12} = 0. \]

If (16) does not hold and (18) holds, then
\[ \lambda_{11} = \mu_{11} N_1, \lambda_{22} = P_2 \lambda_2, \lambda_{12} = P_1 \lambda_1 - \lambda_{11} \text{ and } \lambda_{21} = 0. \]

Remark 8. To calculate revenues under FCFS-T we assign revenue \( r_{ij} \) to a type \( i \) patient that completes his stay in a type \( j \) room regardless of the number and locations of his transfers during his entire stay. When we are maximizing throughput we set \( r_{ij} = 1 \) for all \( (i, j) \) so transfers are irrelevant to counting output. Because we assume that \( r_{11} = r_{12} \) transfers do not affect the calculations of revenue from isolation patients when \( r_{ij} \neq 1 \). The inaccuracy in the calculating revenues with our definition of \( r_{ij} \) comes from type 2 patients who spend part of their visit in each of the two types of room when \( r_{21} > r_{22} \). We do not believe that this inaccuracy is of significance because if we look at the system from the perspective of a room we are calculating revenue by determining the rate at which a particular room type generates revenue based upon the fraction of time it is occupied by type 1 or type 2 patients. An individual patient’s revenue contribution may not be accounting for how he spent time in different room types but the room’s revenue generation is accounted for correctly on average.

4. Asymptotically optimal admission policies

In §3 we proposed a continuous relaxation of REVMAX and derived the associated revenues. In this section we propose a preemptive admission policy which attains this revenue asymptotically. Although the proposed policy may not be very practical, we can use it to benchmark the performance of FCFS-T in our numerical experiments in §6.

The policy we propose is very similar to the FCFS-T admission policy except that it takes a different action when a patient arrives to find no available bed(s) even after transfers are done. The proposed policy is for the case when (10)–(12) hold. We discuss the extension to other cases in Appendix G. If at the time of a type \( i \) patient arrival there is not sufficient space to admit the patient and if there are type \( i' \) patients boarding in a type \( i \) room, one of the type \( i' \) patient’s treatment is preempted and the arriving type \( i \) patient is admitted in the vacated bed. We denote this policy by REVMAX\(_P\). We assume that no revenue is obtained from the patient whose treatment is preempted. Although preempting a patient’s treatment is not practical in most applications, it may be possible to transfer a boarding patient to a step-down unit without any significant adverse medical effects in certain situations.
To prove the accuracy of the proposed solution procedure and the effectiveness of the proposed admission policies in a certain many-server asymptotic regime, we consider a sequence of systems indexed by $n$. Let \( \{k^n\} \) denote a sequence of real numbers such that $k^n \to \infty$ and $k^n/n \to 0$ as $n \to \infty$ and $T^n = k^n T$. Assume that the arrival rate $\Lambda^n$ in the $n$th system satisfies

$$\Lambda^n(t) = n\Lambda\left(\frac{t}{k^n}\right)$$

(21)

for a nonnegative continuous stochastic process $\Lambda$ satisfying

$$\sup_{t \in [0,T]} \|\Lambda(t)\| < M \text{ a.s.}$$

(22)

for some constant $M < \infty$. Therefore, along this sequence, the arrival rate gets larger but the relative rate of change becomes smaller with $n$. We assume that the LOS distributions, hence $\mu_{ij}$'s, are independent from $n$. We also assume that the total space constraint $B^n$ scales up linearly with the demand;

$$B^n = nB$$

(23)

for some $B > 0$.

First we consider the solution of (3) in the $n$th system. We set

$$\bar{\phi}^{*,n}(T^n, \Lambda^n, B^n) = \frac{\phi^*(T^n, \Lambda^n, B^n)}{k^n}$$

and $N^{*,n} = N^*(T^n, \Lambda^n, B^n)$ (recall the definition of $\phi^*$ and $N^*$ in §3.1). The following result is immediate from algebraic manipulations and it shows that the solution of (3) with $\bar{R}_n$ is “scale invariant” in the asymptotic regime we consider. Specifically

$$\phi^*(T, \Lambda, B) = \bar{\phi}^{*,n}(T^n, \Lambda^n, B^n) \text{ and } N^{*,n} = nN^*(T, \Lambda, B).$$

(24)

Now we are ready to show that $\phi^*$ is an asymptotic upper bound for the optimal revenue. We define the scaled optimal revenue for the $n$th system by

$$\bar{\Psi}^n(T^n, \Lambda^n, B^n) = \frac{\Psi(T^n, \Lambda^n, B^n)}{k^n}.$$ 

The following result implies that (3) provides an asymptotic upper bound for the achievable performance.

**Theorem 1.** Consider a sequence of systems indexed by $n$ that satisfy (21)–(23). Then

$$\limsup_{n \to \infty} \bar{\Psi}^n(T^n, \Lambda^n, B^n) \leq \phi^*(T, \Lambda, B).$$
The proofs of the results in this section are placed in the appendix.

Now we show that the configurations found using (3) with \( \psi_\xi = \psi^* \) are asymptotically optimal when used with \( \text{REVMAX}_P \). Let

\[
\tilde{R}_\xi^n(T^n, \Lambda^n, N) = \frac{R_\xi(T^n, \Lambda^n, N)}{k^n}.
\]

**Theorem 2.** Consider a sequence of systems indexed by \( n \) that satisfy (21)–(23). Also assume that (10)–(12) hold. Then,

\[
\liminf_{n \to \infty} \tilde{R}^n_{\text{REVMAX}_P}(T^n, \Lambda^n, N^{*,n}) = \phi^*(T, \Lambda, B).
\]

In words, combined with Theorem 1, Theorem 2 implies that if the configuration is selected according to (3) and the \( \text{REVMAX}_P \) policy is used then the properly scaled revenue is optimized asymptotically. In addition to providing optimal staffing levels, (3) also provides asymptotically correct estimates for the revenue under \( \text{REVMAX}_P \) by (24) and (25).

5. **Derivation of the performance approximations under FCFS-T**

In this section we provide the details of how the approximations proposed in §3.2 for the steady state under FCFS-T are obtained. Our approximations draw upon many-server asymptotic results from Hunt and Kurtz (1994). We also use a novel way to refine these approximations using an approach similar to that in Perry and Whitt (2009) and Perry and Whitt (2011). We do not attempt to prove any limit theorems here (since it is not the main focus of this paper), but we reference the relevant results from the literature when needed.

Fix the arrival rate \( \lambda_i \), \( i = 1, 2 \), and number of beds \( N = (N_1, N_2) \). We use \( z_{ij} \) to denote the expected number of type \( i \) patients boarding in type \( j \) rooms in steady state. By the conservation of flow

\[
P_1 \lambda_1 = \mu_{11} z_{11} + \mu_{12} z_{12},
\]

*Figure 1  Underlying Markov chain for the number of available beds*
The general idea behind our approach is to find an approximation for the $z_{ij}$'s, $i, j = 1, 2$. In order to find these approximations, we use a Markov chain whose transition rates are derived from the many-server fluid limit result, Theorem 3, in Hunt and Kurtz (1994). The results in Hunt and Kurtz (1994) are only applicable to the cases when there is one type of rooms. Hence, we use insights from the many-server fluid analysis of the systems with two types of rooms to adapt the results in Hunt and Kurtz (1994). The analysis of the many-server fluid model of the FCFS-T policy is presented in Appendix J for systems with two types of rooms, see Lemma 7 and Remark 9. Although they are not necessary to understand the rest of this section, those results provide a theoretical basis for our approximations of the patient overflows. Loosely speaking, in many-server fluid analysis the limit of a sequence of systems with arrival rate and number of servers going to infinity at the same rate (while service rates are kept fixed) is analyzed to obtain approximations for the steady state of a system. We also refine these approximations by estimating the expected number of idle beds in steady state using this Markov chain. A similar approach is used in Perry and Whitt (2009) and Perry and Whitt (2011) to build approximations of the steady state of a different system.

We start with the easier cases. If (16) holds and the system is underloaded we assume that all arriving patients are admitted. Therefore, we can set $P_1 = P_2 = 1$ (which is asymptotically correct in the fluid limit.) For the rest of this section we assume that (16) does not hold and the system is overloaded.

When $\frac{N_1}{\rho_{11}} = \frac{N_2}{\rho_{22}}$ both types of rooms are overloaded and there will rarely be opportunities for type $i$ to go to type $i'$ rooms. (One can actually show that $z_{21} \sim o(N)$ and $z_{12} \sim o(N)$ for $N$ large using a many-server fluid analysis.) Therefore we approximate $P_i = \frac{N_i}{\rho_{ii}}$, $i = 1, 2$. This means that $\lambda_{ii} = \mu_{ii}N_i$. 

$$P_2\lambda_2 = \mu_{21}z_{21} + \mu_{22}z_{22}. \quad (27)$$

Figure 2  Underlying Markov chain for the number of available beds
If (17) holds then type 2 patients overflow to type 1 beds but type 1 patients are boarded in type 1 rooms only. (Formally, one can show that \( z_{12} \sim o(N) \) using a many-server fluid analysis.) Hence, we assume that \( z_{12} \approx 0 \), \( z_{21} > 0 \) and so \( z_{11} + z_{21} \approx N_1 \) and \( z_{22} \approx N_2 \). Also, because patients are transferred to a private room whenever possible, private rooms are almost always full. Therefore an incoming patient can only be admitted if there are semi-private rooms available. If an isolation patient arrives and there is an available bed in a semi-private room, one of the non-isolation patients in a private room is transferred to a semi-private room and the arriving patient is admitted. Based on our approximation \( z_{21} > 0 \), there is always a non-isolation patient in a private room. On the other hand, a non-isolation patient can be admitted as long as there is an available bed. This implies that \( P_1 \approx P_2 \). Given the above approximation of how the system behaves (which can be shown to be true for the many-server fluid limit of the system), we can view type 2 beds as a separate system that can be modeled as a simple birth-death Markov chain with \( X_t \) the number of idle type 2 beds (see Figure 1). To make this approach more tractable we will set \( z_{12} = 0 \) and \( P_1 = P_2 = P \). Now the expected number of type 2 idle beds is given by (obtained by solving for the steady state probabilities of the Markov chain \( X_t \)) \( P/(1-P) \) so we can estimate \( z_{22} \) as:

\[
 z_{22} = N_2 - P/(1-P). \tag{28}
\]

The birth rates in the Markov chain \( X_t \) represents the rate at which type 2 patients leave type 2 rooms which is on average \( \mu_{11}z_{11} + \mu_{21}z_{21} + \mu_{22}z_{22} \) because of completed treatment and patient transfers (recall that patients are transferred to a type 1 room when a patient from a type 1 room leaves the system). Based on our assumptions we can also estimate that:

\[
 z_{11} + z_{21} = N_1 \text{ and } z_{12} = 0. \tag{29}
\]

Solving (26)–(29) with \( P = P_1 = P_2 \) for \( P \) we obtain (19).

Now assume that (18) holds. In this case type 1 patients overflow to type 2 rooms, but not the other way around. By similar arguments as above it is reasonable to assume that \( z_{21} \approx 0 \), \( z_{11} \approx N_1 \), and \( z_{21} + z_{22} \approx N_2 \). If \( z_{12} > 0 \), we expect the number of type 1 patients boarding in type 1 rooms to be equal to \( N_1 \) most of the time because of patient transfers. Therefore, both patients types are admitted only to type 2 rooms. This implies that a type 1 patient is admitted only if there are two beds available in type 2 rooms and so the system operates very similarly to a system with only semi-private rooms. Hence we set \( P_1 = (P_2)^2 \) (based on the results from Hunt and Kurtz (1994), Reiman (1991) and Kelly (1986)). To find an approximation for the number of type 2 idle beds we use the Markov chain \( X_t \) with transition rates given in Figure 2. The departure rate for
type 1 and type 2 patients is equal to \( \mu_{11}N_1 + \mu_{12}z_{12} \) and \( \mu_{22}z_{22} \), respectively. If a type 1 patient is discharged, two additional type 2 beds become available (recall our assumption that \( z_{12} > 0 \)), possibly after patient transfers. Because all type 1 patients are admitted to type 2 rooms, if a type 1 patient is admitted then two additional type 2 beds become occupied. On the other hand, type 2 patients only occupy one type 2 bed at a time and hence when a type 1 patient is admitted (discharged), one additional type 2 bed becomes occupied (resp. available). Again, the expected number of idle type 2 beds is equal to \( \frac{P}{1 - P} \) for \( P = P_2 \).

Let \( P \) be the solution. In this solution it is possible that \( z_{12} < 0 \) (which is not feasible) hence the solution needs to be modified because our approximations assume that \( z_{12} > 0 \). This implies that with admission probabilities \( P_1 \) and \( P_2 \) either \( P_2 \approx 1 \) or the system cannot have \( z_{12} > 0 \) and therefore \( z_{12} = 0 \). If \( P_2 = 1 \) then \( \tilde{N}_{12} = N_2 - \rho_{22} \) semi-private rooms must be allocated to type 1 patients. Therefore \( P_1 = (\mu_{11}N_{11} + 0.5\mu_{12}\tilde{N}_{12})/\lambda_1 \). If \( P_2 < 1 \) then \( z_{12} = 0 \) and since \( z_{21} = 0 \) as well, the admission probabilities are given by \( P_i = \frac{N_i}{\rho_i}, i = 1, 2, \) in that case.

6. Numerical Experiments

In this section we first illustrate the accuracy of the proposed approximation method using simulation experiments. Then, in §6.3 we carry out an extensive numerical experiment to gain insights on how to effectively configure a hospital’s rooms and on the effect of admission policies on the optimal revenue. We begin with explaining the structure of a general arrival rate pattern that we use in our experiments.

6.1. Arrival rate patterns

We consider a seasonal arrival pattern whose parameters are derived from the trends reported in the literature (mainly those reported in Langley et al. (1994) and Vayalumkal et al. (2007) in pediatric hospitals) for the demand for beds in hospitals. We assume that the arrival rate and the proportion of infectious patients are at the highest during winter and in our experiments we consider various degrees of changes in rates from winter to summer. Let \( \gamma(t) \in [0, 1] \), with \( \gamma(0) = \gamma(1) = 1 \) denote the “intensity” level of the total arrival rate at time \( t \) and set \( \gamma_m = \gamma(0.5) \) to denote the minimum level. Let \( \lambda(t) \) denote the total arrival rate to the system at time \( t \) and assume that it is given by

\[
\lambda(t) = \lambda_0 \gamma(t), \quad t \geq 0
\]

for \( \lambda_0 > 0 \). For simplicity we assume that \( \gamma \) is piecewise linear in \( t \) and we set

\[
\gamma(t) = \begin{cases} 
1 - 2(1 - \gamma_m)(t - \lfloor t \rfloor), & \text{if } t - \lfloor t \rfloor \leq 1/2 \\
\gamma_m + 2(1 - \gamma_m)(t - \lfloor t \rfloor), & \text{if } t - \lfloor t \rfloor \geq 1/2 
\end{cases}
\]

for \( \gamma_m > 0 \).
Figure 3 Seasonal arrival Rates

Note that $\gamma(t)$ is maximum at time 0, hence, $t = 0$ can be taken as the initial calendar day of a year.

Let $p_i(t)$ denote the probability that a customer arriving at time $t$ is a type $i$ patient. We assume that $p_i$ also follows a piece-wise linear function. Specifically,

$$p_1(t) = 1 - p_2(t) = \begin{cases} p_{\text{max}} - 2(p_{\text{max}} - p_{\text{min}})(t - \lfloor t \rfloor), & \text{if } t - \lfloor t \rfloor \leq 1/2 \\ p_{\text{min}} + 2(p_{\text{max}} - p_{\text{min}})(t - \lfloor t \rfloor), & \text{if } t - \lfloor t \rfloor \geq 1/2 \end{cases}$$

(32)

for $1 > p_{\text{max}} > p_{\text{min}} \geq 0$. In our initial numerical experiments we use two sets of parameters: in the first set we use $\gamma_m = 0.7$, $p_{\text{max}} = 0.3$, $p_{\text{min}} = 0.1$ and in the second one we set $\gamma_m = 0.8$, $p_{\text{max}} = 0.3$ and $p_{\text{min}} = 0.2$. We plot the total arrival rate $\lambda$ along with the arrival rates to each class in Figure 3 for these two cases when $\lambda_0 = 100$ patients per day. For the experiments in §6.3, we consider a variety of combinations of these parameters that are presented in Table 3 below.

6.2. Accuracy of Approximations

In this section we present the results of an initial set of numerical experiments to assess the quality of our solution procedure and the approximations we proposed for the expected revenue under REVMAX$_P$ and FCFS-T. We run two sets of experiments with arrival patterns that are similar to those in Figures 3(a) and 3(b). We consider three different values for $\lambda_0$: 25, 50 and 100 patients per day. In both experiments we set the service rates to $\mu_{11} = \mu_{12} = \mu_{21} = \mu_{22} = 0.5$. Therefore, the average LOS of both types is around 2 days.

Here we assume the LOS distribution is exponential, tests with a log-normal LOS distribution with the same mean LOS and coefficient of variation equal to 0.5 yielded similar results. We set
\( r_{ij} = 1 \) units per patient, for \( i, j = 1, 2 \), hence our focus is on throughput. We set \( c_2 = 1 \) unit per bed and consider three different values for \( c_1 \): 1.25, 1.5 and 1.75 per bed. We assume that there is sufficient space (budget) to build 50 semi-private beds when \( \lambda_0 = 25 \), 100 semi-private beds when \( \lambda_0 = 50 \), and 200 semi-private beds when \( \lambda_0 = 100 \) per day. Therefore we are assuming that the system is heavily loaded some of the time and more lightly loaded at other times.

We proceed as follows. For a fixed parameter set, first we find the estimated optimal configurations by solving (3) for FCFS-T and the REVMAX\(_P\) policy using functions \( \psi_F \) and \( \psi^* \), respectively, to approximate the expected revenue. We then simulate the system with this estimated optimal configuration to determine the actual expected revenue these solutions obtain. We call these the stochastic optimization solutions \(^1\). To obtain an estimate for the “true” optimal configuration and the optimal revenue we simulate each system with all the feasible room combinations under the REVMAX\(_P\) policy, and FCFS-T. Then, we compare the revenues observed in simulations for the estimated optimal configuration found by (3) with the maximum revenue observed among all feasible room combinations in our simulations. For each setting we simulate the system for 20 years and we use the first 10 years as a warm-up period and use five replications. To illustrate the efficiency of the proposed method, we note that it takes around 1:30 hours of CPU time to find the optimal revenue with simulations when \( \lambda_0 = 100 \) for each parameter combination, whereas with our method it takes less than several seconds.

Table 1 shows the average relative differences between results of the stochastic optimization approach and the exhaustive simulations in terms of the optimal total number of beds and the optimal revenues. The averages are taken over three different \( c_1 \) values for a fixed arrival pattern and a fixed \( \lambda_0 \). In Table 1, the first column indicates the arrival pattern used and the second gives the \( \lambda_0 \) value used.

It is clear from Table 1 that in terms of the expected revenue, the configurations found by (3) perform remarkably well for all policies. Even for \( \lambda_0 = 25 \), the average difference between revenues of the configurations found by (3) and the ones found by simulations is below 2.5% for all policies. The difference goes much lower when \( \lambda_0 = 100 \). Better performance in larger systems is expected since our approximations are based on an asymptotic analysis of large systems. In terms of the total number of beds, the results are a little worse, but the average error is still less than 2.5% even when \( \lambda_0 = 25 \). The better accuracy of the proposed method in estimating revenues compared to estimating the optimal total number of beds can be ascribed to the observed flatness of the

\(^1\) In the case of REVMAX\(_P\) we find the optimal configuration using the LP model (4) but simulate the performance of the (preemptive) REVMAX\(_P\) policy in §4 to estimate the expected revenue generated.
expected revenue around the optimal solution in these experiments. More detailed results of the simulations for \( \lambda_0 = 100 \) are presented in Appendix I.

In order to assess the quality of the approximations for the revenue given by (3), we compare the revenues estimated by (3) with the revenues observed in simulations when the optimal configurations found by (3) are used under FCFS-T and REVMAX\( r \) policies. The average differences are presented in Table 2 for three different \( \lambda_0 \)'s under two arrival patterns and the two admission policies. Again, we present the averages that are obtained by using the results under three different \( c_1 \) values. The differences between the estimated revenues and the revenues obtained from the simulations are larger than the differences between the revenues observed in simulations. The differences are smaller for larger systems as expected. In addition, the errors in estimating revenues are smaller in the first arrival pattern than the second one. We conjecture that this is due to the fact that our estimates are based on fluid limits and such estimates are known to be very accurate when the arrivals become more random, see for example Bassamboo et al. (2010). The performance of our procedure in terms of finding effective configurations is thus much better than its performance in estimating the revenue. This can be explained by the fact that our procedure overestimates the revenues in general by a similar amount for room configurations that are similar. In addition, we observed in our experiments that a large portion of the error in estimating the revenue is because of the simple approximations we use when the system is not overloaded. When the system load is light the bed configuration will not have a strong effect on performance.

<table>
<thead>
<tr>
<th>Arr. Pattern</th>
<th>( \lambda_0 )</th>
<th>Revenue</th>
<th># of beds</th>
<th>Revenue</th>
<th># of beds</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25</td>
<td>1.04</td>
<td>2.47</td>
<td>0.29</td>
<td>1.80</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>0.72</td>
<td>2.46</td>
<td>0.08</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.44</td>
<td>1.69</td>
<td>0.14</td>
<td>0.28</td>
</tr>
<tr>
<td>2</td>
<td>25</td>
<td>2.38</td>
<td>1.85</td>
<td>1.33</td>
<td>1.17</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>1.46</td>
<td>2.46</td>
<td>0.34</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.62</td>
<td>1.21</td>
<td>0.15</td>
<td>0.73</td>
</tr>
</tbody>
</table>

Table 1 Average differences between optimal total number of beds and revenues obtained by solving stochastic optimization versus exhaustive simulation.

6.3. Policy Analysis

Having demonstrated the accuracy of our approximations we now explore some policy questions using a more extensive set of parameter combinations. Specifically we focus on two aspects of these systems investigating (i) if there is a significant potential to improve performance by optimizing
Table 2 Average differences between simulated revenues and approximated revenues at optimum solution selected by stochastic optimization

<table>
<thead>
<tr>
<th>Arr. Pattern</th>
<th>$\lambda_0$</th>
<th>REVMAX$_P$</th>
<th>FCFS-T</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25</td>
<td>7.31</td>
<td>3.96</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>4.68</td>
<td>2.33</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>2.88</td>
<td>1.32</td>
</tr>
<tr>
<td>2</td>
<td>25</td>
<td>8.63</td>
<td>10.27</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>5.85</td>
<td>6.11</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>3.74</td>
<td>3.51</td>
</tr>
</tbody>
</table>

The results of the numerical experiments are presented in Table 4. The column “Experiment”
indicates the LOS setting. The differences between optimal revenues under REVMAX_P and FCFS-T policies are presented in row “F-P”. In rows “F-S”, “F-D” and “F-H” we present the differences between revenues in systems with the optimal configuration and with private rooms only (F-S), with semi-private rooms only (F-D), and with at least half of the rooms private (F-H), under FCFS-T. We present the mean difference of revenues across all the parameter combinations in column “Mean”. We also present the standard deviation of the differences and the maximum difference in columns “Stdev” and “Max”. Among all the parameters, c₁, the space required for a private room, seems to have the biggest impact on the differences. Therefore, we also present the average difference when c₁ = 1.5. Finally, in the last column we focus on the throughput and present the differences between revenues when c₁ = 1.5 and r₁ = 1, for the first LOS setting.

In agreement with the results in the previous section, there is a relatively small difference between the performance of FCFS-T and REVMAX_P policies, with average difference around 1.3% and 2.5% in experiments 1 and 2, respectively. This suggests that there is little to be gained from optimizing the admission policy beyond the natural approach of FCFS-T. On the other hand using private rooms only seems to reduce the revenue significantly, in some cases by as much as 45%. Using semi-private rooms only does not have such a significant effect, although it can also reduce revenue by as much as 20.23%. Generally, requiring at least half of the rooms to be private performs better than having semi-private rooms only. In terms of throughput, when c₁ = 1.5, in the first LOS setting, the average differences are very similar to the overall average differences. Therefore, having only private rooms can reduce the throughput significantly as well, i.e. it is bad for access.

In addition to the expected revenue, another performance measure of interest is the admission probabilities for each patient type when different configurations are used under FCFS-T. We report the average admission probabilities for each patient type in Table 5 for the two LOS settings, with Pᵢ the average admission probability for type i patients, i = 1, 2. Each row reports a different configuration, optimal configuration in rows “Opt”, all private rooms configuration in rows “S” and all semi-private rooms configuration in “D”. The average admission probabilities are larger

² We note that it is possible to construct examples when this is not true. For example, consider a system with all private rooms with the same mean LOS for both patient types. If the revenue per patient for one of the patient types, say non-isolation patients, is significantly larger than the revenue for the other patient type, then REVMAX_P policy would admit as many non-isolation patients as possible and depending on the difference in revenues, it would generate a significantly larger revenue than FCFS-T. However, a significant difference between revenues per patient is not very common in practice. In addition, in most of the experiments we considered it is optimal to use a mix of private and semi-private beds and most of the isolation patients are boarded in private rooms. Therefore, when the optimal configuration is used (assuming (10) holds), the impact of using FCFS-T instead of a (less practical) REVMAX_P policy on the revenue is not significant.
for both patient types when the optimal configurations are used than those when only private or semi-private rooms are used demonstrating better access. Also, in all the experiments, the mean admission probabilities under the optimal configurations are larger than those when only private rooms are used for both types of patients. When only semi-private rooms are used the mean admission probabilities for isolation patients are worse than those under the optimal configuration in all experiments. The admission probabilities for non-isolation patients are improved in 1/3 of experiments with the maximum improvement equal to 5.5%, compared with the optimal configuration.

It is argued that the LOS of a patient can be reduced if the patient is boarded in a private room instead of a semi-private room, see, for example, Boardman and Forbes (2011). Next we report the results of additional experiments to assess the impact of a potential reduction in average LOS in private rooms on the capacity decisions. We use the same parameter combinations in the first LOS setting above, but we assume that the rate patients recuperate in private rooms or if they board alone in a semi-private room is increased by 10% first and then by 20%. The results are presented in Table 6 in a way similar to Table 4. The only difference is that instead of two different LOS settings we consider two different “speed-up” values for LOS in private rooms. When the mean LOS is shorter in private rooms, the negative impact of using all private rooms on the revenue is reduced. We see that, the differences between the revenues under all private rooms and under the optimal configurations decrease but are still significant when rates are increased by 10% or 20%. We conclude that even with shorter LOS in private rooms, having all private rooms may still reduce the throughput and the revenue significantly.

6.4. Sensitivity Analysis

Now we look more closely at what the optimal room configurations are and how they depend on the system parameters. We focus on the case when FCFS-T is used and when $\lambda = 50$, $\gamma_m = 0.7$, $p_{max} = 0.3$ and $p_{min} = 0.1$. We consider $c_1 = 1.3, 1.5, 1.7$. We assume that the rate patients recover in private rooms are faster and we denote the increase in the rate by $\alpha$ and we consider $\alpha = 1, 1.2, 1.4$. We set $\mu_{22} = 0.5$ and assume that $\mu_{ij} = 0.5\alpha$ otherwise. We assume that $r_{22} = 1$ and we consider different values for $r_{11} = r_{12} \geq r_{22}$ and $r_{21}$. First we set $r_{21} = 0.5(r_{11} + r_{22})$, referred to as Case A, hence, non-isolation patients who board in a private room pay average of what is charged for a private room for isolation patients and what is charged for a bed in a semi-private room. Then we consider that case when $r_{21} = r_{11}$ (referred to as Case B) that is, we assume that non-isolation patients are charged at the same rate as isolation patients if they are boarded in a private room.

The results for Case A and B are presented in Figures 4 and 6, respectively. In each graph we plot the optimal percentage of private beds vs. $r_{11}$ for a fixed value of $c_1$ and $\alpha$, whose values are
Figure 4 Percentage of private beds vs. $r_{11}$ Case A: $r_{21} = 0.5(r_{11} + r_{22})$

displayed on top of the graph. In Figures 5 and 7 we present the difference between the revenues under the optimal configuration and under the case when only private beds is used, for Cases A and B, respectively.

From these graphs it is obvious that there exists a threshold value for $r_{11}$ beyond which it is optimal to have almost only private rooms. This threshold value seems to be approximately at the point $r_{21} = c_1/\alpha$ which is in agreement with Lemma 1. It is also interesting to observe that across all the parameter values there are effectively two possible optimal configurations a low percentage and nearly 100% private rooms. In all the experiments in this section, the low number of private rooms $N_1$ is usually slightly more than the average load from isolation patients where the average load is equal to average demand times the expected length of stay for these patients. Our model and approximation methods help identify when each of the two configurations is optimal and the loss from erring.

7. Conclusions

There is a trend toward greater use of private rooms in hospitals motivated by very real concerns about privacy, comfort and infection control. In this paper we modeled a hospital department with private and semi-private rooms as a queueing loss system. We account for two types of patients those requiring isolation and those that don’t, as well as seasonal arrival patterns. We also allow for patients having different recovery rates depending upon their room type as well as different
budget. Unfortunately most studies of the optimal configuration of hospital beds have glossed over

Figure 5  Difference between revenues vs. r_{11}  Case A: r_{21} = 0.5(r_{11} + r_{22})

Figure 6  Percentage of private beds vs. r_{11}  Case B: r_{21} = r_{11}

revenue generation. We formulate and solve the bed configuration optimization problem faced by a hospital building a new facility or reconfiguring an existing facility with a fixed space or monetary budget. Unfortunately most studies of the optimal configuration of hospital beds have glossed over
the effect private rooms can have on access because they have not explicitly factored congestion effects or uncertainty in demand into their analyses. Using a queueing framework we are able to estimate the effects of bed configuration on access. To accomplish this goal we develop a novel fluid based approximation of system performance with the FCFS-T admission control policy. We are
able to check the accuracy of this approximation by deriving an asymptotically optimal admissions policy that we use as a benchmark. These are the main technical results of the paper. The technical results allow us to efficiently and accurately test the performance of different bed configurations under a wide range of parameter settings and thus do a thorough policy analysis.

In a pair of articles Stall (2012b) and Stall (2012a) summarizes the case made for 100% private rooms in hospitals. First it is claimed that there is very strong evidence that private rooms reduce infections and that resistance to private rooms and thus reduced infections is driven by hospital profit seeking. Second it is claimed that it is in fact more profitable to use private rooms partly because nosocomial infections are reduced and thus LOS is shortened and because of isolation needs private beds are more heavily utilized than semi-private rooms. Interestingly the cost benefit analysis of private rooms by Boardman and Forbes (2011) that is cited, almost entirely bases the benefits of private rooms on increased revenues from patients for private rooms. Boardman and Forbes (2011) actually voices skepticism about the infection reduction. Furthermore, Boardman and Forbes (2011) and an interviewee in Stall (2012a) both state that it is questionable if hospitals will be able to charge more for private rooms if all rooms are private. As we noted above the claims by Bobrow and Thomas (2000) about occupancy in semi-private rooms are unsubstantiated. They are also based on the assumption that all rooms are semi-private or all rooms are private. Our results in this paper give a more nuanced perspective.

The arguments in favor of 100% private rooms ignore that hospital renovation and construction projects, like all capital projects, have constraints (these can be both budgetary and physical). Our approach is to maximize revenue and/or access within these constraints. Doing this and taking into account the stochasticity of demand for private and semi-private rooms we find through our extensive numerical experiments that often a mix of room types is optimal. By setting aside sufficient rooms to handle isolation patients the remaining space can be optimally utilized with semi-private rooms. A judicious mix eliminates the under utilization of semi-private rooms claimed

<table>
<thead>
<tr>
<th>Speed up</th>
<th>Type</th>
<th>Mean</th>
<th>Stddev</th>
<th>Max</th>
<th>$c_1 = 1.5$</th>
<th>$c_1 = 1.5, r_1 = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>F-P</td>
<td>1.05</td>
<td>1.46</td>
<td>6.49</td>
<td>0.84</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>F-S</td>
<td>20.33</td>
<td>11.53</td>
<td>41.22</td>
<td>18.86</td>
<td>19.07</td>
</tr>
<tr>
<td></td>
<td>F-D</td>
<td>11.68</td>
<td>7.57</td>
<td>35.45</td>
<td>12.63</td>
<td>6.88</td>
</tr>
<tr>
<td></td>
<td>F-H</td>
<td>5.68</td>
<td>3.97</td>
<td>18.27</td>
<td>5.05</td>
<td>5.31</td>
</tr>
<tr>
<td>1.2</td>
<td>F-P</td>
<td>0.75</td>
<td>1.27</td>
<td>6.75</td>
<td>0.51</td>
<td>0.41</td>
</tr>
<tr>
<td></td>
<td>F-S</td>
<td>15.32</td>
<td>11.80</td>
<td>36.63</td>
<td>13.52</td>
<td>13.63</td>
</tr>
<tr>
<td></td>
<td>F-D</td>
<td>10.70</td>
<td>7.16</td>
<td>34.30</td>
<td>11.69</td>
<td>6.33</td>
</tr>
<tr>
<td></td>
<td>F-H</td>
<td>4.66</td>
<td>3.99</td>
<td>16.75</td>
<td>4.04</td>
<td>4.17</td>
</tr>
</tbody>
</table>

Table 6 Results of numerical experiments with faster recovery rates in private rooms
by Bobrow and Thomas (2000). Another insight from our experiments is that using FCFS-T instead of the near optimal REVMAX$_P$ policy does not affect the revenue or the throughput significantly. Finally we see that the extreme solution of using only private rooms may decrease the revenue (and throughput) significantly. The magnitude of the decrease in revenues depends especially on the relative space requirement of a private room compared to that of a semi-private room. This is despite the fact that we account for isolation patients. For a configuration of all private rooms to be optimal it is necessary to make up the difference in capacity through shorter LOS and/or increased revenues.

Stall (2012b) cites increases of approximately 11% in some types of infections from exposure to roommates i.e. from being in a semi-private room versus a private room. As a rough calculation let’s assume that getting a nosocomial infection doubles one’s length of stay and the base rate of these infections in a private room is 50% which is high. Then the ratio of recovery rates in private rooms to semi-private rooms ($\alpha$ in our experiments) would be only: 1.033. From our numerical experiments we can see that this speed up in recovery is not sufficient by itself to overcome the loss of beds when using only private rooms. This rough calculation oversimplifies the clinical benefits of private rooms. Infections can lead to worse clinical outcomes and even mortality and so focusing on differences in length of stay may not be sufficient for comparing the performance of different bed configurations.

We are also skeptical about claims that revenue will increase because patients prefer private rooms. First, there is increasing pressure to reduce healthcare costs not increase them. Second, if private rooms become the norm, competition will drive down the prices hospitals can charge. Thus the revenue boost from private rooms is also very uncertain. We acknowledge that competition may be driving hospitals to all private bed configurations because given a choice most patients prefer private rooms. But as we note above the greater the competition the less likely that hospitals will be able to reap a price premium for private rooms. Modeling the competition for patients is an interesting area for future research and it is possible that such competition is leading to an equilibrium in which there is less capacity.

Given the limited data on the benefits in terms of LOS of private rooms our results suggest that hospital designers may be over investing in private rooms. Our modeling and the solution methodologies developed in this paper provide tools that facilitate a more complete understanding of the effects of bed configuration decisions on both revenue and access. As better data is collected on the clinical benefits of private rooms these tools can be used to optimize bed configurations.
References


Whitt, W. 1991. The pointwise stationary approximation for $M_t/M_t/s$ queues is asymptotically correct as the rates increase. *Management Science* 37(2) 307–314.
Appendix

In Appendix A we prove Lemma 1. Appendices B–E are devoted to the proofs of Theorems 1 and 2. We first present the queueing model equations in Appendix B. We prove Theorem 1 in Appendix C and Theorem 2 in Appendix D. An auxiliary result needed in the proof of Theorem 2 is proved in Appendix E. In Appendices F and G we present a few extensions of our basic model.

A. Proof of Lemma 1

Assume that (10)–(12) and (15) holds. The result follows from the following observation. Let $N = (N_1, N_2)$ and consider a feasible configuration $N' = (N'_1, 0)$ with $N'_1 = N_1 + N_2/c$. Then

$$
\psi^*(\lambda, N) \leq \psi^*(\lambda, N')
$$

(33)

where $N' = (N'_1, 0)$ with $N'_1 = N_1 + N_2/c$.

Next we prove (33). Consider the optimal solution $s^*_{ij}(\lambda, N)$ of (8). First assume that $s^*_{12}(\lambda, N) = 0$. Then, $s_{11} = s^*_{11}(\lambda, N)$, $s_{21} = s^*_{21}(\lambda, N) + \frac{\mu_{22}}{c\mu_{22}} s^*_{22}(\lambda, N)$, $s_{22} = s_{21} = 0$ is a feasible solution of (8) for configuration $N'$. In addition, by (15)

$$
\sum_{i,j=1,2} r_{ij}s_{ij} - \sum_{i,j=1,2} r_{ij}s^*_{ij}(\lambda, N) = \left( r_{21} \frac{\mu_{21}}{c\mu_{22}} - r_{22} \right) s^*_{22}(\lambda, N) \geq 0,
$$

hence (33).

Now assume that $s^*_{12}(\lambda, N) > 0$. Let $N_{12} = 2s^*_{12}(\lambda, N)/\mu_{12}$. Consider the configuration $N''_1 = N_1 + N_{12}/c$, $N''_2 = N_2 - N_{12}$ and the allocations given by $s_{11} = s^*_{11}(\lambda, N) + \mu_{11} N_{12}/c$, $s_{22} = s^*_{22}(\lambda, N)$, $s_{12} = s_{21} = 0$. Clearly these allocations give a feasible solution of (8) for configuration $N''$. Also

$$
\sum_{i,j=1,2} r_{ij}s_{ij} - \sum_{i,j=1,2} r_{ij}s^*_{ij}(\lambda, N) = \left( r_{11} \frac{\mu_{11}}{c} - \frac{r_{12}\mu_{12}}{2} \right) N_{12} \geq 0.
$$

The desired result then follows from the previous part. □

B. Queuing Model

In this section we provide the details of the queueing model equations under Markovian policies and the additional equations satisfied under the REVMAX policy. Let $Z_{ij}(t)$ denote the number of type $i$ patients being treated in type $j$ rooms at time $t$ for $i, j = 1, 2$. Also, we use $a_{ij}(t)$ to denote the number of type $i$ patients who are admitted to a type $j$ room and $T_{ij}(t)$ to denote the number of type $i$ patients who are transferred to a room of type $j'$ from a room of type $j$ by time $t$, for $i, j = 1, 2$. Let $\{S_{ij}; i, j = 1, 2\}$ denote a set of independent Poisson processes with rate 1 that are
also independent from the arrival processes. We use \( V_{ij}(t) \) to denote the number of type \( i \) patients whose service is preempted while they were in a type \( j \) bed by time \( t \). We set

\[
Z_i(t) = Z_{i1}(t) + Z_{i2}(t) \quad \text{and} \quad Z_i(t) = 2Z_{i2}(t) + Z_{i2}(t).
\]

Also, the number of treated type \( i \) patients in a type \( j \) room is given by

\[
D_{ij}(t) = S_{ij} \left( \mu_{ij} \int_0^t Z_{ij}(s) ds \right), \quad i, j = 1, 2,
\]

and so

\[
D_j(t) = \sum_{i=1}^2 D_{ij}(t), \quad j = 1, 2,
\]

is the total number of patients released from type \( j \) rooms.

Under any Markovian policy we have

\[
Z_{ij}(t) = Z_{ij}(0) + a_{ij}(t) - D_{ij}(t) - T_{ij}(t) + T_{ij'}(t) - V_{ij}(t), \quad i, j = 1, 2.
\]

Also because we restrict our attention to policies that only allow transfers between rooms when there is an arrival or a departure, we have

\[
T_{ij}(t_2) - T_{ij}(t_1) \leq (A_i(t_2) - A_i(t_1)) \land \left( \sum_{i=1,2} (D_i(t_2) - D_i(t_1)) \right), \quad \text{for} \quad i, j = 1, 2
\]

\[
V_{ij}(t_2) - V_{ij}(t_1) \leq (A_i(t_2) - A_i(t_1)) \land \left( \sum_{i=1,2} (D_i(t_2) - D_i(t_1)) \right), \quad \text{for} \quad i, j = 1, 2
\]

and

\[
a_{ij}(t_2) - a_{ij}(t_1) \leq A_i(t_2) - A_i(t_1), \quad \text{for} \quad i, j = 1, 2
\]

for \( 0 \leq t_1 \leq t_2 \).

Next we focus on queueing equations that hold under REVMAX\(_P\). We have \( V_{ii}(t) = 0 \) and the treatment of a type \( i' \) patient in a type \( i \) room will be interrupted if a type \( i \) patient arrives to find all the beds full, hence

\[
V_{ii}(t) = \int_0^t 1 \{ Z_i(s) = N_i \} 1 \{ Z_{i'}(s) = N_{i'} \} 1 \{ Z_{i'}(s) > 0 \} dA_i(s), \quad i = 1, 2.
\]

A type \( i \) arrival will be diverted only if all the beds are full and there are no type \( i' \) patients in a type \( i \) room. Hence \( a_{11} \) and \( a_{22} \) satisfy

\[
a_{11}(t) = \int_0^t (1 \{ Z_{21}(s) > 0 \} 1 \{ Z_1(s) = N_i \} + 1 \{ Z_1(s) < N_i \}) dA_i(s),
\]

(42)
\[ a_{22}(t) = \int_{0}^{t} \left( \mathbf{1}\{Z_1(s-) = N_1\} \mathbf{1}\{Z_2(s-) < N_2\} + \mathbf{1}\{Z_1(s-) = N_1\} \mathbf{1}\{Z_2(s-) = N_2\} \mathbf{1}\{Z_{12}(s-) > 0\} \right) dA_2(s). \] (43)

Because patients are transferred from a type 2 to a type 1 room, whenever possible, with priority given to type 1 patients we have

\[ Z_{12}(t) (N_1 - Z_{11}(t)) = 0, \quad \text{for } t \geq 0 \] (44)

and

\[ Z_2(t) (N_1 - Z_1(t)) = 0 \quad \text{for } t \geq 0. \] (45)

Because isolation patients are never transferred from a private room to a semi-private room, we have that \( T_{11}(t) = 0 \). On the other hand, isolation patients are transferred to a private room whenever a room becomes available, hence

\[ T_{12}(t) = \int_{0}^{t} \mathbf{1}\{Z_{12}(s-) > 0\} dD_1(s). \] (46)

In addition,

\[ T_{21}(t) = \int_{0}^{t} \mathbf{1}\{Z_{21}(s-) > 0\} \mathbf{1}\{Z_2(s-) < N_2\} \mathbf{1}\{Z_1(s-) = N_1\} dA_1(s), \] (47)

since type 2 patients are transferred to a type 2 room from a type 1 bed to open up room for a type 1 patient.

An arriving non-isolation patient is admitted to a private room whenever there is availability. Hence,

\[ a_{21}(t) = \int_{0}^{t} \mathbf{1}\{Z_1(s-) < N_1\} dA_2(s). \] (48)

An isolation patient is admitted to a semi-private room only if all the private rooms are occupied by isolation patients, hence \( a_{12} \) is given by

\[ a_{12}(t) = \int_{0}^{t} \mathbf{1}\{Z_{11}(s-) = N_1\} \mathbf{1}\{Z_2(s-) \leq N_2 - 2\} dA_1(s). \] (49)

Because type 2 patients are transferred to type 1 rooms when a type 1 room becomes available and there are no type 1 patients in a type 2 room we have

\[ T_{22}(t) = \int_{0}^{t} \mathbf{1}\{Z_{12}(s-) = 0\} \mathbf{1}\{Z_{22}(s-) > 0\} dD_1(s). \] (50)
C. Proof of Theorem 1

Consider a sequence of systems indexed by \( n \) that satisfy (21)–(23). Note that it is enough to show that given \( \tilde{\delta} > 0 \), under any sequence of Markovian control policies \( \{\xi^n\} \) and any sequence of configurations \( \{N^n = (N_1^n, N_2^n)\} \) with \( c_1N_1^n + c_2N_2^n \leq nB \) for all \( n \), we have

\[
\tilde{R}^n_c(T^a, \Lambda^n, N) \leq \phi^*(T, \Lambda, B) + \tilde{\delta}
\]

for \( n \) large enough.

Fix a sequence of control policies \( \{\xi^n\} \) and a sequence of configurations \( \{N^n = (N_1^n, N_2^n)\} \) with \( c_1N_1^n + c_2N_2^n \leq nB \) for all \( n \). For the rest of the proof we assume that \( c_i \geq 1 \), \( i = 1, 2 \), without loss of generality. Recall that \( L_i \) is a Poisson processes with rate 1 and \( A_i \) is defined by (1), for \( i = 1, 2 \). We define

\[
\tilde{S}^n_{ij}(t) = \frac{S_{ij}(k^nt)}{k^n} \quad \text{and} \quad \tilde{I}^n_i(t) = \frac{L_i(k^nt)}{k^n}, \quad i, j = 1, 2.
\]

Fix \( \delta > 0 \) and let

\[
\mathcal{A}^n(\delta) = \left\{ \sup_{0 \leq t \leq BT} \left| \tilde{S}^n_{ij}(\mu_{ij}t) - \mu_{ij}t \right| < \delta; i, j = 1, 2 \right\} \cup \left\{ \sup_{0 \leq t \leq MT} \left| \tilde{I}^n_i(t) - t \right| < \delta; i = 1, 2 \right\}.
\]

As in Atar et al. (2010), we have

\[
P \mathcal{A}^n(\delta) \to 1 \quad \text{as} \quad n \to \infty.
\]

We append superscript “\( n \)” to all the queueing processes defined in Appendix B associated with the \( n \)th system. Let

\[
U^n_{ij}(t) = \int_0^t Z^n_{ij}(s) ds,
\]

\[
\tilde{U}^n_{ij}(t) = \frac{U^n_{ij}(k^nt)}{k^n}, \quad \tilde{A}^n_i(t) = \frac{A_i(k^nt)}{k^n}, \quad \tilde{a}^n_{ij}(t) = \frac{a^n_{ij}(k^nt)}{k^n},
\]

and

\[
\tilde{T}^n_{ij}(t) = \frac{T^n_{ij}(k^nt)}{k^n}, \quad \tilde{V}^n_{ij}(t) = \frac{V^n_{ij}(k^nt)}{k^n}, \quad \tilde{D}^n_{ij}(t) = \frac{D^n_{ij}(k^nt)}{k^n}, \quad \tilde{Z}^n_{ij}(t) = \frac{Z^n_{ij}(k^nt)}{k^n}.
\]

By (37)

\[
\tilde{Z}^n_{ij}(t) = \tilde{Z}^n_{ij}(0) + \tilde{a}^n_{ij}(t) - \tilde{D}^n_{ij}(t) - \tilde{T}^n_{ij}(t) + \tilde{V}^n_{ij}(t), \quad i, j = 1, 2,
\]

\[
\tilde{a}^n_{i1}(t) + \tilde{a}^n_{21}(t) \leq \tilde{A}^n_i(t), \quad i = 1, 2.
\]
Clearly

$$\sup_{0 \leq t \leq T} \tilde{Z}^n_{ij}(t) \to 0 \text{ as } n \to \infty$$

(59)

by (23). Also on $\mathcal{A}^n(\delta)$ we have

$$\sup_{0 \leq t \leq T} \left| \tilde{A}^n_i(t) - \int_0^t \Lambda_i(s) ds \right| < \delta, \ i = 1, 2$$

(60)

by (21) and

$$\sup_{0 \leq t \leq T} \left| \tilde{D}^n_{ij}(t) - \mu_{ij} \tilde{U}^n_{ij}(t) \right| < \delta, \ i, j = 1, 2.$$  

(61)

By (57)–(61), for all $0 \leq t \leq T$

$$\left| \mu_{ij} \tilde{U}^n_{ij}(t) - \tilde{A}^n_i(t) - \tilde{T}^n_{ij}(t) + \tilde{V}^n_{ij}(t) \right| \leq 2\delta, \ i, j = 1, 2,$$

(62)

$$\tilde{a}^n_{ij}(t) - \tilde{T}^n_{ij}(t) + \tilde{V}^n_{ij}(t) \geq 0, \ i, j = 1, 2,$$

(63)

$$\tilde{a}^n_{ii}(t) + \tilde{a}^n_{ij}(t) < \int_0^t \Lambda_i(s) ds + 2\delta, \ i = 1, 2,$$

(64)

$$\tilde{U}^n_{11}(t) + \tilde{U}^n_{21}(t) \leq \frac{N^n_1}{n} t,$$

(65)

$$2\tilde{U}^n_{12}(t) + \tilde{U}^n_{22}(t) \leq \frac{N^n_2}{n} t,$$

(66)

on $\mathcal{A}^n(\delta)$. By the continuity of an LP on its constraints, (4)–(8), (22) and (61)–(66) imply that

$$\sum_{i,j=1,2} r_{ij} \tilde{D}^n_{ij}(T) \leq \int_0^T \psi^* \left( \Lambda(s), \frac{N^n}{n} \right) ds + b\delta$$

(67)

for some $b > 0$ on $\mathcal{A}^n(\delta)$. By (53), (22) and (67) imply (51) because $\delta > 0$ is arbitrary. 

\section{D. Proof of Theorem 2}

Consider a sequence of systems indexed by $n$ that satisfy (21)–(23). Assume that the configuration for the $n$th system is given by $N^* n$. The proof is based on the following result which we prove in Appendix E. Let $s^*(\lambda, N)$ denote the optimal solution of (4) with arrival rate $\lambda = (\lambda_1, \lambda_2)$ and bed capacity $N = (N_1, N_2)$ defined as in (13) and (14) and set

$$z_{ij}^* = s_{ij}^*(\lambda, N)/\mu_{ij},$$

for $i, j = 1, 2$ and also set $z^* = (z_{ij}^*; i, j = 1, 2)$. Next we define

$$\tilde{Z}^n_{ij}(t) = \frac{Z^n_{ij}(k^n t)}{n}.$$  

We prove the following (state space collapse) result in Appendix E.
Theorem 3. Consider a sequence of systems indexed by $n$ that satisfy (21)–(23) with configuration $N^n$ such that $N^n/n \to N$ as $n \to \infty$. Under REVMAX$_P$ policy

$$
\lim_{n \to \infty} P \left\{ \sup_{0 \leq t \leq T} \left| \tilde{Z}_{ij}^n(t) - s^*_{ij}(A(t), N)/\mu_{ij} \right| > \epsilon \right\} = 0, \ i, j = 1, 2,
$$

for any $T > 0$ and $\epsilon > 0$.

Now we are ready to prove Theorem 2. Consider a sequence of systems satisfying the conditions of the theorem. Fix $\delta > 0$ and define $A^n(\delta)$ as in (52) and consider the scaling defined in (55) and (56). Note that (61) holds under any policy, hence

$$
\sup_{0 \leq t \leq T} \left| \tilde{D}_{ij}^n(t) - \mu_{ij} \int_0^t \tilde{Z}_{ij}^n(s) ds \right| < \delta, \ i, j = 1, 2,
$$
on $A^n(\delta)$.

By (24), (53) and Theorem 3 under REVMAX$_P$ policy

$$
\tilde{D}_{ij}^n(T) \to \int_0^T s^*_{ij}(\Lambda(t), N^*(T, \Lambda, B)) dt, \ i, j = 1, 2,
$$
in probability as $n \to \infty$. The desired result follows from (22), dominated convergence theorem and (68).

E. Proof of Theorem 3

The proof of Theorem 3 is based on analyzing the limits of the processes $Z^n_{ij}$’s in the asymptotic regime described in §4 under the proposed policy. We show that the limiting processes satisfy a set of equations that are similar to many-server fluid model equations (see Dai and Tezcan (2011)). These equations are known as the hydrodynamic equations in the literature, see Bramson (1998) and Dai and Tezcan (2011) and in the current asymptotic regime they agree with the more traditional many-server fluid model equations. We begin the proof of Theorem 3 by analyzing the steady state of these fluid model equations in §E.1. Then in §E.2 we show that the limiting processes satisfy these fluid model equations and complete the proof of Theorem 3 using the established steady state properties of fluid model solutions, in a similar fashion to Besbes and Maglaras (2009).

E.1. Fluid model

Consider the following equations for $t_1 \leq t_2$.

$$
\tilde{Z}_{ij}(t) = \tilde{Z}_{ij}(0) + \tilde{a}_{ij}(t) - \mu_{ij} \int_0^t \tilde{Z}_{ij}(s) ds - \tilde{T}_{ij}(t) - \tilde{V}_{ij}(t), \ \text{for} \ i, j = 1, 2,
$$

$$
\tilde{Z}_i(t) = \sum_{j=1,2} \tilde{Z}_{ij}(t), \ \text{for} \ i = 1, 2,
$$

$$
\tilde{a}_{11}(t) = \lambda_1, \ \text{if} \ \tilde{Z}_1(t) < N_1 \ \text{or} \ \tilde{Z}_{21}(t) > 0,
$$
\begin{equation}
\hat{a}_{21}(t) = \lambda_2, \text{ if } \bar{Z}_1(t) < N_1, \tag{72}
\end{equation}
\begin{equation}
\hat{a}_{22}(t) = \lambda_2, \text{ if } \bar{Z}_1(t) = N_1 \text{ and } \bar{Z}_{12}(t) > 0, \tag{73}
\end{equation}
\begin{equation}
\hat{a}_{i1}(t) + \hat{a}_{i2}(t) = \lambda_i, \text{ if } \bar{Z}_i(t) < N_1 \text{ or } \bar{Z}_{2}(t) < N_2, \ i = 1, 2, \tag{74}
\end{equation}
\begin{equation}
\hat{a}_{i1}(t) + \hat{a}_{i2}(t) \leq \lambda_i, \ i = 1, 2, \tag{75}
\end{equation}
\begin{equation}
\bar{Z}_2(t)(N_1 - \bar{Z}_1(t)) = 0 \text{ and } \bar{Z}_{12}(t)(N_1 - \bar{Z}_{11}(t)) = 0, \tag{76}
\end{equation}
\begin{equation}
\hat{T}_{22}(t) = 0 \text{ if } \bar{Z}_{12}(t) > 0, \tag{77}
\end{equation}
\begin{equation}
\dot{\bar{V}}_{i'}(t) = 0, \text{ if } \bar{Z}_i(t) < N_i \text{ or } \bar{Z}_{i'}(t) < N_{i'}, \text{ for } i = 1, 2, \tag{78}
\end{equation}
\begin{equation}
\dot{T}_{11}(t) = 0, \dot{\bar{V}}_{ii}(t) = 0, \text{ for } i = 1, 2, \tag{79}
\end{equation}
\begin{equation}
\bar{Z}_{ij}(t) \geq 0 \text{ and } \bar{Z}_{i}(t) \leq N_i, \text{ for } i, j = 1, 2, \tag{80}
\end{equation}
\begin{equation}
\bar{a}_{ij}, \bar{T}_{ij} \text{ and } \bar{V}_{ij} \text{ are nondecreasing for } i, j = 1, 2, \tag{81}
\end{equation}
\begin{equation}
\max_{i,j=1,2} \left\{ \bar{a}_{ij}(t_2) - \bar{a}_{ij}(t_1), \bar{T}_{ij}(t_2) - \bar{T}_{ij}(t_1), \bar{V}_{ij}(t_2) - \bar{V}_{ij}(t_1) \right\} \leq |\lambda_1 + \lambda_2 + N_1 + N_2|(t_2 - t_1), \tag{82}
\end{equation}

where \( \dot{f}(t) = df(t)/dt \) for \( f : \mathbb{R} \to \mathbb{R} \) and for \( t \) where \( f \) is differentiable. These equations can be interpreted as the deterministic equivalent of the queuing equations presented in Appendix B. Let \( \bar{X} = (\bar{Z}, \bar{T}, \bar{a}, \bar{V}) \) denote a solution of the fluid model equations (69)–(82), where \( \bar{Z} = (\bar{Z}_{ij}; i, j = 1, 2) \), \( \bar{T} = (\bar{T}_{ij}; i, j = 1, 2) \), \( \bar{a} = (\bar{a}_{ij}; i, j = 1, 2) \) and \( \bar{V} = (\bar{V}_{ij}; i, j = 1, 2) \). It can be shown as in Dai and Tezcan (2011) using (69), (80) and (82) that \( \bar{X} \) is continuous and differentiable a.e. For a fluid model solution \( \bar{X} \), we refer to \( t \) as a regular point of \( \bar{X} \) if \( \bar{X} \) is differentiable at \( t \). We refer to \( (69)\text{–}(82) \) as the REVMAX\(_P\)-fluid model.

**Theorem 4.** For any solution \( \bar{X} \) of the REVMAX\(_P\)-fluid model with initial conditions \( 0 \leq \bar{Z}_i(0) \leq N_i, \) and \( \bar{Z}_{ij}(0) \geq 0, \ i, j = 1, 2, \) and for any \( \delta > 0, \) there exists \( T^* \) such that, for \( t \geq T^* \)

\[
\|\bar{Z}(t) - z^*\| < \delta.
\]

**Proof of Theorem 4:** Let \( \bar{X} \) be a fluid model solution with \( 0 \leq \bar{Z}_i(0) \leq N_i, \ i = 1, 2, \) and \( \bar{Z}_{ij}(0) \geq 0, \ i, j = 1, 2. \) Also we set \( \mu_m = \min_{i,j=1,2} \{ \mu_{ij} \} \) and assume without loss of generality that \( \mu_m > 0. \) We focus on the case when \( z_{11}^* > 0, z_{21}^* > 0 \) and \( N_2 > z_{22}^* > 0. \) We comment on the how to proceed in the other cases at the end of the proof. Throughout the proof we only consider regular points of \( \bar{X}. \)

Fix \( \delta > 0 \) and take \( b \) small enough such that

\[
b < \min \left\{ 0.5\mu_{22}z_{22}^*/(\delta \mu_{11} + \delta \mu_{21}), (N_1 - z_{11}^*)/\delta, (N_1 - z_{21}^*)/\delta, z_{11}^*/\delta, z_{21}^*/\delta \right\}. \tag{83}
\]

It immediately follows from (69), (71), and (79) that

\[
\bar{Z}_{11}(t) > z_{11}^* - b\delta \tag{84}
\]
for \( t > t_0 = N_1/(b\mu_0\delta) \).

Because \( z_{21}^* > 0 \), \( z_{12}^* = 0 \). Also, if \( \bar{Z}_{12}(t) > 0 \), then \( \bar{Z}_{11}(t) = N_1 \) by (76). Therefore, \( \dot{\bar{Z}}_{11}(t) = 0 \). Also by (75)

\[
\dot{\bar{Z}}_{11}(t) + \dot{\bar{Z}}_{12}(t) \leq \lambda_1 - \mu_{11}N_1 - \mu_{12}\bar{Z}_{12}(t) \leq -\mu_{11}(N_1 - z_{11}^*). 
\]

This implies by (83) that \( \bar{Z}_{12}(t) = 0 \) for \( t \geq t_1 = t_0 + N_2/(\mu_{11}(N_1 - z_{11}^*)) \). Then, by (69), (71), (79) and (83)

\[
\bar{Z}_{11}(t) < z_{11}^* + b\delta
\]

for \( t \geq t_2 = t_1 + N_1/(b\mu_{11}\delta) \).

By our assumption \( z_{22}^* > 0 \). Then, if \( \bar{Z}_1(t) < N_1 \), we have by (84), (85), (71) and (72) that for \( t \geq t_2 \)

\[
\dot{\bar{Z}}_1(t) \geq -b\delta\mu_{11} - b\delta\mu_{21} + \mu_{22}z_{22}^*.
\]

Hence, by (83) and (86)

\[
\bar{Z}_1(t) = N_1
\]

for \( t \geq t_3 = t_2 + N_1/(0.5\mu_{22}z_{22}^*) \). This by (84) and (85) also gives

\[
|\bar{Z}_{21}(t) - z_{21}^*| < b\delta
\]

for \( t \geq t_3 \). Finally, we focus on \( \bar{Z}_{22} \). Assume that \( \bar{Z}_{22}(t) < z_{22}^* - b_2\delta \), for \( b_2 = 4\mu_{11}/\mu_{22}b \) and assume also without loss of generality that \( b_2\delta < \min\{z_{22}^*, N_2 - z_{22}^*\} \) by taking \( b \) small enough. Note that, by (84) and (85),

\[
-b\delta\mu_{11} \leq \dot{\bar{Z}}_{11}(t) \leq b\delta\mu_{11}
\]

and by (87) \( \dot{\bar{Z}}_1(t) = 0 \), hence

\[
-b\delta\mu_{11} \leq \dot{\bar{Z}}_{21}(t) \leq b\delta\mu_{11}
\]

for \( t \geq t_3 \). By (74), (78) and (88), this implies for \( t \geq t_3 \) that

\[
\dot{\bar{Z}}_{21}(t) + \dot{\bar{Z}}_{22}(t) \geq -\mu_{11}b\delta + \mu_{22}b_2\delta.
\]

By (89) and (90),

\[
\bar{Z}_{22}(t) \geq z_{22}^* - b_2\delta
\]
for \( t \geq t_4 = t_3 + b_2(N_1 + N_2)/(\mu_m \delta) \). Now assume that \( \hat{Z}_{22}(t) > z_{22}^* + b_2 \delta \). Then it follows using an argument similar to that leading to (91) that

\[
\hat{Z}_{22}(t) \leq z_{22}^* + b_2 \delta
\]  

(92)

for \( t \geq t_5 = t_5 + b_2(N_1 + N_2)/(\mu_m \delta) \) proving the result. If \( z_{11}^* > 0, z_{21}^* > 0 \) and \( z_{22}^* = N_2 \), in the last step we take \( b_2 \delta < z_{22}^* \) and result follows from (91).

Assume now that \( z_{11}^* > 0, z_{21}^* > 0 \) and \( z_{22}^* = 0 \). If \( z_{11}^* + z_{21}^* = N_1 \), then (84), (85), (88) still hold. The result then follows from a similar argument leading to (92). If \( z_{11}^* + z_{21}^* < N_1 \), (84) and (85) still hold. Then, if \( \hat{Z}_{22}(t) > 0 \), we have that \( \hat{Z}_1(t) = N_1 \) by (76). Hence it can be shown using an argument similar to that leading to (90) that \( \hat{Z}_{22}(t) = 0 \) for \( t \) large enough. Then, it can be shown using (84) and (85) that \( |\hat{Z}_{22}(t) - z_{21}^*| < b_2 \delta \) for \( t \) large enough.

Assume now that \( z_{11}^* = N_1, z_{21}^* = 0 \) and \( z_{12}^* \geq 0 \). It follows from a similar argument leading to (85) that \( \hat{Z}_{11}(t) = N_1 \) for \( t \) large enough. Similarly, \( |\hat{Z}_{22}(t) - z_{22}^*| < b_2 \delta \) for \( t \) large follows from (69) and (73). The result for \( \hat{Z}_{12} \) follows from these results. \( \square \)

### E.2. Proof of Theorem 3

In this section we prove Theorem 3. Following Bramson’s notations (cf. [Bramson (1998), equation (5.3)]), we define for \( m = 1, 2, \ldots, [k^a T] \),

\[
\begin{align*}
\hat{Z}_{ij}^{n,m}(t) &= Z_{ij}^n(t+m)/n  \\
\hat{D}_{ij}^{n,m}(t) &= n^{-1}(D_{ij}^n(t+m) - D_{ij}^n(m))  \\
\hat{A}_{ij}^{n,m}(t) &= n^{-1}(A_{ij}^n(t+m) - A_{ij}^n(m))  \\
\hat{a}_{ij}^{n,m}(t) &= n^{-1}(a_{ij}^n(t+m) - a_{ij}^n(m))  \\
\hat{U}_{ij}^{n,m}(t) &= n^{-1}(U_{ij}^n(t+m) - U_{ij}^n(m))  \\
\hat{T}_{ij}^{n,m}(t) &= n^{-1}(T_{ij}^n(t+m) - T_{ij}^n(m))  \\
\hat{V}_{ij}^{n,m}(t) &= n^{-1}(V_{ij}^n(t+m) - V_{ij}^n(m))  \\
\Lambda_{ij}^{n,m}(t) &= \Lambda(t+m)/n,
\end{align*}
\]  

(93)

where \( U_{ij}^{n} \) is defined as in (54). Let \( X^{n,m} = (\hat{Z}^{n,m}, \hat{D}^{n,m}, \hat{A}^{n,m}, \hat{a}^{n,m}, \hat{U}^{n,m}, \hat{T}^{n,m}, \hat{V}^{n,m}) \), with \( \hat{Z}^{n,m} = (\hat{Z}_{ij}^{n,m}; i,j = 1, 2) \) and other vector processes are defined similarly and let \( d \) denote the dimension of \( X^{n,m} \). Also let \( \hat{Z}_{11}^{n,m}(t) = \hat{Z}_{12}^{n,m}(t) + \hat{Z}_{22}^{n,m}(t) \) and \( \hat{Z}_{22}^{n,m}(t) = 2\hat{Z}_{12}^{n,m}(t) + \hat{Z}_{22}^{n,m}(t) \).

Under the policy REVMAX\(_p\), the following equations hold by (37)–(40).

\[
\begin{align*}
\hat{Z}_{ij}^{n,m}(t) &= \hat{Z}_{ij}^{n,m}(0) + \hat{a}_{ij}^{n,m}(t) - \hat{D}_{ij}^{n,m}(t) - \hat{T}_{ij}^{n,m}(t) + \hat{V}_{ij}^{n,m}(t), \quad i,j = 1, 2,  \\
\hat{a}_{11}^{n,m}(t_2) - \hat{a}_{11}^{n,m}(t_1) &= \hat{A}_{11}^{n,m}(t_2) - \hat{A}_{11}^{n,m}(t_1),
\end{align*}
\]  

(94)
if \( \hat{Z}_{1,n}^m(s) < N_1 \) or \( \hat{Z}_{2,n}^m(s) > 0 \) for all \( s \in [t_1, t_2] \),
\[
\hat{a}_{21}^n(t_2) - \hat{a}_{21}^n(t_1) = \hat{A}_2^m(t_2) - \hat{A}_2^m(t_1), \quad \text{if } \hat{Z}_{1,n}^m(s) < N_1 \text{ for all } s \in [t_1, t_2],
\]
\[
\hat{a}_{22}^n(t_2) - \hat{a}_{22}^n(t_1) = \hat{A}_2^m(t_2) - \hat{A}_2^m(t_1), \quad \text{if } \hat{Z}_{2,n}^m(s) > 0 \text{ for all } s \in [t_1, t_2],
\]
\[
\hat{a}_{11}^n(t_2) - \hat{a}_{11}^n(t_1) + \hat{a}_{12}^n(t_2) - \hat{a}_{12}^n(t_1) = \hat{A}_1^m(t_2) - \hat{A}_1^m(t_1),
\]
if \( \hat{Z}_{1,n}^m(s) < N_1 \) or \( \hat{Z}_{2,n}^m(s) < N_2 \) for all \( s \in [t_1, t_2] \), \( i = 1, 2 \),
\[
\hat{a}_{i1}^n(t_2) - \hat{a}_{i1}^n(t_1) + \hat{a}_{i2}^n(t_2) - \hat{a}_{i2}^n(t_1) \leq \hat{A}_i^m(t_2) - \hat{A}_i^m(t_1), \quad \text{for } i = 1, 2,
\]
\[
\hat{Z}_{12}^n(t) \left( N_1^n - \hat{Z}_{11}^n(t) \right) = 0, \quad \hat{Z}_{22}^n(t) \left( N_2^n - \hat{Z}_{21}^n(t) \right) = 0
\]
\[
\hat{T}_{22}^n(t_2) - \hat{T}_{22}^n(t_1) = 0, \quad \text{if } \hat{Z}_{12}^n(s) > 0 \text{ for all } s \in [t_1, t_2],
\]
\[
\hat{V}_{i1}^n(t_2) - \hat{V}_{i1}^n(t_1) = 0, \quad \text{if } \hat{Z}_{i1}^n(s) < N_i \text{ for all } s \in [t_1, t_2], \quad i = 1, 2,
\]
\[
\hat{T}_{11}^n(t) = 0 \quad \text{and } \hat{V}_{i2}^n(t) = 0, \quad i = 1, 2,
\]
\[
\hat{Z}_{ij}^n(t_2) \geq 0 \quad \text{and } \hat{Z}_{ij}^n(t) < N_i, \quad \text{for } i, j = 1, 2,
\]
\[
\hat{a}_{ij}^n, \hat{T}_{ij}^n \text{ and } \hat{V}_{ij}^n \text{ are nondecreasing for } i, j = 1, 2.
\]

We use these equations below to show that \( X_{i,n}^m \) satisfy the fluid model equations given in Appendix E.1 in the limit. Before we can proceed with the proof we need some additional results. First we begin with establishing a result for arrivals and service completion processes.

**Lemma 2.** Fix \( \delta > 0 \) and \( L > 0 \). For \( n \) large enough, under any Markovian policy
\[
P_A \left( \max_{m \leq k \leq T} \left\| \hat{A}_i^m(t) - \int_0^t \Lambda_i^m(u) du \right\|_L > \delta \right) < \delta, \quad i = 1, 2 \tag{106}
\]
\[
P_A \left( \max_{m \leq k \leq T} \left\| \hat{D}_{ij}^m(t) - \mu_{ij} \hat{U}_{ij}^m(t) \right\|_L > \delta \right) < \delta, \quad i, j = 1, 2 \tag{107}
\]
The proof is similar to that of Lemma 1 in Besbes and Maglaras (2009) hence skipped. One can choose a sequence \( \delta(n) \) decreasing to 0 sufficiently slowly such that the inequalities (106) and (107) still hold. For such a sequence and \( C \geq \limsup_{n \to \infty} \left( \frac{\| \Lambda_i^m \|}{n} \right) \mu_{ij} \vee M \), let
\[
K_0^n = \left\{ \sup_{t_1, t_2 \leq L} |X_{i,n}^m(t_2) - X_{i,n}^m(t_1)| \leq C|t_2 - t_1| + \delta(n), \forall m < k^n T \right\},
\]
\[
K_1^n = \left\{ \max_{m \leq k \leq T} \left\| \hat{A}_i^m(t) - \int_0^t \Lambda_i^m(u) du \right\|_L < \delta(n), \max_{m \leq k \leq T} \left\| \hat{D}_{ij}^m(t) - \mu_{ij} \hat{U}_{ij}^m(t) \right\|_L < \delta(n) \right\},
\]
\[
K^n = K_0^n \cap K_1^n.
\]

Paralleling the line of arguments leading to Bramson (1998), Proposition 5.2, Corollary 5.1, one can establish the following result using Lemma 2.

**Lemma 3.** Under any Markovian policy
\[
P(K^n) \to 1 \text{ as } n \to \infty.
\]
Let $E$ be the space of RCLL functions $x : [0, L] \rightarrow \mathbb{R}^d$ and let

$$E' = \{ x \in E : |x(0)| \leq N, |x(t_2) - x(t_1)| \leq C(t_2 - t_1) \text{ for all } t_1, t_2 \in [0, L] \},$$

$$E_0^n = \{ X^{n,m}(\cdot, \omega) : m < k^nT, \omega \in K^n \},$$

$$\mathcal{E}_0 = \{ E_0^n, n \in \mathbb{N}^+ \} ,$$

Similar to Proposition 6.1 in Bramson (1998) we have the following result.

**Lemma 4.** Fix $\delta > 0$, $L > 0$ and choose $n$ large enough. Then for $\omega \in K^n$ and any $m < k^nT$,

$$\left\| X^{n,m}(\cdot, \omega) - \hat{X}(\cdot) \right\|_L \leq \delta$$

for some cluster point $\hat{X}$ of $\mathcal{E}_0$ with $\hat{X} \in E'$, under any Markovian policy.

Next we consider the cluster points of $\mathcal{E}_0$.

**Lemma 5.** Under a Markovian policy any cluster point $\hat{X}$ of $\mathcal{E}_0$ has its component $\hat{\lambda}$ constant.

**Proof of Lemma 5:** Let $\hat{X}$ be a cluster point of $\mathcal{E}_0$. Then, there exists a sequence $(n^\ell, m^\ell, \omega^\ell)$, with $\omega^\ell \in K^\ell$, such that

$$\left\| X^{n^\ell, m^\ell}(\cdot, \omega^\ell) - \hat{X}(\cdot) \right\|_L \to 0$$

as $\ell \to \infty$. Assume for some $t_2 \neq t_1$ that

$$\hat{\lambda}(t_2) > \hat{\lambda}(t_1). \quad (109)$$

Let $\ell$ denote a subsequence such that $m^\ell / k^\ell \to m'$ as $\ell \to \infty$. Existence of a such subsequence is guaranteed by the fact that $m^\ell \leq k^\ell T$ (recall the definition of $m^\ell$). Then by (21) and (93)

$$\Lambda^{n^\ell,m^\ell}(t) = \Lambda \left( \frac{t}{k^\ell} + \frac{m^\ell}{k^\ell} \right).$$

Therefore by the continuity of $\Lambda$, we have

$$\hat{\lambda}(t_2) = \lim_{\ell \to \infty} \Lambda^{n^\ell,m^\ell}(t_2) = \lim_{\ell \to \infty} \Lambda \left( \frac{t_2}{k^\ell} + \frac{m^\ell}{k^\ell} \right) = \Lambda(m') = \lim_{\ell \to \infty} \Lambda \left( \frac{t_1}{k^\ell} + \frac{m^\ell}{k^\ell} \right) = \lim_{\ell \to \infty} \Lambda^{n^\ell,m^\ell}(t_1) = \hat{\lambda}(t_1).$$

Hence, (109) cannot hold. \qed

Finally, we prove that the cluster points under REVMAX$_p$ satisfy the fluid model equations presented in Appendix E.1.

**Lemma 6.** Fix $L > 0$. Then all cluster points $\hat{X}$ of $\mathcal{E}_0$ under REVMAX$_p$ are solutions of the REVMAX$_p$-fluid model equations with $\lambda = \hat{\lambda}$. 

Proof of Lemma 6: Let \( \hat{X} \) be a cluster point of \( E_0 \). For a given \( \eta > 0 \), choose \((n,m,\omega)\) so that

\[
\left\| X^{n,m}(\cdot,\omega) - \hat{X}(\cdot) \right\|_L \leq \eta/4 \tag{110}
\]

and \( \delta(n) < \eta/4 \) in (108).

Consider the policy REVMAX\(_P\). We have that

\[
\left| \hat{D}_{ij}(t) - \mu_{ij} \int_0^t \hat{Z}_{ij}(t) \right| \leq \left| \hat{D}_{ij}(t) - \hat{D}_{ij}^{n,m}(t) \right| + \left| \hat{D}_{ij}^{n,m}(t) - \int_0^t \hat{Z}_{ij}^{n,m}(s) ds \right| + \left| \int_0^t \hat{Z}_{ij}^{n,m}(s) ds - \int_0^t \hat{Z}_{ij}(s) ds \right| < \eta L
\]

by (108) and (110). Hence by (94) and (110) every cluster points satisfies (69).

Next we prove that \( \hat{X} \) satisfies (71) with \( \hat{\lambda} \). Assume that \( \hat{Z}_1(t) < N_1 - \theta \) for some \( \theta < \eta \), by reselecting \( \eta > 0 \) if necessary. Then, by (110), there exists \( t > t_o > 0 \) such that

\[
\hat{Z}_1^{n,m}(u) < N_1 - \theta/2,
\]

for all \( u \in [t - t_o, t + t_o] \). Hence by (95)

\[
\hat{a}_{11}^{n,m}(t + t_o) - \hat{a}_{11}^{n,m}(t - t_o) = \hat{A}_{1}^{n,m}(t + t_o) - \hat{A}_{1}^{n,m}(t - t_o),
\]

giving (71) by (108) and (110) if \( \hat{Z}_1(t) < N_1 \). Similarly, if \( \hat{Z}_{21}(t) > \theta \) for \( \theta > 4\eta \), there exists \( t > t_o > 0 \) and such that

\[
\hat{Z}_{21}^{n,m}(u) > \theta/2,
\]

for all \( u \in [t - t_o, t + t_o] \). Giving (71) when \( \hat{Z}_{21}(t) > 0 \) by (95), (108) and (110). The fact that the cluster points satisfy the other REVMAX\(_P\)-fluid model equations is proved similarly using (94)–(105). \( \square \)

Now we are ready to prove Theorem 3.

Proof of Theorem 3: Fix \( \eta, \epsilon > 0 \). Let \( L = T^* + 1 \), where \( T^* \) is chosen as in Theorem 4. Choose \( r \) large enough so that \( L/k^n < \eta \). For \( u \in (L/k^n, T] \), let

\[
m_n(u) = \min\{m \in \mathbb{N} : m \leq k^n u \leq m + L\}.
\]

Also note that \( \tau' = k^n u - m_n(u) \geq L - 1 \geq T^* \). We have

\[
\hat{Z}_{ij}^n(u) = \frac{Z_{ij}^n(k^n u)}{n} = \frac{Z_{ij}^n(\tau' + m_n(u))}{n} = \hat{Z}_{ij}^{n,m_n(u)}(\tau').
\]
For all \( \omega \in K^n \) and for all \( u \in (\eta, T] \),
\[
\left| \hat{Z}_{ij}^n(u) - s^*(\Lambda(u), N) \right| = \left| \hat{Z}_{ij}^{n,m}(u)(\tau') - s^*_ij(\Lambda^{n,m}(u)(\tau'), N) \right|
\leq \left| Z_{ij}^{n,m}(u)(\tau') - \hat{Z}_{ij}(\tau') \right| + \left| \hat{Z}_{ij}(\tau') - s^*_ij(\hat{\Lambda}(\tau'), N) \right| + \left| s^*_ij(\hat{\Lambda}(\tau'), N) - s^*_ij(\Lambda^{n,m}(u)(\tau'), N) \right|
\leq 3\epsilon,
\]
for \( n \) large enough, where \( (a) \) follows from Lemmas 4 and 6, Theorem 4, and the fact that \( s^* \) is a continuous function. The desired result now follows from Lemma 3. \( \Box \)

**F. Extensions of the basic model**

In this section we discuss two extensions of our basic model to include the features explained in Remarks 2 and 3. We first focus on the case when an arriving non-isolation patient can only be admitted with probability \( 0 < p_l < 1 \) when there is only one bed left in a semi-private room. Then, we consider the case when a certain fraction of non-isolation patients prefer to board alone.

**F.1. Partial admission**

In our original model we assume that when there is only one semi-private bed left an arriving non-isolation patient is always admitted. However, as described in Remark 2, this is not the case in certain situations, for example because of gender differences between patients in an adult ward. Assume that an arriving non-isolation patient can only be admitted with probability \( 0 < p_l < 1 \) when there is only one bed left in a semi-private room. (Obviously, in our original model \( p_l = 1 \).) Also, if \( p_l = 0 \) then it is possible to obtain simple approximations because the admission probabilities for each class would be the same in all situations.

For the rest of this section we focus on the case when \( 0 < p_l < 1 \) which is common in practice. For the sake of simplicity first assume that \( N_1 = 0 \). Using the approach described in §5, the number
of available beds process can be modeled as a Markov chain, denoted by $X_t$, with transition rates given in Figure 8. Unfortunately, unlike the case in §5, this Markov chain is not time reversible and it is not possible to find closed form solutions for its steady state. (It is easy to show that $X_t$ is positive recurrent.) However, numerical methods can be used to find the stationary distribution of $X_t$ using the transition rates and the flow balance equations as we explain next. Let $\{\pi_i, i = 1, 2, \ldots\}$ denote the stationary distribution of $X_t$. In the current case the flow balance equations can be written as follows.

$$\lambda_1 (1 - \pi_0 - \pi_1) = \mu_{21} z_{21},$$

$$\lambda_2 (1 - \pi_0 - (1 - p_l) \pi_1) = \mu_{22} z_{22}. \tag{111}$$

Also,

$$z_{21} + z_{22} = N - \sum_{i=1}^{\infty} i \pi_i. \tag{113}$$

Using the solutions to recursive equations or working with the general balance equations and using (111) and (112), it is possible to express $\pi_n$, for $n \geq 2$, in terms of $\pi_0$, $\pi_1$, and the service and arrival rates. We do not pursue the details in this paper. A solution for $\pi_0$ and $\pi_1$ then can be obtained using (113) and the fact that $\sum_i \pi_i = 1$.

When there are two types of rooms, approximations for the admission probabilities can also be obtained using the Markov chain $X_t$ described above (assuming the same transfer procedures described in §3.2 are used). Specifically, if (17) holds, then the approximations for the admission probabilities are still given by (19) and if $\frac{N_i}{\rho_{i1}} = \frac{N_2}{\rho_{22}}$, we set $P_i = N_i / \rho_{ii}$, $i = 1, 2$. If (18) holds, approximations for the admission probabilities $\hat{P}_1$ and $\hat{P}_2$ can be found by solving for the stationary distribution of the Markov chain $X_t$ with transition rates given in Figure 8. Also, as in §3.2, if $\hat{P}^2 \lambda_i > N_i / \rho_{i1}$, then we set $P_1 = \hat{P}^2$ and $P_2 = \hat{P}$, otherwise we set $P_1 = \frac{\mu_{11} N_1 + 0.5 \rho_{i2} \hat{N}_{12}}{\lambda_i}$ and $P_2 = (N_2 / \rho_{22}) \wedge 1$, where $\hat{N}_{12} = (N_2 - \rho_{22})^+$. It can be shown that the REVMAX policy is asymptotically optimal for the case when $0 < p_l < 1$ as well.

F.2. Three types of patients

In this section we consider the case when a certain fraction of non-isolation patients prefer to board in private rooms. To model these patients we assume that besides isolation and non-isolation patients there is a third type of patients, referred to as private non-isolation patients. Although these patients do not have infectious diseases, we assume that they prefer not to share a room with another patient. Unlike isolation patients, private non-isolation patients can be boarded in a
semi-private room with another patient. Hence, if there is only one bed available in a semi-private room, a certain proportion of private non-isolation patients may choose to be admitted (recall that isolation patients cannot be admitted in that case). However, under this assumption the analysis becomes even more complicated than that described in Appendix F.1. To obtain simple approximations, we assume that these patients are only admitted if there is an available private room or two semi-private beds instead, as is the case for isolation patients. Let $\lambda_3$ denote the arrival rate and $1/\mu_{3i}$ denote the average LOS for this patient type in room type $i = 1, 2$.

We make the following assumptions regarding the revenues and service rates corresponding to (10)–(12) for the ease of exposition (other cases can be handled similarly). We assume that it is more profitable to board isolation and private non-isolation patients in private rooms and to board non-isolation patients in semi-private rooms. In addition we assume that private non-isolation patients are more profitable than isolation patients. Therefore,

$$r_{i1}\mu_{i1} \geq r_{21}\mu_{21} \text{ and } 0.5r_{i2}\mu_{i2} \leq r_{22}\mu_{22}, \text{ for } i = 1, 3$$

(114)

and

$$r_{1j}\mu_{1j} \leq r_{3j}\mu_{3j}, \text{ for } j = 1, 2.$$ 

(115)

We also assume boarding patients of all types in a private room is more profitable, that is,

$$r_{i1}\mu_{i1} \geq r_{i2}\mu_{i2} \text{ for } i = 1, 2, 3.$$ 

(116)

Under (114)–(116), we use the same transfer procedures described in §3.2 with the following adjustments. When a patient is discharged from a private room, a patient from one of the semi-private rooms is transferred to that private room with preference given to private non-isolation patients. If an isolation or a private non-isolation patient arrives to find all private rooms taken, a non-isolation patient boarding in a private room (if there is any) is transferred to a semi-private room with an empty bed (if there is any available) and the arriving patient is admitted to the private room that became available.

To obtain simple approximations for the performance of the FCFS-T policy we use the following procedure. Because patients arrive according to Poisson processes, we can consider type 1 and 3 patients together as one patient type and consider a weighted average service time. Let

$$\mu'_{1j} = \left(\frac{\lambda_1}{\lambda_1 + \lambda_3 \mu_{1j}} + \frac{\lambda_3}{\lambda_1 + \lambda_3 \mu_{3j}}\right)^{-1}, \text{ for } j = 1, 2.$$
By using the combined service rates $\mu'_{ij}$ and the total arrival rate $\lambda'_1 = \lambda_1 + \lambda_3$ instead of $\mu_{1j}$ and $\lambda_1$, respectively, we can use the approximations in §3.2 to estimate the admission probabilities. Obviously, the service times of the combined patient type are not exponential. However, based on our numerical experiments the service time distribution does not have a significant impact on the admission probabilities beyond its mean. Hence, the suggested approximations for this three-class case should be reasonably accurate.

In the presence of private non-isolation patients, the asymptotically optimal policies depend on the revenue and the service rates of the three patient types. Under the assumptions (114)–(116), the REVMAX$_P$ policy should be modified as follows. First the transfer procedures explained above should be used. Also, the private non-isolation patients should be given preemptive priority over other patient types in private rooms and non-isolation patients should be given preemptive priority over other patient types in semi-private rooms. In addition, isolation patients should be given preemptive priority over non-isolation patients in private rooms and private non-isolation patients should be given preemptive priority over isolation patients in semi-private rooms.

G. Other Extensions

So far our main focus in this paper is on the case when (10) holds. In this section we provide the generalizations for our approximations under FCFS-T and provide asymptotically optimal policies for other cases. Our goal is here is not to be exhaustive but to illustrate that our approach is general and can be used in other cases as well. For simplicity we still assume that (12) holds, hence boarding in a private room is always preferred over boarding in a semi-private room for both patient types.

First assume that it is more profitable to admit non-isolation patients in both types of rooms, therefore,

$$r_{21}\mu_{21} \geq r_{11}\mu_{11} \text{ and } 0.5r_{12}\mu_{12} \leq r_{22}\mu_{22}.$$ 

There are two additional situations that need to be considered when isolation patients are present based on whether it is more profitable to board them in private rooms or semi-private rooms instead of non-isolation patients. For simplicity we only consider the former, that is,

$$r_{11}\mu_{11} + 2r_{22}\mu_{22} > r_{12}\mu_{12} + r_{21}\mu_{21}.$$ 

In this particular case, the same transfer procedures described in the original case in §3.2 should be used and the approximations proposed in that section for FCFS-T are still valid. For asymptotically
optimal policies, the same patient transfer procedures used under FCFS-T should be used and
the treatment of isolation patients are preempted, preference given those boarding in semi-private
rooms, whenever necessary.

Now assume that instead of (10) and (11) the following holds

\[ r_{21} \mu_{21} \geq r_{11} \mu_{11} \text{ and } 0.5 r_{12} \mu_{12} \geq r_{22} \mu_{22}. \]  

(117)

That is, it is more profitable to board type 1 patients in semi-private rooms and type 2 patients
in private rooms. (We do not claim that this parameter regime is realistic.) First we present the
version of FCFS-T that should be used and then the approximations for its performance.

When (117) holds, the same transfer and admission procedures under FCFS-T should be used
except that if a patient is discharged from a private room then a patient from one of the semi-
private rooms is transferred to that private room with preference given to non-isolation patients.
Also, if a non-isolation patient arrives to find all private rooms taken, an isolation patient boarding
in a private room (if there is any) is transferred to an empty semi-private room (if there is any
available) and the arriving non-isolation patient is admitted to the vacated private room. Recall
that for an isolation patient to be admitted to a semi-private room both beds in that room must
be empty. Using an approach similar to that in §5, the following approximations for the FCFS-T
admission policy are obtained. If

\[ \frac{\mu_{21}}{\mu_{22}} (\rho_{21} - N_1)^+ \leq (N_2 - 2 \rho_{12})^+ \text{ and } \frac{\mu_{12}}{\mu_{11}} (\rho_{12} - 0.5 N_2)^+ \leq (N_1 - \rho_{21})^+, \]  

(118)

then \( P_1 = P_2 = 1 \). If (118) does not hold and \( \frac{N_1}{\rho_{21}} = \frac{N_2}{2 \rho_{12}} \) we set \( P_1 = \frac{N_1}{\rho_{21}} \) and \( P_2 = \frac{N_2}{2 \rho_{12}} \). If

\[ \frac{N_1}{\rho_{21}} > \frac{N_2}{2 \rho_{12}} \]

we set \( P_1 = P_2 = P \), where \( P \) is the unique root of

\[ -\left( \rho_{12} + \frac{\mu_{11}}{\mu_{12}} \rho_{21} \right) P^2 + \left( \rho_{12} + \frac{\mu_{11}}{\mu_{12}} \rho_{21} + \frac{\mu_{11}}{\mu_{12}} N_1 + \frac{N_2}{2} + 1 \right) P - \left( \frac{\mu_{11}}{\mu_{12}} N_1 + \frac{N_2}{2} \right) P = 0 \]

between 0 and 1. If (118) does not hold and \( \frac{N_1}{\rho_{21}} < \frac{N_2}{2 \rho_{12}} \) let \( \hat{P} \) denote the unique root of

\[ -2 \rho_{12} P^3 + (2 \rho_{12} - \rho_{22}) P^2 + \left( \rho_{22} + 1 + \frac{\mu_{21}}{\mu_{22}} N_1 + N_2 \right) P - \left( \frac{\mu_{21}}{\mu_{22}} N_1 + N_2 \right) = 0, \]

between 0 and 1. We set \( P_1 = \hat{P}^2 \) and \( P_2 = \hat{P} \). The REVMAX\(_P\) policy should be modified as
follows. The same patient transfer procedures described above for FCFS-T should be used. Also
the treatment of isolation patients in private rooms should be preempted for non-isolation patients
and non-isolation patients in semi-private rooms for isolation patients, whenever necessary.
H. Patient transfers

One important performance characteristic we ignored in our model is the number of transfers. A patient transfer adds to the staff workload and so imposes a cost on the system. Transfers may also disrupt patient care and in the case of infectious patients expose more people to the virus. In this section we attempt to assess the impact of room configuration on the number of patient transfers. First we focus on the FCFS-T policy we described in Section 3.2 and then we discuss a slight modification. Under FCFS-T there are three different scenarios when a patient must be transferred. i) If a patient is discharged from a private room, one of the patients from semi-private rooms (if there are any) is transferred to that private room, ii) if an infectious patient arrives to find all private rooms full, one of the non-infectious patients in a private room, if there are any, is transferred to a semi-private room, iii) if an infectious patient is admitted to a semi-private room, a non-infectious patient in a semi-private room may need to be transferred to another semi-private room, to open up a room for the infectious patient. The transfers under scenarios (i) and (ii) are easy to keep track of, however the transfers under case (iii) require more attention. Throughout the paper we do not consider which semi-private room a non-infectious patient is admitted to when there are multiple available rooms, mainly because this is irrelevant for capacity decisions. For the purpose of evaluating the the total number of transfers under the FCFS-T policy we derive upper and lower bounds for the number of transfers caused by the third scenario. An upper bound for the number of transfers in the third scenario is found by assuming that a non-infectious patient is always needed to be transferred when an infectious patient is admitted to a semi-private room. A lower bound is obtained by assuming that such a transfer is never needed. In practice the number of patient transfers under the third scenario should clearly be between these two bounds.

Now we simulate a hospital ward with parameters used in the numerical experiments described in §6.2. We use the arrival rates given on left side of Figure 3, assuming that $\lambda_0 = 100$, $\gamma_m = 0.7$, $p_{max} = 0.3$, $p_{min} = 0.1$. We set $\mu_{ij} = 0.5$, for $i,j = 1, 2$, $c_1 = 1.5$, $c_2 = 1$, $r_1 = r_2 = 1$ and we assume there is sufficient space to build 200 semi-private beds. (The other parameters in §6.2 give similar results.) The results are reported in Figure 9(a). In this figure we show the upper and the lower bounds for the number of transfers divided by the number of admitted patients vs. the percentage of private rooms. We also show the admission ratio. The optimal configuration (in terms of admission probability) is approximately 25% private rooms. Near the optimal solution, the number of transfers is approximately 40% of admitted patients, which is fairly high.

The main reason why so many transfers are required is that non-infectious patients are always admitted to a private-room first and they are also moved to a private room whenever such a room
opens up. We observed in our simulations that a large proportion of non-infectious patients who are moved to a private room have to be transferred back to semi-private rooms when an infectious patient arrives. From the perspective of transfers this is clearly inefficient and is a result of how FCFS-T is designed. FCFS-T is designed this way to enable our asymptotic analysis however it is possible to decrease the number of transfers by slightly modifying the proposed policies. In the modified policy we assume that non-infectious patients are always admitted to semi-private rooms whenever it is possible and they are not transferred to a single room whenever a single room becomes available. Also, we assume that an infectious patient in a semi-private room is transferred to a private room only when all semi-private rooms are full and a non-infectious patient needs to be admitted. If the rate a non-infectious patient recovers is independent of the room type that patient is boarded, it is easy to show that such a modification does not affect the probability that a patient is admitted (we intrinsically assume that an infectious patient recovers at the same rate in either type of room since an infectious patient always boards alone). Even if patients recover faster in private rooms, we expect that when the system is overloaded, all the private rooms are full. Hence, the system should still have the same capacity after this slight modification.

Next we run the same simulation experiment as above under the modified FCFS-T and report the results in Figure 9(b). Obviously, after the prescribed modification, the transfer ratio is decreased significantly. At the optimal configuration ($N_1 = 40, N_2 = 140$) the upper bound for the transfer ratio is around 4% and the lower bound is less than 2%. The convergence of the upper and lower
bounds indicates that the third transfer scenario becomes very infrequent with the modified policy. In addition, it is possible to decrease the upper bound to less than 2% by building a few more private rooms than the optimal solution. For example, if $N_1 = 60$ and $N_2 = 110$, the admission probability is decreased by 2% but both the lower and the upper bound for transfer ratio are at 2.1%. Such a result follows from the fact that because of excess private room capacity, infectious patients are always admitted to private rooms. Note that the latter may have additional benefits because it may be more risky to transfer infectious patients than non-infectious ones. When we repeat the same simulation experiments after setting the recovery rate 20% faster in private rooms than the original case, we obtain similar results and the original and the modified FCFS-T have virtually the same optimal revenues. Therefore we conclude that it is possible to operate with a mixed bed configuration without many patient transfers and reap the benefits of greater access.

I. Detailed simulation results

In this section we present the detailed simulation results for two parameter settings discussed in §6.2. We only present the results for $\lambda_0 = 100$ the results for the other cases are similar. In Table 7, we present the results for $\gamma_m = 0.7$, $p_{max} = 0.3$ and $p_{min} = 0.1$ and in Table 8 for $\gamma_m = 0.8$, $p_{max} = 0.3$ and $p_{min} = 0.2$. (All the other parameters are the same with those in §6.2.)

In each table we present the optimal configuration $N_1$ and $N_2$ for policies FCFS-T and REVMAX$_P$. The results for the case when all the rooms are private are also included for reference. In those cases $N_2 = 0$. We also present the optimal revenue (recall that it is also equal to the throughput in this case) and the admission probabilities for infectious ($P_1$) and non-infectious patients ($P_2$).

<table>
<thead>
<tr>
<th>Policy</th>
<th>$c_1$</th>
<th>$N_1$</th>
<th>$N_2$</th>
<th>Revenue</th>
<th>$P_1$</th>
<th>$P_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCFS-T</td>
<td>1.25</td>
<td>48</td>
<td>140</td>
<td>82.143</td>
<td>0.944</td>
<td>0.974</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>42</td>
<td>137</td>
<td>80.346</td>
<td>0.907</td>
<td>0.957</td>
</tr>
<tr>
<td></td>
<td>1.75</td>
<td>36</td>
<td>137</td>
<td>78.546</td>
<td>0.868</td>
<td>0.94</td>
</tr>
<tr>
<td>REVMAX$_P$</td>
<td>1.25</td>
<td>48</td>
<td>140</td>
<td>82.435</td>
<td>0.942</td>
<td>0.979</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>40</td>
<td>140</td>
<td>80.978</td>
<td>0.884</td>
<td>0.972</td>
</tr>
<tr>
<td></td>
<td>1.75</td>
<td>32</td>
<td>144</td>
<td>79.948</td>
<td>0.809</td>
<td>0.977</td>
</tr>
</tbody>
</table>

Table 7 Case 1: $\gamma_m = 0.7$, $p_{max} = 0.3$ and $p_{min} = 0.1$
Table 8  Case 2: $\gamma_m = 0.8$, $p_{max} = 0.3$ and $p_{min} = 0.2$

<table>
<thead>
<tr>
<th>Policy</th>
<th>$c_1$</th>
<th>$N_1$</th>
<th>$N_2$</th>
<th>Revenue</th>
<th>$P_1$</th>
<th>$P_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCFS-T</td>
<td>1.25</td>
<td>50</td>
<td>137</td>
<td>85.597</td>
<td>0.935</td>
<td>0.961</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>160</td>
<td>0</td>
<td>77.145</td>
<td>0.853</td>
<td>0.863</td>
</tr>
<tr>
<td></td>
<td>1.75</td>
<td>44</td>
<td>134</td>
<td>82.926</td>
<td>0.89</td>
<td>0.936</td>
</tr>
<tr>
<td></td>
<td></td>
<td>133</td>
<td>0</td>
<td>65.201</td>
<td>0.719</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td></td>
<td>39</td>
<td>131</td>
<td>79.937</td>
<td>0.839</td>
<td>0.909</td>
</tr>
<tr>
<td></td>
<td></td>
<td>114</td>
<td>0</td>
<td>56.097</td>
<td>0.618</td>
<td>0.628</td>
</tr>
<tr>
<td>REVMAX</td>
<td>1.25</td>
<td>48</td>
<td>140</td>
<td>86.12</td>
<td>0.926</td>
<td>0.971</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>160</td>
<td>0</td>
<td>77.447</td>
<td>1</td>
<td>0.818</td>
</tr>
<tr>
<td></td>
<td>1.75</td>
<td>40</td>
<td>140</td>
<td>83.968</td>
<td>0.849</td>
<td>0.966</td>
</tr>
<tr>
<td></td>
<td></td>
<td>133</td>
<td>0</td>
<td>65.584</td>
<td>1</td>
<td>0.641</td>
</tr>
<tr>
<td></td>
<td></td>
<td>36</td>
<td>137</td>
<td>81.725</td>
<td>0.788</td>
<td>0.952</td>
</tr>
<tr>
<td></td>
<td></td>
<td>114</td>
<td>0</td>
<td>56.525</td>
<td>1</td>
<td>0.506</td>
</tr>
</tbody>
</table>

J. Fluid model of FCFS-T systems

In this section we consider the fluid model of FCFS-T systems. Our goal is to show that the approximations used for the overflow of patients in §5 are asymptotically valid. Before we study the fluid model we first present the additional queueing equations under FCFS-T. Clearly equations (37)–(40), (44)–(50) hold under FCFS-T as well. Because there are no preemptions, $V_{ij}(t) = 0$, for $i,j = 1,2$ and $t \geq 0$. Note that under FCFS-T an arriving patient is admitted to a type 1 room if there is an available room, after possible transfers. Therefore,

$$a_{11}(t) = \int_0^t \left(1 \{Z_1(s^-) < N_1\} + 1 \{Z_1(s^-) = N_1\} 1 \{Z_2(s^-) < N_2\} 1 \{Z_2(s^-) > 0\}\right) dA_1(s).$$

(119)

Also, if all type 1 beds are occupied, a patient is routed to a type 2 room if there is sufficient number of available beds. Therefore,

$$a_{22}(t) = \int_0^t 1 \{Z_1(s) = N_1\} 1 \{Z_2(s^-) \leq N_2 - 1\} dA_2(s).$$

(120)

In §5 we provide approximations for the steady state of the system under FCFS-T when the arrival rates are fixed. To obtain these approximations we consider a sequence of systems indexed by $n$; the total number of beds in system. Let the arrival rates be given by

$$\lambda_i^n = \lambda_i n, \ i = 1,2$$

(121)

and assume that the number of beds of each type satisfies

$$N_i^n = n N_i, \ i = 1,2$$

(122)

in the $n$th system, so $N_1 + N_2 = 1$. The asymptotic regime described in §4 is very similar to the current regime except the fact that there the arrivals rates are assumed to be random.
We append \( n \) to the queueing processes; \( Z_{ij}^n, a_{ij}^n \) and \( T_{ij}^n \), \( i,j = 1,2 \), which still have the same interpretations explained previously, associated with the \( n \)th system in this sequence. Define

\[
Z_{ij}^n(t) = \frac{Z_{ij}^n(t)}{n}, \quad a_{ij}^n(t) = \frac{a_{ij}^n(t)}{n}, \quad T_{ij}^n(t) = \frac{T_{ij}^n(t)}{n}, \quad i,j = 1,2.
\]  
(123)

It can be shown using the queueing equations (37)–(40), (44)–(50), (119) and (120) as in Appendix B of Dai and Tezcan (2011) that the sequence \( (Z_{ij}^n, a_{ij}^n, T_{ij}^n, i,j = 1,2) \) is tight a.s. and every limit \( (\bar{Z}_{ij}, \bar{a}_{ij}, \bar{T}_{ij}, i,j = 1,2) \) satisfies the following fluid model equations in addition to (72), (74)–(77) and (79)–(82),

\[
\begin{align*}
\bar{Z}_{ij}(t) &= \bar{Z}_{ij}(0) + \bar{a}_{ij}(t) - \mu_{ij} \int_0^t \bar{Z}_{ij}(s)ds + \bar{T}_{ij}(t) - \bar{T}_{ij}(t), \quad i,j = 1,2,
\hat{a}_{11}(t) &= \lambda_1, \quad \text{if } \bar{Z}_1(t) < N_1 \text{ or } \bar{Z}_2(t) < N_2 \text{ and } \bar{Z}_{21} > 0,
\hat{a}_{22}(t) &= \lambda_2, \quad \text{if } \bar{Z}_{12} > 0 \text{ and } \bar{Z}_2(t) < N_2,
\end{align*}
\]  
(124)

Note also that \( \bar{V}_{ij}(t) = 0, i,j = 1,2 \), for all \( t \geq 0 \). We refer to (72), (74)–(77), (79)–(82) and (124)–(126) as the FCFS-T fluid model. Let \( \bar{X} = (\bar{Z}, \bar{T}, \bar{a}) \) denote a solution of the fluid model under FCFS-T, where \( \bar{Z} = (\bar{Z}_{ij}; i,j = 1,2) \), \( \bar{T} = (\bar{T}_{ij}; i,j = 1,2) \) and \( \bar{a} = (\bar{a}_{ij}; i,j = 1,2) \). It can be shown as in Dai and Tezcan (2011) that \( \bar{X} \) is continuous and differentiable a.e.

We also note that, if \( \bar{Z}_1(t) = N_1, \bar{Z}_2(t) = N_2 \) and \( \bar{Z}_{21} > 0 \), then type 1 patients are admitted only to a type 1 room at time \( t \), therefore (recall that arrivals are Poisson)

\[
\frac{\hat{a}_{11}(t)}{\lambda_1} = \frac{\hat{a}_{21}(t) + \hat{a}_{22}(t)}{\lambda_2},
\]  
(127)

If \( \bar{Z}_1(t) = N_1, \bar{Z}_2(t) = N_2 \) and \( \bar{Z}_{12} > 0 \), then new patients are admitted to type 2 rooms only. In addition, because type 1 patients require two type 2 beds to be admitted to type 2 rooms

\[
\frac{\hat{a}_{21}(t)}{\lambda_1} \leq \frac{\hat{a}_{22}(t)}{\lambda_2},
\]  
(128)

and \( \hat{a}_{11}(t) = 0 \).

The proof of the fact that fluid limits satisfy the fluid model equations is standard, hence is skipped. Equation (124) follows from (37), (125) and (126) follow from (119) and (120), respectively. Equation (72) follows from (48). Also (74) follows from (48), (49), (119), and (120), (75) follows from (121), (76) follows from (44) and (45), (77) follows from (45) and (50), (79) follows from the transfer procedures, (80) follows from (122) and (123), (81) follows from the definition and (82) follows from (38)–(40).

Fix the arrival rates \( \lambda = (\lambda_1, \lambda_2) \) and the capacity \( N = (N_1, N_2) \). Given a FCFS-T fluid model solution \( \bar{X} \), let \( \bar{z}_{ij} = \lim_{t \to \infty} \bar{Z}_{ij}(t), \quad i,j = 1,2 \), denote the steady state of this solution. Next we
establish certain properties of the steady state of the fluid model solutions. We use these results when we build approximations for the system performance under FCFS-T. Let \((s^*_i, i, j = 1, 2)\) be defined as in (13) and (14) (we omit \(\lambda\) and \(N\) from the notation for simplicity). Clearly \(s^*_i + s^*_j = \lambda_i\), for \(i = 1, 2\) when (16) holds.

**Lemma 7.** Let \(\bar{X}\) be a FCFS-T fluid model solution for fixed \(\lambda\) and \(N\).

(i) If (16) holds then \(\bar{z}_{ij} = s^*_ij/\mu_{ij}\), for \(i, j = 1, 2\).

(ii) If (16) does not hold, then \(\bar{z}_{ii} + \bar{z}_{ii'} = N_i\), for \(i = 1, 2\).

(iii) If (16) does not hold and \(\frac{N_i}{\rho_{11}} \geq \frac{N_2}{\rho_{22}}\), then \(\bar{z}_{12} = 0\).

(iv) If (16) does not hold and \(\frac{N_i}{\rho_{11}} > \frac{N_2}{\rho_{22}}\), then \(\bar{z}_{21} > 0\).

(v) If (16) does not hold and \(\frac{N_i}{\rho_{11}} = \frac{N_2}{\rho_{22}}\), then \(\bar{z}_{21} = 0\).

(vi) If (16) does not hold and \(\frac{N_i}{\rho_{11}} < \frac{N_2}{\rho_{22}}\), then \(\bar{z}_{21} = 0\) and \(\bar{z}_{12} > 0\).

**Remark 9 (Interpretation of Lemma 7).** The conditions for the different cases in §3.2 are obtained from the results of Lemma 7. The condition in (i) corresponds to the fact that the system has enough capacity. Note that in the fluid model all the patients are admitted by (i) in this case. Result (ii) implies that all the rooms are full when the system is overloaded. Results (iii)–(iv) imply that when (17) holds, type 2 patients overflow to type 1 rooms and type 1 patients are boarded in type 1 rooms only. Result (vi) implies that when (18) holds, type 1 patients overflow to type 2 rooms and type 2 patients are boarded in type 2 rooms only. Result (v) corresponds to the case when there is no overflow.

**Proof of Lemma 7**

Let \(\bar{X}\) be a FCFS-T fluid model solution satisfying (72), (74)–(77), (79)–(82) and (124)–(126) for fixed \(\lambda\) and \(N\). We start the proof of Lemma 7 with the following elementary result.

**Lemma 8.** The steady state quantities satisfy

\[
\mu_{11}\bar{z}_{11} + \mu_{12}\bar{z}_{12} \leq \lambda_1.
\]

**Proof of Lemma 8:** It is enough to show that for any given \(\epsilon > 0\) we have

\[
\mu_{11}\bar{Z}_{11}(t) + \mu_{12}\bar{Z}_{12}(t) \leq \lambda_1 + \epsilon
\]

(129)

for \(t\) large enough. To prove (129) assume that for a regular point \(t\)

\[
\mu_{11}\bar{Z}_{11}(t) + \mu_{12}\bar{Z}_{12}(t) > \lambda_1 + \epsilon
\]

(130)
and $\bar{Z}_{12}(t) > 0$. This implies that $\bar{Z}_{11}(t) = N_1$ by (76). In addition, by the fact that $\bar{Z}_{11}$ attains a maximum at a regular point $t$, this implies that $\bar{Z}_{11}(t) = 0$. Therefore, by (124) and (75)

$$\hat{\bar{Z}}_{12}(t) \leq \lambda_1 - \mu_{11}\bar{Z}_{11}(t) - \mu_{12}\bar{Z}_{12}(t).$$

This gives

$$\mu_{11}\hat{\bar{Z}}_{11}(t) + \mu_{12}\hat{\bar{Z}}_{12}(t) \leq \mu_{12}(\lambda - \mu_{11}\hat{\bar{Z}}_{11}(t) - \mu_{12}\hat{\bar{Z}}_{12}(t)) < -\mu_{12}\epsilon. \quad (131)$$

If (130) holds and $\bar{Z}_{12}(t) = 0$ we have $\hat{\bar{Z}}_{12}(t) = 0$ for any regular point $t > 0$. Therefore (131) holds in this case as well by (124) and (75), giving (129). □

Proof of Lemma 7 (i): Assume that (16) holds and consider the solution $s_{ij}^*$, for $i, j = 1, 2$, of (13) and (14). We next show that for any $\epsilon > 0$, we have

$$\mu_{11}\bar{Z}_{11}(t) + \mu_{12}\bar{Z}_{12}(t) \geq \lambda_1 - \epsilon \quad (132)$$

for $t$ large enough. Assume on the contrary that

$$\mu_{11}\bar{Z}_{11}(t) + \mu_{12}\bar{Z}_{12}(t) < \lambda_1 - \epsilon \quad (133)$$

for a regular point $t$. To conclude (132) it is enough to show that

$$\mu_{11}\hat{\bar{Z}}_{11}(t) + \mu_{12}\hat{\bar{Z}}_{12}(t) > c\epsilon \quad (134)$$

for some $c > 0$. If $\bar{Z}_1(t) < N_1$ or $\bar{Z}_2(t) < N_2$, the result is immediate from (124), (125) and (76), therefore assume that all the beds are full;

$$\bar{Z}_1(t) = N_1 \text{ and } \bar{Z}_2(t) = N_2. \quad (135)$$

For a regular point $t$ this gives

$$\hat{\bar{Z}}_1(t) = \hat{\bar{Z}}_2(t) = 0, \quad (136)$$

since $\bar{Z}_1$ and $\bar{Z}_2$ attain their maximum values at time $t$. Also, when $\bar{Z}_1(t) = N_1$ and $\bar{Z}_2(t) = N_2$, by (13), (14) and (76), (133) implies that

$$\mu_{21}\bar{Z}_{21}(t) + \mu_{22}\bar{Z}_{22}(t) > \lambda_2 + c\epsilon \quad (137)$$

for reselected $c > 0$. If

$$\bar{Z}_{21}(t) = 0, \quad (138)$$
then $\dot{Z}_{21}(t) = 0$ since $t$ is a regular point and so

$$\dot{Z}_{22}(t) < -\epsilon/\mu_{22},$$

(139)

by (137). Also (138) implies by (136) that $\dot{Z}_{11}(t) = N_1$ and so $\dot{Z}_{11}(t) = 0$. This with (139) and (136) gives $\dot{Z}_{12}(t) > \epsilon/\mu_{22}$. Hence, (134) holds when (138) holds. If

$$\dot{Z}_{21}(t) > 0,$$

(140)

then by (76) $\dot{Z}_{12}(t) = 0$ and so by (135) $\dot{Z}_{22}(t) = N_2$. Hence $\dot{Z}_{22}(t) = 0$ and so $\dot{Z}_{21}(t) < -\epsilon/\mu_{21}$ by (137) and $\dot{Z}_{12}(t) = 0$. Therefore, $\dot{Z}_{11}(t) > \epsilon/\mu_{21}$ by (136) and so (134) holds if (140) holds also.

By (76), (129) and (132) imply that for $t$ large enough $|Z_{1i}(t) - s_{1i}/\mu_{1i}| < \epsilon$, $i = 1, 2$. By similar arguments using (13), (14), (72), (74), (76) and (126) we have $|\dot{Z}_{2i}(t) - s_{2i}/\mu_{2i}| < \epsilon$, for $i = 1, 2$. Because $\epsilon$ is arbitrary, this completes the proof. \hfill \Box

Next we focus on the departure rate for the second patient type.

**Lemma 9.** The steady state quantities satisfy

$$\mu_{21}\bar{z}_{21} + \mu_{22}\bar{z}_{22} \leq \lambda_2.$$

**Proof of Lemma 9:** When (16) holds the result follows from part (i), hence we focus on the case when (16) does not hold. We show that for $\epsilon > 0$

$$\mu_{21}\bar{Z}_{21}(t) + \mu_{22}\bar{Z}_{22}(t) \leq \lambda_2 + \epsilon,$$

(141)

for $t$ large enough. We first show that if

$$\frac{N_1}{\rho_{11}} > \frac{N_2}{\rho_{22}}$$

(142)

then

$$\bar{Z}_{11}(t) \geq z'_{11} - \epsilon$$

(143)

for $t$ large enough and $\epsilon < z'_{11}$, where

$$z'_{11} = \frac{N_1 + \frac{\mu_{22}}{\mu_{21}}N_2}{\rho_{11} + \rho_{21}}\rho_{11}. $$

(144)

Let $P = z'_{11}/\rho_{11}$. It is easily checked that $z'_{11} < N_1$, $P < 1$, and $P\lambda_2 = \mu_{21}(N_1 - z'_{11}) + \mu_{22}N_2$ if (142) holds. Hence, (143) (with reselected $\epsilon$) gives (141) by (142) and Lemma 8.
Next we prove (143). Assume that \( \dot{Z}_{11}(t) < z'_{11} - \epsilon \) for a regular point \( t \). Then, if \( \dot{Z}_1(t) < N_1 \) or \( \dot{Z}_2(t) < N_2 \), it follows from (124), (125) and (76) that \( \dot{Z}_{11}(t) > c\epsilon \) for some constant \( c \) and any regular point \( t \). So assume that \( \dot{Z}_1(t) = N_1 \) and \( \dot{Z}_2(t) = N_2 \). Hence,

\[
\dot{Z}_1(t) = \dot{Z}_2(t) = 0. \tag{145}
\]

Because \( \dot{Z}_{11}(t) < z'_{11} - \epsilon \), we have that \( \dot{Z}_{21}(t) > 0 \). \tag{146}

(145) and (146) imply by (124) and (127) that

\[
\begin{align*}
\dot{Z}_{11}(t) &= p\lambda_1 - \mu_{11}\dot{Z}_{11}(t) \tag{147} \\
\dot{Z}_{21}(t) &= p\lambda_2 - \mu_{21}\dot{Z}_{21}(t) - \mu_{22}\dot{Z}_{22}(t) \tag{148}
\end{align*}
\]

for some \( 0 \leq p < 1 \) and

\[
\dot{Z}_{11}(t) = -\dot{Z}_{21}(t). \tag{149}
\]

Assume that \( \dot{Z}_{11}(t) \leq 0.5\epsilon(\mu_{11} \land \mu_{21}) \). Then by (147)

\[
p \leq \rho_{11}^{-1}\dot{Z}_{11}(t) + 0.5\rho_{11}^{-1}\epsilon < P
\]

and by (148)

\[
p \geq \rho_{21}^{-1}\dot{Z}_{21}(t) + \rho_{22}^{-1}N_2 - \lambda_2^{-1}\dot{Z}_{11}(t) \geq P,
\]

obviously a contradiction. Thus, \( \dot{Z}_{11}(t) \geq 0.5\epsilon(\mu_{11} \land \mu_{21}) \) giving (143).

We now show that if (142) does not hold, then

\[
\dot{Z}_{11}(t) = N_1. \tag{150}
\]

Let \( \delta = \rho_{22}^{-1}N_2 - \rho_{11}^{-1}N_1 \). Note that \( \delta > 0 \) if (142) does not hold. Assume that \( \dot{Z}_{11}(t) < N_1 \). If \( \dot{Z}_1(t) < N_1 \) or \( \dot{Z}_2(t) < N_2 \), it follows from (124), (125) and (76) that \( \dot{Z}_{11}(t) > c\epsilon \) for some constant \( c \). So assume that \( \dot{Z}_1(t) = N_1 \) and \( \dot{Z}_2(t) = N_2 \). Proceeding as above if \( \dot{Z}_{11}(t) < \delta (\lambda_1 \land \lambda_2) / 4 \) we have

\[
p \leq \rho_{11}^{-1}N_1 + \delta / 4
\]

and

\[
p > \rho_{22}^{-1}N_2 - \delta / 4,
\]

hence a contradiction by definition of \( \delta \), giving (150). It is now easy to show that (141) holds for \( i = 2 \) using (150) when (142) does not hold because (150) implies that \( \dot{Z}_{21}(t) = 0 \), and so \( \dot{Z}_{21}(t) = 0 \) and \( \dot{Z}_{22}(t) \leq \lambda_2 - \mu_{22}\dot{Z}_{22}(t) \) for any regular \( t \). \( \square \)
Proofs of Lemma 7 (ii)–(vi): Assume that (16) does not hold. Given $\epsilon > 0$, consider the following LP

\[
\begin{align*}
\min_{z_{ij} \geq 0, i, j = 1, 2} & \sum_{i=1}^{2} \lambda_i - \sum_{i,j=1}^{2} \mu_{ij} z_{ij} \\
\text{s.t.} & \quad z_{11} + z_{21} \leq N_1, \\
& \quad 2z_{12} + z_{22} \leq N_2, \\
& \quad \mu_{ii} z_{ii} + \mu_{ii} z_{ii}^t \leq \lambda_i + \epsilon, \text{ for } i = 1, 2, \\
& \quad z_{ij} \geq 0, \text{ for } i, j = 1, 2.
\end{align*}
\]

If (16) does not hold, by the continuity of an LP on the values of the constraints, for $\epsilon$ small enough the optimal objective function value $\theta^*(\epsilon)$ of (151) is greater than $\delta$ for some $\delta > 0$. Fix $\epsilon > 0$ such that $\theta^*(\epsilon) > \delta > 0$, for some $\delta > 0$.

Proof of (ii): Next we show that

\[
\bar{Z}_{ii}(t) + \bar{Z}_{ii'}(t) = N_i
\]

for $t$ large enough. Assume that (156) does not hold. Then, for $t$ large enough, $z_{ij} = \bar{Z}_{ij}(t), i, j = 1, 2$, satisfy the constraints (152)–(154) of the LP (151) by Lemmas 8 and 9. Hence for any regular $t$ by (74) and (124)

\[
\sum_{i=1}^{2} \sum_{j=1}^{2} \dot{\bar{Z}}_{ij}(t) = \lambda_1 + \lambda_2 - \sum_{i=1}^{2} \sum_{j=1}^{2} \mu_{ij} \bar{Z}_{ij}(t) \geq \delta, \quad i = 1, 2
\]

giving (156) for $t$ large enough.

Proof of (iii): Next we show that if (16) does not hold and (142) holds then $\dot{\bar{Z}}_{12}(t) \to 0$ as $t \to \infty$. Fix a large regular point $t$ so that (156) holds. If $\dot{\bar{Z}}_{12}(t) > 0$, then $\bar{Z}_{11}(t) = N_1$ by (76) and so $\dot{\bar{Z}}_{11}(t) = 0$. Because $\dot{\bar{Z}}_{2}(t) = N_2$ by part (ii), for all $t$ large,

\[
2\dot{\bar{Z}}_{12}(t) + \dot{\bar{Z}}_{22}(t) = 0.
\]

By (124), (77) and (76), if $\dot{\bar{Z}}_{12}(t) > 0$, then

\[
\dot{\bar{Z}}_{12}(t) = \dot{a}_{12}(t) - \mu_{12} \bar{Z}_{12}(t) - \mu_{11} N_1, \\
\dot{\bar{Z}}_{22}(t) = \dot{a}_{22}(t) - \mu_{22} \bar{Z}_{22}(t).
\]

Hence, if $\dot{\bar{Z}}_{12}(t) > 0$

\[
\frac{\dot{a}_{22}(t)}{\lambda_2} - \rho^{-1}_{22} \bar{Z}_{22}(t) \geq \frac{\dot{a}_{22}(t)}{\lambda_2} - \rho^{-1}_{22} (N_2 - \alpha) \geq \frac{\dot{a}_{12}(t)}{\lambda_1} - \rho^{-1}_{11} N_1,
\]

(158)
for a constant $c > 0$, where (a) follows from the fact that $\dot{Z}_{12}(t) > 0$ and (b) follows from (142) and (128). By (157)–(158), $\dot{\tilde{Z}}_{12}(t) < -ca$, for some $c > 0$, giving the desired result.

**Proofs of (iv)–(vi):** Next we show that if $\frac{N_1}{\rho_{11}} > \frac{N_2}{\rho_{22}}$, then $\tilde{z}_{21} > 0$. Let $z_{11}^*$ be defined as in (144). We now show that for any $\epsilon > 0$

$$\dot{Z}_{11}(t) \leq z_{11}^* + \epsilon$$

for $t$ large enough, which gives the desired result by part (ii). Assume that $\dot{Z}_{11}(t) > z_{11}^* + \epsilon$, for $z_{11}^* + \epsilon \leq N_1$ and a regular point $t$ large enough so that (156) holds. Then, by (147)–(149), using an argument similar to that succeeding (147)–(149), it can be shown that if $\dot{Z}_{11}(t) > -0.5\epsilon(\mu_{11} \wedge \mu_{21})$ then

$$p \geq \rho_{11}^{-1} z_{11}^* + 0.5 \rho_{11}^{-1} \epsilon > P$$

and

$$p < \rho_{21}^{-1} z_{21}^* + \rho_{22}^{-1} N_2 - 0.5 \rho_{21}^{-1} \epsilon < P,$$

giving the desired result. If $\frac{N_1}{\rho_{11}} = \frac{N_2}{\rho_{22}}$, the result follows from the fact that $z_{11}^* = N_1$. Part (vi) is proved similarly using (127). \qed