Balancing Admission Control, Speedup, and Waiting in Service Systems

Carri W. Chan
Columbia Business School, New York, NY cwchan@columbia.edu

Galit Yom-Tov
Israel Institute of Technology, Haifa, Israel galit@technion.edu

November 11, 2014

In a number of service settings, customer waiting, admission control, and speedup of service rates can occur during periods of congestion. For example, in a healthcare setting, this means that patients who require care may be sent to other, less ideal service outlets or hospital units. As expected, this comes at a cost to patient outcomes. In this work, we examine a multi-server queueing system which allows for admission control and speedup. We use dynamic programming to characterize properties of the optimal control and find that in some instances, the optimal policy has a simple form of a threshold policy. Leveraging this insight, we examine a queueing system where speedup is used when the number in the system exceeds some threshold and admission control is used when the number in the system exceeds some (potentially different) threshold. Using fluid analysis and a loss model, we establish approximations for the probability of speedup, the probability of admission control, and the expected queue length. We use the approximation analysis to characterize the region of the optimal solution and develop a greedy heuristic to derive a near optimal solution to the original optimization problem. We use simulation to demonstrate the quality of these approximations and find they can be quite accurate and robust. This analysis can provide insight to system administrators as they evaluate how to balance admission and speedup control—deciding when and to what extent to use each.

Key words: Queueing models, admission control, service rate control, dynamic programming, state-dependent queues, healthcare operations

1. Introduction

Providing high quality service is of paramount importance for many service systems. Unfortunately, when a system becomes congested, this is not always possible. A number of approaches have been considered and adopted to navigate these periods of congestion. For instance, admission control whereby customers are denied service (presumably finding service elsewhere or returning later) has arisen in hospital settings (Kim et al. 2014), call-centers (Ormeci 2004), and general service systems (Ata and Shneorson 2006). Alternatively, increasing service rate (with a sacrifice to quality) has also been considered in Intensive Care Units (ICUs) (Kc and Terwiesch 2012), production lines (Powell and Schultz 2004), email contact centers (Hasija et al. 2010), and general service systems (Ata and Shneorson 2006). In this work, we consider how to balance admission and service rate control in a multi-server setting in order to provide high quality service to as many customers as possible.

Our main motivating application is healthcare environments where congestion is pervasive. For example, many hospital ICUs have insufficient capacity to manage all of the demands of critical patients (Green 2003). These units are often congested and physicians have adopted a number of adaptive mechanisms—delays (Chalfin et al. 2007), admission control (Kim et al. 2014), speedup (Kc and Terwiesch 2012), ambulance diversion (Allon et al. 2013), etc.—to manage access to ICU care. Both speedup (Kc and Terwiesch 2012, Chan et al. 2012) and admission control (Kim et al. 2014, Shmueli et al. 2003) have been examined from both an empirical and analytic viewpoint. Patients who are sped-up may suffer from physiologic deterioration due to shorter intensive care. On the other hand, denying ICU admission to critical patients may also result in worse patient outcomes as well as loss of financial compensation to the hospital. Most of the work in the healthcare Operations Management and medical literature has examined each mechanism individually. To the best of our knowledge, this work is a first step to examine the pros and cons of joint speedup and admission control in the healthcare setting.

Related to our examination of speedup and its impact on service quality, Anand et al. (2010) examine the trade-off between quality and service rate in a queueing game framework. They consider a single server system which can modify its service rate and price. They find that the trade-off between quality and service speed are critical components of equilibrium prices, congestion, and service. In contrast, we examine a multi-server setting which includes admission control. We provide characterizations of the optimal policy, approximations of performance metrics of interest, and examine methods to use these approximations effectively to find a near optimal policy. Moreover, in a healthcare setting, patients can be relatively price-insensitive, as long as they have insurance coverage and the hospital in question is within their network. More often, quality of care is a stronger consideration. That said, hospitals operate under very real budget constraints. In this work, we provide approach that a) evaluates the performance of different admission and service strategies and b) finds a near optimal policy under appropriate (monetary and/or clinical) cost metrics.

A few works consider joint admission and service rate control (see for example, Adusumilli and Hasenbein (2010), Ata and Shneorson (2006), Lee and Kulkarni (2014)). Ata and Shneorson (2006) examine joint arrival and service rate control in an M/M/1 queue. They also consider how to set prices for service, when customers are price and delay sensitive. However, we consider a multi-server setting, provide structural properties of the optimal solution, and develop approximations for cost. We do not consider price-setting, as it is not a main driver in our healthcare application of interest. The properties we identify reveal policies which are both simple to implement as well as simple to estimate the impact on patient flows of the proposed policies. These characteristics are useful to help facilitate adoption of speedup and admission control policies by physicians and hospital administrators. Perhaps most closely related to our work is that of

Lee and Kulkarni (2014), which examines arrival and service rate control for a multi-server system. They also characterize properties of the optimal policy. However, we consider a slightly different cost setting (concave rather than convex cost functions) and are able to further characterize the optimal policy as having a threshold property. First, we find that it is optimal to only use the maximum or minimum arrival and service rates. Second, we leverage this fact and a monotonicity property to conclude the optimal policy can be defined by two threshold, N_s and N_a, such that if the number of customers in the system is larger than N_s (N_a) it is optimal to use service rate (admission) control. While we are able to identify settings in which threshold policies are optimal, studying such policies is of broader interest as they are simple to implement and are often used in practice (e.g. Allon et al. (2013)).

Via our analysis, we find that under the optimal policy, the admission and service rates depend on the amount of congestion in the system. Thus, we evaluate performance metrics of a system with these state-dependent dynamics. Bekker and Boxma (2007) and Bekker et al. (2008) consider the steady-state distribution of a single-server queue with workload-dependent service rates. In earlier work, Bekker et al. (2004) considers both arrival and service rates which depend on the workload in the system. Bekker and Borst (2006) considers admission control for a system with workload-dependent service rates. All of these works consider a single-server setting. Our work is quite different in that we consider a multi-server setting with both admission and service rate control; the dependence on workload is driven by properties of the optimal control policy, which we derive; and, we utilize fluid analysis and the methods of di Bernardo et al. (2008) and Filippov (1988) to provide approximations for the performance metrics of interest: the probability of speedup and admission control as well as the expected queue length under our control policy. Similar to Chan et al. (2014), we utilize fluid models with discontinuous differential equations. However, here we consider a system with admission control and speedup, but without feedback. In Section 6, we consider extensions to include customer returns.

We are motivated by the following questions: Under what conditions is speedup beneficial? When should patients' service rate should be accelerated? Similarly, when should admission control be used to manage patient demand? What is the trade-off between ensuring quality care for admitted patients versus providing access to care for incoming patients?

In examining joint admission and service control, we introduce a queueing model which extends prior work. In particular, we consider a system with multiple servers, a modified policy space to capture constraints in a healthcare setting—such as requiring non-zero arrival rates—and a combination of optimization and performance evaluation to provide more insight into the management of a service system with adjustable arrival and service rates. Arrival and service rates can be adjusted dynamically over a closed-set of possible

rates. Increasing service rate comes at a cost, while decreasing arrival rate comes as a cost. In addition, costs are incurred for each customer who enters the system and has to wait—longer waits result in larger costs.

We start with a stochastic optimization framework and characterize properties of the optimal policy. Some of these properties are similar to those established in Ata and Shneorson (2006) for a single-server system and Lee and Kulkarni (2014) for a multi-server system. We further identify properties of the optimal policy under characterizations of the system's cost functions, which was not considered in these prior works. Specifically, we are able to demonstrate the optimality of a threshold policy. We find that the optimal thresholds can be highly dependent on system dynamics—which are likely possible to estimate from empirical data—and cost functions—which may be possible to coarsely estimate, but difficult to compare across the different sources of costs, i.e. admission rate versus service rate versus queue length. Due to the potential challenges of precisely quantifying the relationship across these different costs, we examine performance evaluation under the restriction of operating the queueing system under a threshold policy. In doing so, system administrators can assess the impact of their control decisions, while being assured that the policy they are considering lies within the space of optimal policies. Taking this a step further, we propose a heuristic algorithm to determine thresholds for admission control and speedup and demonstrate via simulation that this heuristic can have very good performance. In this work, we make the following key contributions:

- We derive properties of the optimal control of a multi-server queueing model with joint admission and service rate control. By considering concave cost functions, we demonstrate the optimality of a policy which only utilizes the maximum and minimum service and arrival rates. Thus, we are able to characterize the optimal policy via a simple policy which is defined by two thresholds, N_a and N_s.
- We leverage our results from the stochastic optimization to characterize the impact of the thresholds on performance metrics of interest. In particular, we use fluid analysis to provide approximations for the probability of speedup and admission control.
- Because the fluid model provides little insight into the behavior of the queue length of our system, we develop an approximation for that measure via a loss model whose parameters are based on the equilibrium analysis of the fluid model.
- Our approximations allow for performance analysis given speedup and admission control thresholds. Moreover, they also enable one to solve a constraint satisfaction problem. In particular, a hospital manager can utilize our performance approximations to determine threshold values which will satisfy constraints in the amount of admission control and/or speedup used as well as the expected queue length.
- We then utilize our performance measure approximations to solve a cost minimization problem, where costs are appropriately defined based on clinical and financial considerations. We find a set of solutions which appear to be 'zero cost'. This set of solutions suggests that–for all system parameters–admission

control and/or speedup should begin before a queue builds. We find that it is important to carefully select thresholds, N_a and N_s, within this set and develop a greedy heuristic to do so that prioritizes admission control and speedup based on relative costs. Using simulation, we compare the performance of our heuristic to the optimal solution found via exhaustive search. We find that our heuristic is quite accurate and provides a near optimal solution to our problem. Moreover, it is quite robust to misspecifications in system cost parameters.

The rest of the paper proceeds as follows. We introduce our stochastic queueing model in Section 2. We then characterize properties of the optimal arrival and service rates in Section 3. In Section 4, we introduce a fluid model and use it to develop approximations for performance metrics of interest. In Section 4.3, we utilize these approximations to establish a heuristic solution to our original optimization problem. We do numerical analysis to examine the quality of the approximations and the heuristic solution in Section 5. In Section 6, we consider an extension of our model which incorporates customer returns to service. Finally, Section 7 provides some concluding remarks and discussion.

2. Model

We now formally introduce our queueing model. This stochastic model captures the possibility of admission and speedup control. We consider a queueing system as depicted in Figure 1. The station models a Medical Unit (MU) with N beds, such as an Intensive-Care Unit (ICU), where patients are treated. The queue at Station 1 captures the time a patient spends waiting when an admission request was made until the patient is finally admitted to the ICU. Let $\mathbb{Q}$ denote the stochastic number of patients in our system.

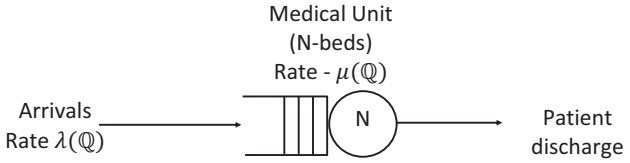


Figure 1 **ICU model: N-server system with arrival rate, $\lambda(\mathbb{Q})$, and service rate, $\mu(\mathbb{Q})$, which can depend on the number of customers in the system, $\mathbb{Q}$.**

When the ICU becomes congested, the hospital can utilize *admission control*, for instance by utilizing ambulance diversion or surgical cancellations, and *patient speedup*, by discharging patients from the ICU at a faster rate. Neither of these options is desirable. On one hand, admission control results in a loss of revenue and potentially degraded care for the patient who is treated in a non-ICU (or less medically desirable) unit. On the other hand, speedup impacts quality of patient care and increases the risk of physiologic deterioration when discharging the patient earlier than would normally occur. Our goal is to understand how and when each mechanism should be utilized.

Remark 1 *While our model presentation focuses on the ICU setting, we note that such dynamics can occur when considering not only any other MU but also the entire hospital, with N capturing the number of hospital beds. Moreover, other service settings may also demonstrate such dynamics. However, to streamline the discussion, we will focus on the ICU setting.*

Similar to Lee and Kulkarni (2014), we consider a continuous time infinite horizon, discounted cost formulation. The arrival rate of critical patients is dependent on the selection of $\lambda \in [\lambda_L, \lambda_H]$. Hence, we consider the situation where admission control is possible. The nominal arrival rate is λ_H; if admission control is in place, the arrival rate is reduced. Note that even with admission control, which can be achieved via ambulance diversion and rerouting patients to other units, the arrival rate is likely to be non-zero as there may be walk-ins or very severe patients who cannot be rerouted. If admission control is employed, a cost rate of $\phi(\lambda)$, which is non-increasing in λ, is incurred. This cost can capture the clinical cost (e.g. the increased mortality risk or readmission load) of denied service. Note that while the patient is 'denied service' in our queueing model, in practice, this patient will be treated in another unit or at another hospital. Thus, this cost can also incorporate any financial losses due to not treating another patient.

Patient service is completed at the nominal service rate μ_L. The ICU can employ speedup which increases the service rate to $\mu \in [\mu_L, \mu_H]$. When speedup is utilized, a cost rate of $\xi(\mu)$, which is non-decreasing in μ, is incurred. This captures the undesirability of speedup. For instance, it can account for the increased mortality risk or readmission load due to speedup (see Chan et al. (2012) for a discussion of different clinically relevant cost functions). Note that, similar to Chan et al. (2012), we do not explicitly model patients being readmitted and this cost ξ serves to capture this phenomenon. We examine this connection more extensively in Section 6.

Finally, if there are $\mathbb{Q}$ critical patients in the system, a cost rate of $h(\mathbb{Q})$, which is non-decreasing in $\mathbb{Q}$, is incurred. Without loss of generality, let $h(0) = 0$. This cost function can capture the clinical cost of waiting in various ways. For example, if a waiting cost c_w is incurred for each patient who is waiting to be treated in the ICU, then $h(\mathbb{Q}) = c_w(\mathbb{Q} - N)^+$. Similarly, if there is simply a cost for having a queue, $h(\mathbb{Q}) = c1_{\{\mathbb{Q}>N\}}$. Our goal is to minimize the expected discounted cost incurred over an infinite horizon. Let $\mathbb{Q}(t)$ be the state at time t, i.e. the number of patients in the system. Our goal is to determine policy $u(t)$—which may depend on $\mathbb{Q}(t)$—such that:

$$\lim_{i\to\infty} E\left[\int_0^{t_i} e^{-\beta t} g(\mathbb{Q}(t), u(t))dt\right] \tag{1}$$

is minimized, where the cost rate is given as:

$$g(\mathbb{Q}, u) = h(\mathbb{Q}) + \phi(\lambda(u)) + \xi(\mu(u)).$$

Throughout our analysis we assume that there are enough servers to satisfy all demand, irrespective of what control is employed. Thus, our control is about ensuring service quality, rather than stability.

Assumption 1 *We make the following assumption about the number of servers in the system:*

$$N > \lambda_H/\mu_L.$$

3. Characterizing the Optimal Policy

We now turn our attention to characterizing the optimal policy which minimizes the average cost, given in (1). Some of these results are similar to those derived in Ata and Shneorson (2006) and Lee and Kulkarni (2014), which we include here for completeness. However, we also establish new properties of the optimal control, which are not included in the prior works. These properties, which emit a simple, easily implementable policy, are vital for the performance evaluation and optimization discussed in Section 4.

Using the uniformization technique (Bertsekas 2001), we transform our continuous time problem into a discrete time equivalent model. In particular, we can see that for any action $u = (\lambda, \mu)$, the rate to the next state transition in state $\mathbb{Q}$ is given as:

$$v_{\mathbb{Q}}(u) = \begin{cases} \lambda(u), & \mathbb{Q} = 0; \\ \lambda(u) + (\mathbb{Q} \wedge N)\mu(u), & \mathbb{Q} \geq 1. \end{cases}$$

Hence, the maximum possible transition rate is $v = \lambda_H + N\mu_H$. We can write the Bellman equation for this optimization problem. The minimum discounted cost-to-go is:

$$\begin{aligned} J(0) &= \frac{1}{\beta + v} \min_{\lambda \in [\lambda_L, \lambda_H]} \{\phi(\lambda) + (v - \lambda)J(0) + \lambda J(1)\} \\ J(\mathbb{Q}) &= \frac{1}{\beta + v} \min_{\lambda \in [\lambda_L, \lambda_H], \mu \in [\mu_L, \mu_H]} \{h(\mathbb{Q}) + \phi(\lambda) + \xi(\mu) + \\ &\qquad \lambda J(\mathbb{Q}+1) + (\mathbb{Q} \wedge N)\mu J(\mathbb{Q}-1) + (v - \lambda - (\mathbb{Q} \wedge N)\mu)J(\mathbb{Q})\}. \end{aligned}$$

We define the following differential of the optimal discounted cost:

$$\Delta(\mathbb{Q}) = J(\mathbb{Q}) - J(\mathbb{Q}-1)$$

where by convention we define $\Delta(0) = 0$. Hence, the Bellman's equation can be rewritten as:

$$J(\mathbb{Q}) = \frac{1}{\beta + v}\left[h(\mathbb{Q}) + vJ(\mathbb{Q}) + \min_{\lambda}\{\phi(\lambda) + \lambda\Delta(\mathbb{Q}+1)\} + \min_{\mu}\{\xi(\mu) - (\mathbb{Q} \wedge N)\mu\Delta(\mathbb{Q})\}\right].$$

The optimal policy is then

$$u^*(\mathbb{Q}) = (\lambda^*(\mathbb{Q}), \mu^*(\mathbb{Q})) = \arg\min_{\lambda,\mu}\{\phi(\lambda) + \lambda\Delta(\mathbb{Q}+1) + \xi(\mu) - (\mathbb{Q} \wedge N)\mu\Delta(\mathbb{Q})\}.$$

Our goal is to understand properties of the optimal policy. In particular, we will show that the optimal policy is monotonic in the number of patients in the system. That is, the optimal service rate $\mu^*(\mathbb{Q})$ is increasing in $\mathbb{Q}$ and the optimal arrival rate $\lambda^*(\mathbb{Q})$ is decreasing in $\mathbb{Q}$. This result is similar to that in Lee and Kulkarni (2014). The proof is provided in the Appendix for completeness.

Theorem 1 *The optimal policy is monotonic in $\mathbb{Q}$. That is, if it is optimal to use speedup (admission control) in state $\mathbb{Q}$, it is also optimal to use speedup (admission control) in state $\mathbb{Q}+1$. We have the following two results:*

1. *The optimal service rate, $\mu^*(\mathbb{Q})$, is non-decreasing in $\mathbb{Q}$.*
2. *The optimal admission rate, $\lambda^*(\mathbb{Q})$, is non-increasing in $\mathbb{Q}$.*

We now consider a special case of the cost functions $\phi(\lambda)$ and $\xi(\mu)$. In this case, we can further characterize the optimal policy as having binary notions of speedup and admission control. Note that this characterization of the cost functions was not considered in Ata and Shneorson (2006) or Lee and Kulkarni (2014); thus, the corresponding results are new.

Assumption 2 *We make the following concavity assumptions about our cost functions.*

1. *The cost function $\phi(\lambda) \geq 0$ is concave and non-increasing in λ.*
2. *The cost function $\xi(\mu) \geq 0$ is concave and non-decreasing in μ.*

We first consider the arrival rate cost function, $\phi(\lambda)$. One could consider a linear function ϕ which would capture the clinical (or financial) cost associated with each denied admission. Generalizing to a concave cost function would imply that the differential cost of reducing the arrival rate is highest when starting to use admission control. This may hold when considering financial or operational costs. Reducing the arrival rate can be done in a number of way; for instance, via ambulance diversion or canceling surgeries. If one considers there is administrative overhead to start canceling surgeries, it may be reasonable to assume that once the initial set up cost is incurred, further cancellations come at a lower cost.

Similar to before, a linear function for the service rate cost function, $\xi(\mu)$, is also reasonable if ξ captures the clinical cost of a patient being 'discharged early'. In so far as the service rate dictates the expected service time, the concavity assumption for the service rate cost function implies that the relative change in service time is a stronger indicator of costs than the absolute change. For instance, staying 1 less day for a patient who is expected to stay for 10 days is much less traumatic than for a patient who is expected to stay for 2 days.

Under these assumptions, we can establish the following property of the optimal policy, which makes it highly desirable for implementation.

Theorem 2 *Given Assumptions 1 and 2, the optimal admission control and speedup policy will only use the maximum and minimum arrival and service rates. That is:*

1. $\lambda^*(\mathbb{Q}) \in \{\lambda_L, \lambda_H\}$.
2. $\mu^*(\mathbb{Q}) \in \{\mu_L, \mu_H\}$.

Any $\mu \in (\mu_L, \mu_H)$ or $\lambda \in (\lambda_L, \lambda_H)$ is sub-optimal.

The proof is provided in the Appendix. Of course linear cost functions also satisfy Assumption 2; thus, Theorem 2 also holds for functions of this form. Theorem 2 implies that the optimal policy can be defined by two parameters, N_a and N_s, which represent thresholds at which to begin admission control and speedup, respectively. That is, the optimal policy is such that:

- **Admission Control:** $\lambda^*(\mathbb{Q}) = \begin{cases} \lambda_L, \text{ if } \mathbb{Q} < N_a; \\ \lambda_H, \text{ if } \mathbb{Q} \geq N_a. \end{cases}$
- **Speedup Control:** $\mu^*(\mathbb{Q}) = \begin{cases} \mu_L, \text{ if } \mathbb{Q} < N_s; \\ \mu_H, \text{ if } \mathbb{Q} \geq N_s. \end{cases}$

Remark 2 *We have identified conditions under which threshold policies are optimal. However, understanding the behavior of threshold policies is of broader interest as there is evidence that such policies are often used in practice. For instance, hospitals often go on ambulance diversion (altering the arrival rate between λ_H and λ_L) once the number of patients waiting exceeds some predefined threshold. Additionally, speedup in the ICU has been shown to take place once the number of available beds goes below some value (Kc and Terwiesch 2012).*

3.1. Average Cost Problem

Thus far, we have considered the infinite-horizon, discounted cost problem. It turns out that our structural results for the discounted cost problem also extend to the average cost problem. In this case, the objective is to minimize the average cost per-stage:

$$\lim_{i \to \infty} \frac{1}{E[t_i]} E\left[\int_0^{t_i} g(\mathbb{Q}(t), u(t)) dt\right],$$

where the cost rate is the same as before.

Proposition 1 *Given Assumptions 1 and 2, the optimal admission control and speedup policy which minimizes the average cost will only use the maximum and minimum arrival and service rates. That is:*

1. $\lambda^*(\mathbb{Q}) \in \{\lambda_L, \lambda_H\}$
2. $\mu^*(\mathbb{Q}) \in \{\mu_L, \mu_H\}$

Any $\mu \in (\mu_L, \mu_H)$ or $\lambda \in (\lambda_L, \lambda_H)$ is suboptimal.

The proof is provided in the Appendix.

4. Performance Evaluation and Cost Minimization: Fluid Analysis

Now that we have characterized the optimal policy, a natural next step is to determine the thresholds, N_a and N_s, which specify when admission control and speedup should be used. As expected, these thresholds are highly dependent on system parameters ($\lambda_L, \lambda_H, \mu_L, \mu_H, N$) as well as the cost functions (ϕ, ξ, h).

Since there exist methodologies to estimate many of these parameters and cost functions (see, for instance, Kim et al. (2014), Kc and Terwiesch (2012), Chan et al. (2013)), in this work, we assume that they are given. In addition, since we know the optimal policy is of threshold type, and in light of Remark 2, we restrict our analysis to policies of this form. We then use various approximations to examine the effect of the thresholds on the performance metrics of interest: the expected queue length, $E[\mathbb{Q}]$, the probability of speedup, $P(\mathbb{Q} \geq N_s)$, and the probability of admission control, $P(\mathbb{Q} \geq N_a)$. Not only does this provide performance evaluation approximations, but also optimizing over these approximations will provide thresholds which approximately minimize the system operating costs. We will start with deriving these approximations here and then use simulation to examine the quality of the approximations in Section 5.

We now have a state-dependent queueing system, which can be quite cumbersome to analyze. Thus, we start by considering a fluid approximation for our system. We denote the fluid function of our queueing network by $Q = \{Q(t), t \geq 0\}$. Here $Q(t)$ is the fluid content of patients in the system at time t. We derive the fluid formula directly. We assume that arrivals and departures occur deterministically at the specified rates and also regard the number of patients and beds as continuous quantities. Thus, the fluid arrives deterministically and continuously at a state dependent rate $\lambda(Q)$. Fluid is served in the ICU deterministically at rate $\mu(Q)(Q \wedge N)$, where $(Q \wedge N)$ is the number of occupied beds in the ICU. The arrival rate function $(\lambda(\cdot))$ and the service rate function $(\mu(\cdot))$ are discontinuous. These functions are given by (2) and (3), respectively

$$\lambda(Q) = \begin{cases} \lambda_H \ , \text{ if } Q < N_a, \\ \lambda_L \ , \text{ if } Q \geq N_a, \end{cases} \tag{2}$$

and

$$\mu(Q) = \begin{cases} \mu_L \ , \text{ if } Q < N_s, \\ \mu_H \ , \text{ if } Q \geq N_s. \end{cases} \tag{3}$$

The dynamics of our model can be captured by the following Ordinary Differential Equations (ODE) with discontinuous Right Hand Side (RHS):

$$\dot{Q}(t) = 1_{\{Q(t)<N_a\}}\lambda_H + 1_{\{Q(t)\geq N_a\}}\lambda_L - 1_{\{Q(t)<N_s\}}\mu_L(Q(t) \wedge N) - 1_{\{Q(t)\geq N_s\}}\mu_H(Q(t) \wedge N). \tag{4}$$

This discontinuous ODE is discontinuous in Q, but continuous in t. From (4), it is easy to see that the derivative values, $\dot{Q}$, which specify the flow dynamics are discontinuous at $Q(t) = N_a$ and $Q(t) = N_s$. We will analyze the long-term behavior of this fluid system, i.e. the behavior as $t \to \infty$. Let $\bar{q}$ be the steady-state value such that:

$$\lim_{t\to\infty} Q(t) = \bar{q}$$

In theory, this limit may be infinite and/or may not be unique (e.g. it may depend on initial conditions). As we will see later, the limit is finite and unique under Assumption 1.

We begin by defining the following parameters, which will be useful in describing the system dynamics:

$$\begin{aligned} q^{LL} &= \frac{\lambda_L}{\mu_L}, \quad q^{HL} = \frac{\lambda_H}{\mu_L}, \\ q^{LH} &= \frac{\lambda_L}{\mu_H}, \quad q^{HH} = \frac{\lambda_H}{\mu_H}. \end{aligned} \tag{5}$$

One can think of these parameters as the offered-load at the ICU under different arrival and service rate dynamics, i.e. when admission and/or speedup control is always/never used. Note that by assumption, the following relationship holds:

$$q^{LH} < q^{LL}, q^{HH} < q^{HL}.$$

We start by analyzing the long-term behavior of the fluid model presented in Equation (4). The main challenge is the discontinuities at $Q = N_a$ and $Q = N_s$. As expected, the long-term behavior is highly dependent on system parameters for arrival and service times, as well as the control variable for when to begin speedup and admission control, (N_a, N_s). Speedup and admission control may begin before a queue forms, if the thresholds are less than N, or after, if they are greater than N. The proof of this result can be found in the Appendix and utilizes Lyapunov techniques under the Filippov (1988) and di Bernardo et al. (2008) approach for differential equations with discontinuous RHS. This approach uses a smoothing technique for the ODE around the points of discontinuity, which results in a probabilistic version of the fluid model.

Theorem 3 *Under Assumption 1, the long-term behavior of the fluid queueing system in* (4) *is broken into the following cases:*

1. ***Case 1—Admission Control First (ACF) (*** $N_a < N_s$***):***
   - *1.1* q^{HL} *is a globally stable equilibrium if* $q^{HL} \leq N_a$.
   - *1.2* N_a *is a globally stable equilibrium if* $q^{LL} \leq N_a \leq q^{HL}$.
   - *1.3* q^{LL} *is a globally stable equilibrium if* $N_a \leq q^{LL} \leq N_s$.
   - *1.4* N_s *is a globally stable equilibrium if* $q^{LH} \leq N_s \leq q^{LL}$.
   - *1.5* q^{LH} *is a globally stable equilibrium if* $N_s \leq q^{LH}$.
2. ***Case 2—Speedup Control First (SCF) (*** $N_s < N_a$***):***
   - *2.1* q^{HL} *is a globally stable equilibrium if* $q^{HL} \leq N_s$.
   - *2.2* N_s *is a globally stable equilibrium if* $q^{HH} \leq N_s \leq q^{HL}$.
   - *2.3* q^{HH} *is a globally stable equilibrium if* $N_s \leq q^{HH} \leq N_a$.
   - *2.4* N_a *is a globally stable equilibrium if* $q^{LH} \leq N_a \leq q^{HH}$.
   - *2.5* q^{LH} *is a globally stable equilibrium if* $N_a \leq q^{LH}$.

3. ***Case 3—Simultaneous Admission and Speedup Control (SASC) ($N_s = N_a$):***

   3.1 q^{HL} *is a globally stable equilibrium if* $q^{HL} \leq N_s = N_a$.

   3.2 $N_a = N_s$ *is a globally stable equilibrium if* $q^{LH} \leq N_a = N_s \leq q^{HL}$.

   3.3 q^{LH} *is a globally stable equilibrium if* $N_s = N_a \leq q^{LH}$.

Figure 2 summarizes the equilibria of Theorem 3, demonstrating its behavior as a function of the thresholds N_a and N_s.

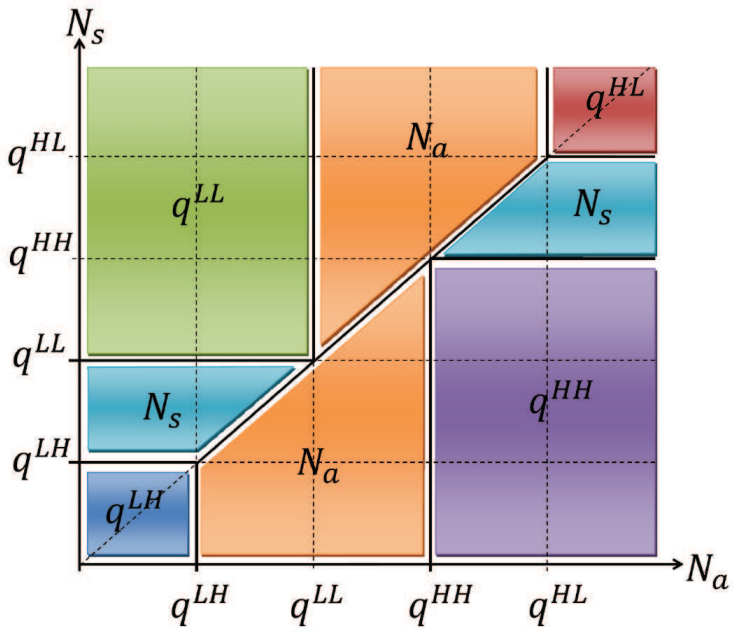


Figure 2 Equilibrium for various admission control and speedup threshold values, N_a and N_s.

4.1. Admission Control and Speedup Approximations

While the equilibrium values of the fluid model are interesting in their own right and provide some insight into the behavior of the stochastic model, this does not yet provide insight into the performance metrics of interest or, ultimately, the cost function we are interested in minimizing. Fortunately, as a byproduct of our fluid analysis via Filippov (1988) techniques, we can derive approximations for the probability of speedup and admission control of the original stochastic model.

As a corollary to Theorem 3, we establish the proportion of time the fluid content is above the speedup threshold, and hence, using speedup:

$$\lim_{T\to\infty} \frac{1}{T} \int_0^T 1_{\{Q(t)\geq N_s\}} dt \tag{6}$$

and admission control threshold, and hence, using admission control:

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T 1_{\{Q(t)\geq N_a\}}dt \tag{7}$$

We formally provide the statement for the ACF case ($N_a < N_s$) and note that the other two cases follow similarly and will be summarized later in Table 1.

Corollary 1 *Under Assumption 1 and ACF case ($N_a < N_s$), the proportion of time the fluid process is above the admission control threshold is given by:*

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T 1_{\{Q(t)\geq N_a\}}dt = \begin{cases} 0, & q^{HL}\leq N_a; \\ \frac{\lambda_H-\mu_L(N\wedge N_a)}{\lambda_H-\lambda_L}, & q^{LL}\leq N_s\leq q^{HL}; \\ 1, & N_a\leq q^{LL}\leq N_s; \\ 1, & N_s\leq q^{LH}; \\ 1, & q^{LH}\leq N_s\leq q^{LL}. \end{cases} \tag{8}$$

Similarly, the proportion of time the fluid process is above the speedup control threshold is given by:

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T 1_{\{Q(t)\geq N_s\}}dt = \begin{cases} 0, & q^{HL}\leq N_a; \\ 0, & q^{LL}\leq N_s\leq q^{HL}; \\ 0, & N_a\leq q^{LL}\leq N_s; \\ \frac{\lambda_L-\mu_L(N\wedge N_s)}{(\mu_H-\mu_L)(N\wedge N_s)}, & q^{LH}\leq N_s\leq q^{LL}. \\ 1, & N_s\leq q^{LH}. \end{cases} \tag{9}$$

Consequently, these values can provide approximations for the probability that speedup and admission control are used in our original stochastic system from Section 2. In particular, we approximate the following probabilities of our original stochastic model as:

$$P(Speedup) = P(\mathbb{Q}\geq N_s)\approx\lim_{T\to\infty}\frac{1}{T}\int_0^T 1_{\{Q(t)\geq N_s\}}dt \tag{10}$$

$$P(Admission\ Control) = P(\mathbb{Q}\geq N_a)\approx\lim_{T\to\infty}\frac{1}{T}\int_0^T 1_{\{Q(t)\geq N_a\}}dt \tag{11}$$

Table 1 summarizes the approximations for the performance metrics P(Admission Control) and P(Speedup) in each subcase. In Section 5, we will use simulation to examine the accuracy of these approximations and see that they can be quite accurate. Recall that our original optimization model considered the use of admission control, speedup control, and waiting times for new patients. We have just established closed-form expressions to approximate the probability admission and speedup control will be used. What remains is to understand how the thresholds, N_a and N_s, impact the queue length.

4.2. Queue Length Approximation

In our fluid model, the queue length $(Q-N)^+$ is always 0. As such, fluid analysis does not appear to provide much insight into the behavior of the queue. So, we take a different approach, which accounts for the stochasticity in the queue length process, but also utilizes the results of our fluid analysis.

Case		P(Admission Control)	P(Speedup)
(1) ACF	1.1	0	0
	1.2	$\frac{\lambda_H-\mu_L(N\wedge N_a)}{\lambda_H-\lambda_L}$	0
	1.3	1	0
	1.4	1	$\frac{\lambda_L-\mu_L(N\wedge N_s)}{(\mu_H-\mu_L)(N\wedge N_s)}$
	1.5	1	1
(2) SCF	2.1	0	0
	2.2	0	$\frac{\lambda_H-\mu_L(N\wedge N_s)}{(\mu_H-\mu_L)(N\wedge N_s)}$
	2.3	0	1
	2.4	$\frac{\lambda_H-\mu_H(N\wedge N_a)}{\lambda_H-\lambda_L}$	1
	2.5	1	1
(3) SASC	3.1	0	0
	3.2	$\frac{\lambda_H-\mu_L(N\wedge N_a)}{\lambda_H-\lambda_L-(\mu_L-\mu_H)(N\wedge N_a)}$	$\frac{\lambda_H-\mu_L(N\wedge N_a)}{\lambda_H-\lambda_L-(\mu_L-\mu_H)(N\wedge N_a)}$
	3.3	1	1

Table 1 Performance level approximations for the probability of speedup and admission control in each subcase. The approximations come from the derived proportion of time the fluid content is about the speedup and/or admission control thresholds.

We start by considering the extreme cases where the fluid analysis suggests that speedup and/or admission control is always or never used. In such a scenario, it is conceivable that very limited information is lost by ignoring the change in dynamics due to the thresholds. As an example, consider the ACF case where $q^{HL} \leq N_a$ (case 1.1). In this case, the fluid analysis suggests that neither speedup or admission control is ever used. If this were truly the case, the system would evolve as an M/M/N queue with arrival rate $\hat{\lambda} = \lambda_H$ and service rate $\hat{\mu} = \mu_L$. Using standard approaches, we can then get an approximation for the queue length given by the analysis of this M/M/N queue. Via a similar argument, we could do the same in case 1.3 with $\hat{\lambda} = \lambda_H$ and $\hat{\mu} = \mu_H$.

When the equilibrium is on a control threshold (either N_a or N_s), it is certain that the dynamics are changing due to the threshold. In fact, they are changing very rapidly, so that the fluid content remains on the threshold boundary. In these cases (1.2, 1.4, 2.2, 2.4, and 3.2), we still use an M/M/N queue to approximate the dynamics; however, the state *independent* arrival and departure rates will be given by the average arrival and departure rates as approximated by the fluid analysis. Hence, for all cases we will consider the following arrival and departure rates: $\hat{\lambda} = \text{P(Admission Control)} \times \lambda_L + (1 - \text{P(Admission Control)}) \times \lambda_H$ and $\hat{\mu} = \text{P(Speedup)} \times \mu_H + (1 - \text{P(Speedup)}) \times \mu_L$, where P(Admission Control) and P(Speedup) are given by our fluid analysis as summarized in Table 1.

Using the analysis of an M/M/N queue, with arrival rate $\hat{\lambda}$ and departure rate $\hat{\mu}$, can provide very reasonable estimates of the simulated queue length of our state-dependent queueing system when when N_a and N_s are smaller than N. However, when both of the thresholds are larger than N, we find that the approximation via the M/M/N queue overestimates the queue length. That is because the change in arrival and service

rates "push" the derivative ($\dot{Q}$) down and decrease the queue length. To capture this strong push, we instead use an M/M/N/K queue (a loss model), where $K = \min\{N_a, N_s\}$, when $\min\{N_a, N_s\} \geq N$ and $K = N$, otherwise. The loss model queue length is derived by solving the local balance equations:

$$\begin{aligned}
\pi_i &= \frac{1}{i!}\left(\frac{\hat{\mu}}{\hat{\lambda}}\right)^{-i}\pi_0 \quad 0 \leq i \leq N \\
\pi_i &= \frac{N^N}{N!}\left(\frac{\hat{\lambda}}{N\hat{\mu}}\right)^{i}\pi_0 \quad N < i \leq K \\
\pi_0 &= \left[\sum_{i=0}^{N}\frac{1}{i!}\left(\frac{\hat{\mu}}{\hat{\lambda}}\right)^{-i} + \sum_{i=N+1}^{K}\frac{N^N}{N!}\left(\frac{\hat{\lambda}}{N\hat{\mu}}\right)^{i}\right]^{-1}.
\end{aligned} \tag{12}$$

Finally, we have that

$$E\left[[\mathbb{Q}-N]^+\right] \approx E[Queue_{M/M/N/K}] = \sum_{i=N}^{K}(i-N)\pi_i.$$

In Section 5, we will see that this approximation works very well. To provide some intuition as to why this seems to provide a very accurate approximation, we consider the impact of the finite buffer in the loss model. When there are K jobs in an M/M/N/K queue, the loss of any new job 'forces' the system away from this boundary. The speedup and/or admission control thresholds have a similar effect. When the number of patients in the system crosses one of these thresholds, the change in dynamics due to increased service rate and/or decreased arrival rate also 'forces' the system down and away from that threshold. When these thresholds are less than N, this strong push is active prior to a queue forming (i.e. when $\mathbb{Q} < N$). Since we are interested in examining the queue length, $\mathbb{Q} > N$, ignoring the change in dynamics before the queue forms does not degrade our approximation. If we were interested in approximating the precise distribution of patients in the system, it is likely the M/M/N/K approximation is too coarse; however, it seems to work quite well in approximating the mean queue length.

Remark 3 *One can utilize the derived approximations for the mean queue length and the probability of admission control and speedup to do performance analysis given thresholds N_a and N_s. Moreover, it is possible to determine a feasible set of N_a and N_s such that various constraints on these performance measures are satisfied. For example, if hospital management set a limit on the proportion of time that admission control is utilized, our approximations would provide a set of thresholds to satisfy such a constraint.*

4.3. Cost Minimization

Now that we have derived approximations for the different performance metrics, we are in position to consider our original optimization problem from (1). Because we've established the optimality of threshold policies in Section 3, our optimization problem can be reduced to:

$$\min_{N_a,N_s}\{h(E\left[[\mathbb{Q}-N]^+\right]) + P(\mathbb{Q} \geq N_a)\phi(\lambda_L) + P(\mathbb{Q} < N_a)\phi(\lambda_H)$$

$$+ P(\mathbb{Q} \geq N_s)\xi(\mu_H) + P(\mathbb{Q} < N_s)\xi(\mu_L)\}. \tag{13}$$

With our approximations from Sections 4.1 and 4.2, we now have closed form expressions for approximations to the optimization problem in (13). Without loss of generality, we set $\phi(\lambda_H) = 0$ and $\xi(\mu_L) = 0$, so there is no cost associated with the nominal system arrival and service rates. We also consider a linear function for the queue length costs. Hence, our optimization model is:

$$\min_{N_a, N_s} \{c_w E\left[[\mathbb{Q} - N]^+\right] + c_a P(\mathbb{Q} \geq N_a) + c_s P(\mathbb{Q} \geq N_s)\} \tag{14}$$

where c_w is the per-patient waiting cost rate, $c_a = \phi(\lambda_L)$ is the cost rate for admission control, and $c_s = \xi(\mu_H)$ is the cost rate for speedup.

Observation 1 *Using the approximations in Section 4.1 and 4.2 we find regimes in which our approximations suggest the cost in* (14) *to be zero. In particular, this occurs in cases 1.1, 2.1 and 3.1. Thus, the cost* (14) *is zero when:*

$$q^{HL} \leq N_a \wedge N_s \leq N.$$

This implies that it always 'optimal' to use at least one form of congestion control before *a queue builds* regardless *of the exact system parameters; a zero cost solution will never have both* $N_a, N_s > N$.

There has been some evidence that hospitals do use speedup and/or admission control *before* reaching full capacity (e.g. Kim et al. (2014) and Kc and Terwiesch (2012)). However, in other cases, such as ambulance diversion, admission control is not used until a queue builds (Allon et al. 2013).

As we will see in Section 5.2, one must be prudent with how to select between these seemingly zero-cost solutions.

4.3.1. A Greedy Heuristic We suggest the following *Greedy* heuristic to select the (N_a, N_s) amongst the potentially numerous solutions with the minimal approximated cost as indicated in Observation 1. In general, this heuristic prioritizes the use of the speedup and admission control in decreasing order of the cost measure. In order to ensure the costs are comparable, we use normalized costs. While, P(Speedup) and P(Admission Control) are naturally normalized to the range $[0, 1]$, the range of $E[Queue_{M/M/N/K}]$ can vary dramatically. Hence, we normalize the waiting costs by dividing them by the maximum expected queue length, denoted by $Q_{\max}$, which is obtained when speedup or admission control are never used.

The Greedy heuristic then selects among the potentially numerous solutions with approximated zero costs. For example, if speedup has the highest costs (c_s is maximal) then N_s should be as large as possible. This implies that only admission control is used; additionally, in light of Observation 1, we know that $q^{HL} \leq N_a \leq N$. The value of N_a will then be selected based on the relative costs between admission control and

waiting. Under a similar argument, if admission control is most expensive (c_a is maximal) then N_a should be as large as possible and $q^{HL} \leq N_s \leq N$. Finally, if waiting is most costly ($c_w/Q_{\max}$ is maximal), then N_a and N_s should be as small as possible. The Greedy heuristic is defined more formally by the pseudo-code in Algorithm 1. In Section 5.2, we will use simulation to examine how well such an approach performs.

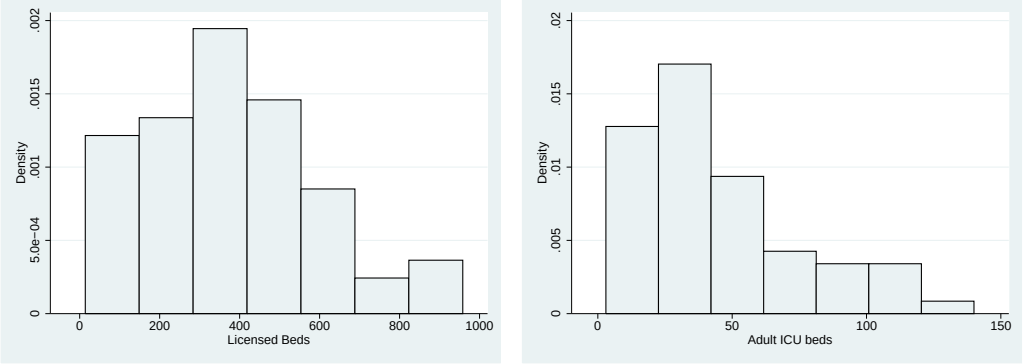
5. Numerical Results

In this section we examine the accuracy of our approximations. Subsection 5.1 presents the accuracy of the fluid approximations for $P(Speedup)$, $P(AdmissionControl)$ and the M/M/N/K model approximation for $E[Queue]$. Assuming these are reasonably accurate, the next step is to consider the performance of decisions which are optimized over these approximations—this is presented in Subsection 5.2. In our numeric analysis, we consider two system sizes: a small system representing an ICU setting and a large system representing the entire hospital. As our approximations are based on fluid analysis, we expect that they will be more accurate for the larger system.

5.1. Numerical Results: Performance Measure

We calibrate the parameters of our model according to typical healthcare environments. We used publicly available data from State of California Office of Statewide Health Planning & Development (2010-2011) which keeps track of all hospitals in California. We only considered short-term, acute hospitals with 24 hour emergency care coverage, trauma designation, and ICUs. Figure 3 shows the empirical distribution of the number of licensed (a) hospital beds and (b) adult ICU beds in these hospitals. The median number of licensed hospitals beds is 377 and the median number of licensed adult ICU beds is 38.5; as such, we will consider a hospital with $N = 400$ beds and an ICU with $N = 40$ beds for our simulations.

Figure 3 **Empirical distribution of the number of licensed beds in California hospitals, as reported in State of California Office of Statewide Health Planning & Development (2010-2011).**



(a)Number of licensed hospital beds

(b)Number of licensed Adult ICU beds

In calibrating the remaining parameters of our model, we start by considering the Length-of-stay (LOS) of a patient. A typical average LOS varies between 2 to 9 days, depending on the hospital unit considered

Algorithm 1 (N_a^g, N_s^g)= Greedy Heuristic(c_s, c_a, c_w)

$N\_MIN\_FLUID \leftarrow$ all (N_a, N_s) pairs with fluid cost approximation equal to the minimum value, i.e. $C_{fluid}(N\_MIN\_FLUID) = \min_{\{N_a,N_s\}} C_{fluid}(N_a, N_s)$.
if $\max(c_s, c_a, c_w) = c_s$ **then**
  $N_s^g \leftarrow \max(N\_MIN\_FLUID\{N_s\})$
  $N\_MAX\_NS \leftarrow$ all $(N_a, N_s = N_s^g) \in N\_MIN\_FLUID$
  if $\max(c_a, c_w) = c_a$ **then**
    $N_a^g \leftarrow \max(N\_MAX\_NS\{N_a\})$
  else
    $N_a^g \leftarrow \min(N\_MAX\_NS\{N_a\})$
  end if
else if $\max(c_s, c_a, c_w) = c_a$ **then**
  $N_a^g \leftarrow \max(N\_MIN\_FLUID\{N_a\})$
  $N\_MAX\_NA \leftarrow$ all $(N_a = N_a^g, N_s) \in N\_MIN\_FLUID$
  if $max(c_s, c_w) = c_s$ **then**
    $N_s^g \leftarrow \max(N\_MAX\_NA\{N_s\})$
  else
    $N_s^g \leftarrow \min(N\_MAX\_NA\{N_s\})$
  end if
else ($\max(c_s, c_a, c_w) = c_w$)
  if $\max(c_s, c_a) = c_s$ **then**
    $N_a^g \leftarrow \min(N\_MIN\_FLUID\{N_a\})$
    $N\_MIN\_NA \leftarrow$ all $(N_a = N_a^g, N_s) \in N\_MIN\_FLUID$
    $N_s^g \leftarrow \min(N\_MIN\_NA\{N_s\})$
  else
    $N_s^g \leftarrow \max(N\_MIN\_FLUID\{N_s\})$
    $N\_MIN\_NS \leftarrow$ all $(N_a, N_s = N_s^g) \in N\_MIN\_FLUID$
    $N_a^g \leftarrow \min(N\_MIN\_NS\{N_a\})$
  end if
end if

(see, for example, Table 3 in de Bruin et al. (2010)). Hence, we chose a lower value of 3 hospital days as the LOS under speedup and 5 days as the LOS under 'unstressed', nominal conditions. The arrival rates are chosen in order to have approximately 20% bed turnover per day under high arrival rates and 10% under low arrival rates. For our simulations, we use the following parameters for an 'ICU' (i.e. small system): $\lambda_L = 4, \lambda_H = 7.5, \mu_L = 0.2, \mu_H = 0.286, N = 40$. For an average sized hospital (i.e. large system), we use: $\lambda_L = 50, \lambda_H = 78, \mu_L = 0.2, \mu_H = 0.286, N = 400$. Note that the parameters chosen here satisfy Assumption 1, so that the system is stable irrespective of whether or not speedup and/or admission control are used.

Figure 4 presents the performance metrics given by our approximations and simulation results for the large system (left column) and small system (right column) as we vary the thresholds N_a and N_s[1]. These figures are meant to illustrate the typical behavior and effect of the control thresholds. The results are very similar across all combinations of the thresholds, N_a and N_s.

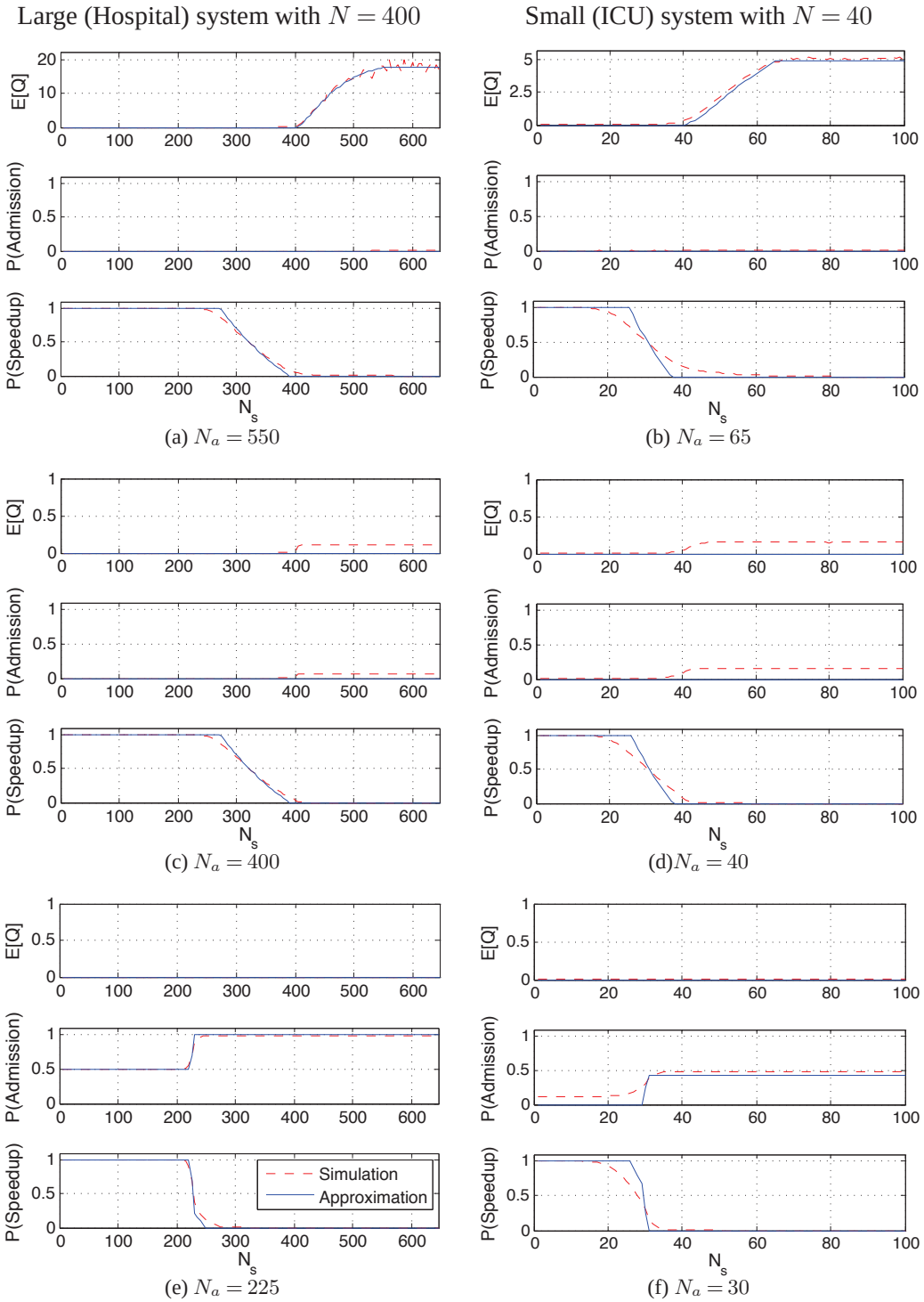
As expected, the approximations are more accurate for the large system. Still, they can be quite accurate for the small system as well. We observe that the approximations are very good in most cases. Some gaps can be observed when speedup and/or admission control is used for some (but small) proportion of the time (e.g. 5–10%) or when it is used most (but not all) of the time (e.g. 90–95%). This is more pronounced in the small system. In such situations, the fluid approximations for P(Speedup) and/or P(Admission Control) are not very accurate. This can result in degradation in the queue length approximation. For example, in Figure 5(d), the approximation for the probability of admission control is 0, while the simulation suggests the true probability is around 10%. Because our queue length approximation uses the probability of admission control to derive an 'effective' arrival rate, $\hat{\lambda}$, poor estimates for P(Admission Control) also result in poor estimates for $E[\mathbb{Q}]$. This is not always the case. In Figure 5(f), the approximation of P(speedup) under estimates the simulated value. However, in this case, the expected queue length is effectively 0 in the simulation and via our approximation.

5.2. Numerical Results: Approximation-Based Cost Minimization

As the performance metric approximations appear to be quite accurate, the next step is to consider the performance of policies resulting from solving an optimization problem based on them. The normalization factor $Q_{\max}$ in our examples are: in the large system $Q_{\max} = 19.7$, and in the small system $Q_{\max} = 8.9$. Here we find N_a and N_s which minimize the approximated costs and use simulation to compare the resulting cost to the minimum cost achieved via exhaustive search over many N_a and N_s combinations. We also consider a number of benchmarks for comparison:

[1] Note that, due to numeric issues, in order to calculate the expected queue length for large systems we used Stirling's formula $\left(i! \sim \left(\frac{i}{e}\right)^i \sqrt{2\pi i}\right)$ (Hazewinkel 2001) to calculate $E[Queue]$.

Figure 4 **Approximation vs. simulation results as a function of the Speedup threshold (N_s) for some fixed Admission threshold (N_a) values**



(a) $N_a = 550$ (b) $N_a = 65$

(c) $N_a = 400$ (d) $N_a = 40$

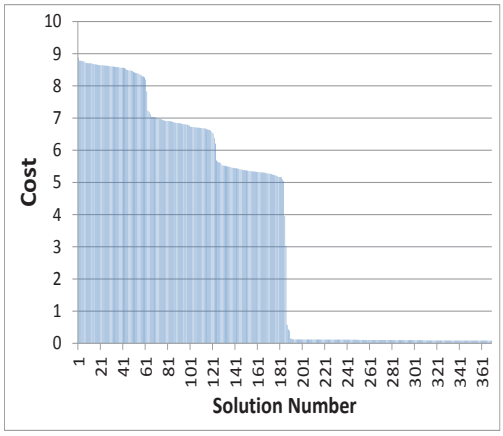
(e) $N_a = 225$ (f) $N_a = 30$

- **Never** use speedup or admission control: $N_a = N_s = \infty$
- Use speedup and admission control as soon as **all beds are filled**: $N_a = N_s = N$

- **Always** use speedup or admission control: $N_a = N_s = 0$

We derive the optimal performance via exhaustive search; for the large system we checked integer thresholds between 0 to 700 in jumps of 5[2], and for the small systems we search all combinations of integer thresholds between 0 to 100. The fluid approximation is derived by solving the optimization problem in (14) using the approximations for P(Speedup), P(Admission Control) and E[Queue] as given in Sections 4.1 and 4.2.

Figure 5 Distribution of simulation costs when approximated costs equal 0 (large system, scenario 11).



As noted in Section 4.3, when considering the fluid approximated costs, there may be many solutions with zero cost. For example, in the large system there are 369 solutions as such. Hence, our first question is are they all "practically" equivalent? As an example, lets assume that $c_s = 100$, $c_a = 1$, and $c_w = 10$ ($c_w/Q_{\max} = .51$). Figure 5 shows the actual costs (via simulation) for each solution that has approximated cost equal 0. We observe that although many of the solutions have simulated cost very close to 0, in others it is very different; nevertheless, about half of the solutions are indeed very close the optimal performance of 0.09.

Observation 2 *While the approximated costs in* (14) *are 0 for a (potentially large) set of N_a and N_s pairs, the* actual *cost associated with these solutions can differ significantly. However, we find that almost half of the solutions are indeed very close to optimal.*

Hence, the question of how to select between the minimal *fluid* solutions is highly pertinent. Choosing randomly amongst the 369 solutions with zero fluid cost would result in an average cost of 3.53, which is far from optimal. The solution with the minimum cost within the set has cost of 0.09 which is exactly

[2] The exhaustive search was computationally intensive, requiring nearly two weeks to complete on an Intel Xeon E5-2470, 2.3Ghz, 16 core CPU. Thus, providing more granularity was computationally limiting.

equal to the optimal value. Yet, finding this solution requires simulating all 369 zero fluid cost solutions. Although this is not as extensive as the run that is required for an exhaustive search, it is still computationally intensive. With this in mind, we consider the Greedy heuristic presented in Section 4.3. We next examine its performance for different cost scenarios.

Table 2 summarizes the cost scenarios we consider. We chose these scenarios in order to examine what happens where the costs have the same order of magnitude (scenarios 1–6), but still different order of importance; or when they are of different orders of magnitude (scenarios 7–12), highly emphasizing one measure in particular.

Cost Scenario	Speedup Control c_s	Admission Control c_a	Waiting Cost (before normalization) c_w
1	1	2	3
2	1	3	2
3	2	1	3
4	2	3	1
5	3	1	2
6	3	2	1
7	1	10	100
8	1	100	10
9	10	1	100
10	10	100	1
11	100	1	10
12	100	10	1

Table 2 Cost parameters for different cost minimization scenarios.

Tables 3 and 4 present the performance of the different strategies, for the various cost scenarios. The first thing we note is the *robustness* of the optimal solution. The left column—*optimal*—presents the minimal solution found using exhaustive search for each cost scenario. Analyzing this solution, we note that the minimal cost is robust. For example, for the large system there are between 8–14 solutions within 5% of the minimal one, and 57–105 solutions within 10% of it. (The numbers for the small systems are 6–176 solutions within 5% and 15–477 within 10% of the optimal performance). Note also that the "optimal" value was found by simulation; hence, numerical errors may indicate that one of the close solutions is indeed the optimal one. Under the *via Fluid Approximation* column we present the minimal simulated cost within the set of zero-value fluid solutions (Min), the average performance of these solutions (Avg), and the performance of the solution selected by the Greedy heuristic (Greedy). We make the following observations:

1. In most cases, the optimal value is within the set of zero fluid cost solutions.
2. The difference between the minimal performance of the zero fluid cost solution set and the optimal one is very small (often times it is 0).

3. The average cost of the zero fluid cost solution set may be quite far from optimal; hence, it is important to choose wisely within this range.
4. The Greedy heuristic is very close to optimal, achieving in most cases the minimal performance. Thus, prioritizing admission control and/or speedup based on relative costs can be very cost-effective.

The last three columns show the performance of the benchmark policies: Never, Always, and All beds filled. We see that the simple benchmarks fail dramatically in all scenarios; the only reasonable one is the "All beds filled" policy. Still, while the Greedy heuristic failed only in one scenario (in scenario 9 of the small system), the "All beds filled" performs well only for large systems where costs are close in magnitude (Scenarios 1–6); even then, it never achieves the optimal cost. We surmise that using the combination of the fluid approximation with the Greedy heuristic is a very good way to find a solution with near optimal performance.

Cost Scenario	Policy						
	Optimal	via Fluid Approximations			Never: $N_a = N_s = \infty$	Always: $N_a = N_s = 0$	All beds filled: $N_a = N_s = N$
		Min	Avg	Greedy			
1	0.07	**0.07***	0.12	**0.07***	3.06	3.00	0.09
2	0.06	**0.06***	0.16	**0.06***	2.04	4.00	0.12
3	0.08	**0.08***	0.12	**0.08***	3.06	3.00	0.09
4	0.10	**0.11**	0.19	**0.11**	1.02	5.00	0.15
5	0.07	**0.07***	0.15	**0.07***	2.04	4.00	0.12
6	0.12	**0.13**	0.19	**0.13**	1.02	5.00	0.15
7	0.10	**0.10***	0.66	**0.10***	101.84	11.00	0.45
8	0.07	**0.07***	4.17	0.11	10.18	101.00	3.03
9	0.10	**0.10***	0.60	0.11	101.84	11.00	0.45
10	0.35	**0.51**	446	0.52	1.02	110.00	3.29
11	0.09	**0.09***	3.53	0.11	10.18	101.00	3.03
12	0.40	**0.61**	3.88	**0.61**	1.02	110.00	3.29

Bold font indicates the method that got the minimal cost value; * indicates optimal costs

Table 3 Large system: Performance of different strategies for cost scenarios 1–12.

5.3. Cost Misspecification: Robustness of Proposed Greedy Policy

We also consider the robustness of our proposed greedy heuristic to miss-estimates in the cost parameters: c_w, c_a and c_s. Certainly, if the optimization is done over costs which are incorrectly specified, the resulting policy will be suboptimal. The question is how much worse will the performance be. Additionally, since we know the greedy heuristic is suboptimal, how will its performance be impacted by such misspecification?

To examine this, we consider the thresholds, N_a and N_s, selected under the Greedy Heuristic and Optimal Policy when the costs are misspecified by plus or minus 10%, 20%, 30%, 40%, and 50%. The thresholds

Cost Scenario	Policy						
	Optimal	via Fluid Approximations			Never: $N_a = N_s = \infty$	Always: $N_a = N_s = 0$	All beds filled: $N_a = N_s = N$
		Min	Avg	Greedy			
1	0.24	**0.24***	0.32	0.26	2.76	3.00	0.29
2	0.22	**0.22***	0.40	**0.22***	1.84	4.00	0.37
3	0.20	**0.20***	0.32	**0.20***	2.76	3.00	0.29
4	0.29	**0.34**	0.47	0.35	0.92	5.00	0.45
5	0.18	**0.18***	0.40	0.19	1.84	4.00	0.37
6	0.27	**0.32**	0.47	**0.32**	0.92	5.00	0.45
7	0.62	**0.97**	2.73	1.42	91.80	11.00	1.54
8	0.38	**0.39**	9.29	0.40	9.27	101.00	9.12
9	0.37	**0.56**	2.73	1.42	91.80	11.00	1.54
10	0.69	**1.56**	9.95	1.59	1.01	110.00	9.87
11	0.26	**0.26***	9.30	**0.26***	9.27	101.00	9.12
12	0.67	**1.53**	9.96	1.54	1.01	110.00	9.87

Bold font indicates the method that got the minimal cost value; * indicates optimal costs

Table 4 Small system: Performance of different strategies for cost scenarios 1–12.

of the Optimal Policy under cost misspecification are determined via exhaustive search over the incorrect cost parameters. The performance of these policies are then evaluated using simulation over the correct cost parameters. The worse case performance over all possible misspecifications is reported.

Table 5 summarizes the robustness results for our 12 cost scenarios in terms of a relative ratio between the average cost of the polices under misspecified costs to the true optimal solution (without misspecification). We can see that for up to 20% (and often 30+%) errors in cost estimates, the performance of the Greedy Heuristic is very robust. If the initial performance of the Greedy Heuristic, under perfect cost information, was reasonable, then this will still be the case, even if the costs have moderate misspecification. Of course, if the performance is poor under perfect cost information and/or the cost misspecification is very high, then the performance of the Greedy Heuristic can degrade substantially. Interestingly, the Greedy Heuristic is much more robust than the optimal policy. In some instances, the performance of the Greedy Heuristic under cost misspecification is better than that of the Optimal Policy under cost misspecification. For example, under cost scenario 3, the Greedy Heuristic achieves the minimum cost with errors up to 30% and it outperforms the misspecified Optimal Policy for up to 50% errors. This robustness feature, along with the simplicity of the heuristic, is another desirable property of the proposed heuristic.

6. A System with Returns to Service

Thus far, we have only accounted for the undesirability of delays, admission control and speedup via a cost function, which can capture clinical and/or monetary costs. However, it is known that these dynamics can reduce quality of service and that the deterioration of a patient's physiologic state may require a return to

Cost Scenario	Relative performance of 'Optimal Policy' with cost misspecifications						Relative performance of Greedy Heuristic with cost misspecifications					
	0 %	10 %	20 %	30 %	40 %	50 %	0%	10 %	20 %	30 %	40 %	50 %
1	1.000	1.001	1.001	1.399	1.399	2.055	1.019	1.019	1.019	1.019	2.055	2.055
2	1.000	1.000	1.000	1.074	1.074	1.074	1.018	1.018	1.018	1.018	1.018	3.147
3	1.000	1.000	1.225	1.539	1.716	1.716	1.000	1.000	1.000	1.000	1.565	1.565
4	1.000	1.027	1.027	1.633	1.815	1.815	1.043	1.043	1.043	1.815	1.815	1.815
5	1.000	1.000	1.000	1.000	1.717	2.282	1.000	1.000	1.000	1.000	1.000	1.000
6	1.000	1.102	1.180	1.292	1.292	1.346	1.046	1.046	1.046	1.312	1.312	1.312
7	1.000	1.007	1.048	1.066	1.066	1.402	1.045	1.045	1.045	1.045	5.615	5.615
8	1.000	1.000	1.000	1.000	1.000	1.150	1.456	1.456	1.456	1.456	1.456	1.456
9	1.000	1.002	1.007	1.025	1.199	1.297	1.041	1.041	1.041	1.041	5.352	5.352
10	1.000	1.000	1.085	1.140	1.269	1.343	1.479	1.479	1.479	1.479	1.479	1.479
11	1.000	1.000	1.000	1.000	1.003	1.309	1.309	1.309	1.309	1.309	1.309	1.309
12	1.000	1.016	1.045	1.045	1.215	1.257	1.535	1.535	1.535	1.535	1.535	1.535

Table 5 Large system: Robustness of greedy policy—Relative performance of Optimal Policy and Greedy Heuristic when cost parameters are misspecified to the true minimum cost.

service. A common quality measure used in practice is readmission rates. Using simulation, we examine an extended model that incorporates patients' readmissions explicitly, and use our original model (without readmissions) to determine policies which minimize the readmission rate for the extended model.

To incorporate readmissions, we assume that waiting for service and/or using speedup or admission control increases the likelihood of readmission. Without loss of generality, we assume the readmission risk to be 0 if neither speedup or admission control are used (i.e. $\mu = \mu_L, \lambda = \lambda_H$) and the new patient does not have to wait to begin service ($\mathbb{Q}_t < N$ where t is the patient's arrival time). If a patient would have arrived under the nominal arrival rate, λ_H, but was blocked due to admission control, this patient may return to service with probability p_λ^R after some time which is exponentially distributed with mean $1/\delta_a$. Similarly, if a patient is discharged under speedup, his probability of a return to service increases by p_μ^R; he returns after $1/\delta_s$ units of time (on average). If the patient arrives to the system with $\mathbb{Q}$ patients waiting in front of him for service, his probability of return to service increases by $p_w^R \times (\mathbb{Q} - N)^+$, where $p_w^R \times (\mathbb{Q}_{\max} - N)^+ \ll 1$. Thus, if a patient arrives with q patients in the system and then is discharged under speedup, his probability of return to service is $p_\mu^R + p_w^R \times (q - N)^+$. On the other hand, if the same patient is discharged under the nominal service rate, his probability of return to service is $p_w^R \times (q - N)^+$. We simulate such a model for each N_a and N_s combination with 40 iterations of 100 days each[3]. We then use exhaustive search to find the thresholds, N_a and N_s, which minimize the readmission rate.

[3] Given the computational complexity of this simulation–we must keep track of the number of customers in the system upon arrival for each customer–the number of repetitions was limited to 40.

As a comparison, we use the analysis of our original model from Section 2 *without* readmission, to determine thresholds, N_a and N_s. We will then use these thresholds in the model with readmissions and simulate the resulting readmission rate of this policy. To do this, we need to appropriately define our cost parameters for our original model to capture the *increase in readmission rate* due to admission control, speedup and waiting. Doing so results in $c_a = p_\lambda^R, c_s = p_\mu^R, c_w = p_w^R$.

For illustrative purposes, we let $p_\lambda^R = 0.05, p_\mu^R = 0.0667, p_w^R = 0.001$ (see, for example, Kim et al. (2014), Chan et al. (2014)). Table 6 compares the simulated return rates of an exhaustive search over the systems with and without readmissions. We observe very small differences in performance; for the small system, the difference is not even statistically significant. As seen in Figure 6, poor selection of thresholds can result in increases in readmission rates of up to 8%. Additionally, we find that the structure of the readmission rate as a function of N_a and N_s is very similar to that of the cost function in our original model without returns. In particular, the optimal regime of (N_a, N_S) which minimizes readmission rates for the system with returns is practically identically to the optimal regime which minimizes costs for the original model presented in Section 2. Additionally, we find that the minimum readmission rates are quite robust, as in Section 5.3. Hence, incorporating readmissions into our original model through appropriately defined cost factors seems to work quite effectively and avoids that complexity associated with explicitly including returns to service in the model.

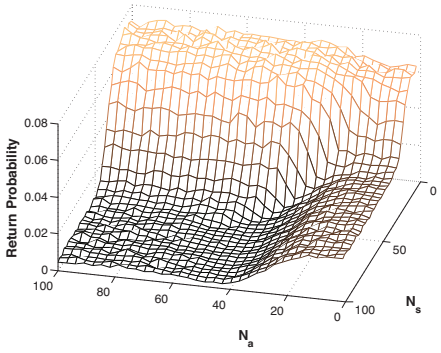
	Original Model (without returns)	Extended Model (with returns)
ICU setting ($N = 40$)	0.23%	0.18%
95% confidence interval	[0.16,0.31]	[0.11,0.24]
Hospital setting ($N = 400$)	0.138%	0.085%
95% confidence interval	[0.12,0.16]	[0.07,0.1]

Table 6 A system with returns to service: Comparison of return rates for solution which ignores returns to service but has appropriately define cost measures (Original Model) to solution established via exhaustive search over a model which explicitly incorporates readmissions.

7. Conclusion

In this paper, we examined the trade-off between delaying customers, blocking them via an admission control policy and speeding up services. We showed that if the cost function is either linear or concave the optimal policy has a distinct structure of a threshold policy. We then investigate the dynamics of a service system with such threshold policy using fluid approximations, and retrieve approximations for the main performance measures of the system: the proportion of time admission control is used, the proportion of time speedup is used, and the expected queue length. Using simulation, we found that these approximations can

Figure 6 Small System: Readmission rates as a function of N_s and N_a.



be very accurate. We then used these approximations in the original cost minimization problem, identified a set of solutions with seemingly zero costs, and developed a heuristic that achieves near optimal performance. Our results can be utilized in two ways: 1) to estimate the performance of a specific admission control and speedup policy, or 2) to find a reasonable admission control and speedup policy. Our proposed heuristic is based on fluid estimates and seems very robust to cost misspecifications.

One potential future direction for further exploration is to consider what happens when Assumption 1 is relaxed. We believe that many of the structural results should hold for any $N > \lambda_H/\mu_L$. However, we expect that the performance measure approximations (especially $E[Queue]$) and the optimal solution are very different from what we showed here. Different techniques are likely necessary to develop an understanding of such systems. Still, it is highly undesirable to operate a system which is unstable under nominal control; thus, we believe that understanding the behavior of our system under the assumption that $N < \lambda_H/\mu_L$ is an important first step.

While we did not explicitly consider returns to service in our analytic model, we find that, with appropriately defined cost parameters, our model–without readmissions–can perform reasonably well. It would be interesting to explore a model which explicitly incorporates readmissions. We note that this is done in Chan et al. (2014) *without* admission control and absent of an optimization framework and it is not obvious how those techniques (which are also used here) can be extended to this more complex flow model.

Finally, we note that time-varying arrival rates can arise in hospital settings. As seen in Chan et al. (2014) and Yom-Tov and Mandelbaum (2014), when the time scale of variation is short compared to the service time (LOS), then ignoring the time-variation can result in very reasonable performance. This is likely to be the case in our settings where the average LOS in the ICU and hospital are on the order of days, while the time-variation is on the order of hours. Of course, in an Emergency Department setting, where both the service time and time-variation are on the order of hours, accounting for the time-variability may be essential.

References

Adusumilli, K. M., J. J. Hasenbein. 2010. Dynamics admission and service rate control of a queue. *Queueing Systems* **66** 131–154.

Allon, G., S. Deo, W. Lin. 2013. The impact of hospital size and occupancy of hospital on the extent of ambulance diversion: Theory and evidence. *Operations Research* **61**(3) 544–562.

Anand, K., M. F. Pac, S. Veeraraghavan. 2010. Quality-speed conundrum: Tradeoffs in customer-intensive services. *Management Science* **57** 40–56.

Ata, B., S. Shneorson. 2006. Dynamic control of an M/M/1 service system with adjustable arrival and service rates. *Management Science* **52** 1778–1791.

Bekker, R., S.C. Borst. 2006. Optimal admission control in queues with workload-dependent service rates. *Probability in the Engineering and Informational Sciences* **20** 543–570.

Bekker, R., S.C. Borst, O.J. Boxma, O. Kella. 2004. Queues with workload-dependent arrival and service rates. *Queueing Systems* **46** 537–556.

Bekker, R., O.J. Boxma. 2007. An M/G/1 queue with adaptable service speed. *Stochastic Models* **23** 373–396.

Bekker, R., O.J. Boxma, J.A.C. Resing. 2008. Queues with adaptable service speed. *Statistica Neerlandica* **62** 441–457.

Bertsekas, D. 2001. *Dynamic Programming and Optimal Control*. Athena Scientific.

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

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

Chan, C. W., V. F. Farias, G. Escobar. 2013. The Impact of Delays on Service Times in the Intensive Care Unit. *Working Paper, Columbia Business School* .

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

de Bruin, A.M., R. Bekker, L. van Zanten, G.M. Koole. 2010. Dimensioning hospital wards using the Erlang loss model. *Annals of Operations Research* **178** 23–43.

di Bernardo, M., C.J. Budd, A.R. Champneys, P. Kowalczyk. 2008. *Piecewise-smooth dynamical systems: Theory and applications*. Springer.

Filippov, A.F. 1988. *Differential equations with discontinuous righthand sides*. Kluwer Academic Publishers, Dortrecht.

Green, L. V. 2003. How many hospital beds? *Inquiry* **39** 400–412.

Hasija, S., E. Pinker, R. A. Shumsky. 2010. OM PracticeWork Expands to Fill the Time Available: Capacity Estimation and Staffing Under Parkinson's Law. *MSOM* **12** 1–18.

Hazewinkel, M., ed. 2001. *Stirling formula, Encyclopedia of Mathematics*. Springer.

Kc, D., C. Terwiesch. 2012. An econometric analysis of patient flows in the cardiac intensive care unit. *Manufacturing & Service Operations Management* **14**(1) 50–65.

Kim, S.-H., C. W. Chan, M. Olivares, G. Escobar. 2014. ICU Admission Control: An Empirical Study of Capacity Allocation and its Implication on Patient Outcomes. *Working Paper, Columbia Business School* .

Lee, N., V.G. Kulkarni. 2014. Optimal arrival rate and service rate control of multi-server queues. *Queueing Systems* **76** 37–50.

Ormeci, E. L. 2004. Dynamic admission control in a call center with one shared and two dedicated service facilities. *IEEE Transactions on Automatic Control* **49** 1157–1161.

Powell, S. G., K. L. Schultz. 2004. Throughput in serial lines with state-dependent behavior. *Management Science* **50** 1095–1105.

Shevitz, Daniel, Brad Paden. 1994. Lyapunov stability theory of nonsmooth systems. *IEEE Transactions on Automatic Control* **39**(9) 1910–1914.

Shmueli, A., C.L. Sprung, E.H. Kaplan. 2003. Optimizing admissions to an intensive care unit. *Health Care Management Science* **6**(3) 131–136.

State of California Office of Statewide Health Planning & Development. 2010-2011. Annual Financial Data. URL `http://www.oshpd.ca.gov/HID/Products/Hospitals/AnnFinanData/CmplteDataSet/index.asp`.

Yom-Tov, G., A. Mandelbaum. 2014. Erlang-R: A time-varying queues with reentrant customers, in support of healthcare staffing. *MSOM* **16** 283–299.

Appendix

A. Miscellaneous Proofs

PROOF OF THEOREM 1: The proof of this theorem requires an intermediate result on the differential discounted cost, Δ.

Proposition 2*[Differential Monotonicity] The differential discounted cost function, $\Delta(\mathbb{Q})$, is non-decreasing in the number of jobs, $\mathbb{Q}$. That is, Let $\bar{\mathbb{Q}} > \mathbb{Q}$, then:*

$$\Delta(\mathbb{Q}) \leq \Delta(\bar{\mathbb{Q}}).$$

PROOF: The proof is via the value iteration method and induction. We generate a sequence of functions J_k starting with $J_0(\mathbb{Q}) = 0$ for all $\mathbb{Q} \geq 0$. Then for each $k > 0$ we have:

$$J_{k+1}(0) = \frac{1}{\beta + v} \min_{\lambda} \{\phi(\lambda) + (v - \lambda) J_k(0) + \lambda J_k(1)\}$$

Additionally, for $\mathbb{Q} > 0$, we have:

$$J_{k+1}(\mathbb{Q}) = \frac{1}{\beta+v} \min_{\lambda,\mu}\{h(\mathbb{Q}) + \phi(\lambda) + \xi(\mu) + \lambda J_k(\mathbb{Q}+1) + (\mathbb{Q} \wedge N)\mu J_k(\mathbb{Q}-1) + (v - \lambda - (\mathbb{Q} \wedge N)\mu)J_k(\mathbb{Q})\}$$

For $k \geq 0$ and $\mathbb{Q} > 0$, let

$$\Delta_k(\mathbb{Q}) = J_k(\mathbb{Q}) - J_k(\mathbb{Q}-1)$$

with $\Delta_k(0) = 0$. Using value iteration, we have that $J(\mathbb{Q}) = \lim_{k\to\infty} J_k(\mathbb{Q})$. It follows that $\Delta(\mathbb{Q}) = \lim_{k\to\infty} \Delta_k(\mathbb{Q})$. If we can show that $\Delta_k(\mathbb{Q})$ is non-decreasing in $\mathbb{Q}$ for every k, then the proposition is true. To do this, we use induction. The base case is trivially true for $k = 0$, where $\Delta_0(\mathbb{Q}) = 0$ for all $\mathbb{Q}$.

We will assume that the assertion is true for k and will show that it is also true for $k+1$. We denote $u_k(\mathbb{Q}+1) = (\lambda_k(\mathbb{Q}+1), \mu_k(\mathbb{Q}+1)) = \arg\min_{\lambda,\mu}\{\phi(\lambda) + \lambda\Delta_k(\mathbb{Q}+1) + \xi(\mu) - (\mathbb{Q} \wedge N)\mu\Delta_k(\mathbb{Q})\}$ as the strategy used in iteration $k+1$.

$$\begin{aligned}
\Delta_{k+1}(\mathbb{Q}+1) &= J_{k+1}(\mathbb{Q}+1) - J_{k+1}(\mathbb{Q}) \qquad (15)\\
&= \frac{1}{\beta+v}[h(\mathbb{Q}) + \phi(\lambda_k(\mathbb{Q}+1)) + \xi(\mu_k(\mathbb{Q}+1)) + vJ_k(\mathbb{Q}+1)\\
&\quad + \lambda_k(\mathbb{Q}+1)\left(J_k(\mathbb{Q}+2) - J_k(\mathbb{Q}+1)\right) - (\mathbb{Q} \wedge N)\mu_k(\mathbb{Q}+1)\left(J_k(\mathbb{Q}+1) - J_k(\mathbb{Q})\right) - J_{k+1}(\mathbb{Q})]\\
&\geq \frac{1}{\beta+v}[h(\mathbb{Q}+1) - h(\mathbb{Q}) + \lambda_k(\mathbb{Q}+1)\Delta_k(\mathbb{Q}+2)\\
&\quad + (v - \lambda_k(\mathbb{Q}+1) - (\mathbb{Q} \wedge N)\mu_k(\mathbb{Q}+1))\Delta_k(\mathbb{Q}+1) + (\mathbb{Q} \wedge N)\mu_k(\mathbb{Q}+1)\Delta_k(\mathbb{Q})]
\end{aligned}$$

where the last inequality comes from the fact that we can use the policy $u_k(\mathbb{Q}+1)$ at iteration $k+1$ in state $\mathbb{Q}$ and incur cost which is no less than $J_{k+1}(\mathbb{Q})$. Similarly, we can use the policy $u_k(\mathbb{Q}-1)$ in state $\mathbb{Q}$ at iteration $k+1$ and incur cost no less than $J_{k+1}(\mathbb{Q})$:

$$\begin{aligned}
\Delta_{k+1}(\mathbb{Q}) &= J_{k+1}(\mathbb{Q}) - J_{k+1}(\mathbb{Q}-1) \qquad (16)\\
&\leq \frac{1}{\beta+v}[h(\mathbb{Q}) - h(\mathbb{Q}-1) + \lambda_k(\mathbb{Q}-1)\Delta_k(\mathbb{Q}+1)\\
&\quad + (v - \lambda_k(\mathbb{Q}-1) - (\mathbb{Q}-1)\mu_k(\mathbb{Q}-1))\Delta_k(\mathbb{Q}) + (\mathbb{Q}-1)\mu_k(\mathbb{Q}-1)\Delta_k(\mathbb{Q}-1)].
\end{aligned}$$

Combining equations (15) and (16), for $\mathbb{Q} \geq 1$ we have that:

$$\begin{aligned}
(\beta+v)(\Delta_{k+1}(\mathbb{Q}+1) - \Delta_{k+1}(\mathbb{Q})) &\geq h(\mathbb{Q}+1) - h(\mathbb{Q}-1)\\
&\quad + [v - (\mathbb{Q} \wedge N)\mu_k(\mathbb{Q}+1) - \lambda_k(\mathbb{Q}+1)](\Delta_k(\mathbb{Q}+1) - \Delta_k(\mathbb{Q}))\\
&\quad + \lambda_k(\mathbb{Q}+1)(\Delta_k(\mathbb{Q}+2) - \Delta_k(\mathbb{Q}+1))\\
&\quad + (\mathbb{Q}-1)\mu_k(\mathbb{Q}-1)(\Delta_k(\mathbb{Q}) - \Delta_k(\mathbb{Q}-1))\\
&\geq 0 \text{ by the induction hypothesis.}
\end{aligned}$$

For the differential function when $\mathbb{Q} = 1$, we will use a suboptimal policy in state 0 so that the arrival and service rates are the same as those used in state 1, $\lambda_k(1)$ and $\mu_k(1)$, so that:

$$\begin{aligned}
\Delta_{k+1}(1) &= J_{k+1}(1) - J_{k+1}(0)\\
&\geq \frac{1}{\beta+1}[\xi(\mu_k(1)) + (v - \mu_k(1))\Delta_k(1) + \lambda_k(1)(\Delta_k(2) - \Delta_k(1))]
\end{aligned}$$

$$\geq 0 = \Delta_{k+1}(0)$$

where the first inequality follows from the induction hypothesis. This completes the proof that for all $\mathbb{Q}$ and k, $\Delta_{k+1}(\mathbb{Q}+1) \geq \Delta_{k+1}(\mathbb{Q})$, and so is also true in the limit as $k \to \infty$. □

Now, we consider the Bellman equation, where the arrival rate and service rate decisions can be separated:

$$J(\mathbb{Q}) = \frac{1}{\beta+v}\left[h(\mathbb{Q}) + vJ(\mathbb{Q}) + \min_{\lambda}\{\phi(\lambda) + \lambda\Delta(\mathbb{Q}+1)\} + \min_{\mu}\{\xi(\mu) - (\mathbb{Q}\wedge N)\mu\Delta(\mathbb{Q})\}\right].$$

- **Admission Control:** We first consider the optimization of the arrival rate. Our goal is to find $\lambda^*(\mathbb{Q})$ such that:

$$\lambda^*(\mathbb{Q}) = \arg\min_{\lambda}\{\phi(\lambda) + \lambda\Delta(\mathbb{Q}+1)\}.$$

By Proposition 2, we have that $\Delta(\mathbb{Q})$ is non-decreasing in $\mathbb{Q}$. By assumption, $\phi(\lambda)$ is non-increasing in λ. Hence, λ^* is also non-increasing in $\mathbb{Q}$.

- **Speedup:** We now consider the optimization of the service rate. Our goal is to find $\mu^*(\mathbb{Q})$ such that:

$$\mu^*(\mathbb{Q}) = \arg\min_{\mu}\{\xi(\mu) - (\mathbb{Q}\wedge N)\mu\Delta(\mathbb{Q})\}.$$

By Proposition 2, we have that $x\Delta(\mathbb{Q})$ is non-decreasing in $\mathbb{Q}$. By assumption, $\xi(\mu)$ is non-decreasing in μ. Hence, μ^* is also non-decreasing in $\mathbb{Q}$.

□

PROOF OF THEOREM 2: We again turn back to Bellman's equation:

$$J(\mathbb{Q}) = \frac{1}{\beta+v}\left[h(\mathbb{Q}) + vJ(\mathbb{Q}) + \min_{\lambda}\{\phi(\lambda) + \lambda\Delta(\mathbb{Q}+1)\} + \min_{\mu}\{\xi(\mu) - (\mathbb{Q}\wedge N)\mu\Delta(\mathbb{Q})\}\right].$$

- **Admission Control:** We first consider the optimization of the arrival rate. Our goal is to find $\lambda^*(\mathbb{Q})$ such that:

$$\lambda^*(\mathbb{Q}) = \arg\min_{\lambda\in[\lambda_L,\lambda_H]}\{\phi(\lambda) + \lambda\Delta(\mathbb{Q}+1)\}.$$

Consider a fixed $\mathbb{Q}$. Then $\Delta(\mathbb{Q}+1)$ is some non-negative constant. By assumption, $\phi(\lambda)$ is concave; hence, the portion of the cost function associated with the arrival rate:

$$\phi(\lambda) + \lambda\Delta(\mathbb{Q}+1) \text{ is concave.}$$

Since we are minimizing a concave function over a finite interval, the optimal admission rate must be at the boundary. Therefore, we must have that $\lambda^*(\mathbb{Q}) = \lambda_L$ or $\lambda^*(\mathbb{Q}) = \lambda_H$.

- **Speedup:** We now consider the optimization of the service rate. Our goal is to find $\mu^*(\mathbb{Q})$ such that:

$$\mu^*(\mathbb{Q}) = \arg\min_{\mu}\{\xi(\mu) - (\mathbb{Q}\wedge N)\mu\Delta(\mathbb{Q})\}.$$

Consider a fixed $\mathbb{Q}$. Then $(\mathbb{Q}\wedge N)\Delta(\mathbb{Q})$ is some non-negative constant. By assumption, $\xi(\mu)$ is concave; hence, the portion of the cost function associated with the arrival rate:

$$\xi(\mu) - (\mathbb{Q}\wedge N)\mu\Delta(\mathbb{Q}) \text{ is concave.}$$

Again, since we are minimizing a concave function over a finite interval, the optimal service rate must be at the boundary. Therefore, we must have that $\mu^*(\mathbb{Q}) = \mu_L$ or $\mu^*(\mathbb{Q}) = \mu_H$.

□

PROOF OF PROPOSITION 1: This is a direct result of Proposition 4.3.3 in Bertsekas (2001) on Blackwell Optimal policies. Note that the results of Theorem 2 (and Proposition 2) are independent of the exact value of the discount factor, β. Because there exists a Blackwell optimal policy, it must satisfy the structural properties derived in Theorem 2. Additionally, because a Blackwell optimal policy is also optimal for the average cost problem, the properties hold for the average cost problem. □

PROOF OF THEOREM 3: Our system is a piecewise-smooth set of ordinary differential equations. We will take a similar approach to that in Chan et al. (2014); however, our system has two regions of discontinuity, $Q = N_a$ and $Q = N_s$, and is one-dimensional. Still, we can utilize generalize Lyapunov techniques for discontinuous differential equations outlines in Filippov (1988) and di Bernardo et al. (2008). The main idea behind the Filippov (1988) approach is to use a 'smoothed' version of the ODE at the points of discontinuity, by using a convex combination of the surrounding smooth ODEs.

To show globally asymptotic stability, we need to identify a Lyapunov function and prove that for all $Q \geq 0, Q \neq \bar{q}$, the derivative of the Lyaponov function is strictly negative. We use the following Lyapunov function:

$$V(Q) = |Q - \bar{q}| \tag{17}$$

where $\bar{q}$ is the specified equilibrium. The main challenge here is that the ODE (4) is discontinuous. Hence, we need to use a generalized Lyapunov theory which utilizes Filippov solutions as done in Shevitz and Paden (1994). We use the Filippov methodology, which redefines the ODE at the points of discontinuity, N_a and N_s, as the set-valued function which is now equal to the convex combination of the surrounding smooth ODEs in (4). In order to establish global asymptotic stability, we need to show that the set value map for our generalized Lyapunov derivative is negative for all states not equal to the equilibrium (Shevitz and Paden 1994).

First, we introduce some notation to help as we define the generalize Lyapunov derivative. We consider the differential equations under policies which either 1) never use admission control or speedup 2) always use admission control and speedup 3) always use admission control, but never use speedup, and, finally 4) never use admission control, but always use speedup:

1. [Never use admission control or speedup:] $\dot{Q}^{HL}(t) \triangleq \lambda_H - \mu_L(Q(t) \wedge N)$.
2. [Always use admission control and speedup:] $\dot{Q}^{LH}(t) \triangleq \lambda_L - \mu_H(Q(t) \wedge N)$.
3. [Always use admission control. Never use speedup:] $\dot{Q}^{LL}(t) \triangleq \lambda_L - \mu_L(Q(t) \wedge N)$.
4. [Never use admission control. Always use speedup:] $\dot{Q}^{HH}(t) \triangleq \lambda_H - \mu_H(Q(t) \wedge N)$.

We can now define our set value map, generalized Lyapunov derivative. This requires considering a number of cases depending on the whether Q is on a point of discontinuity.

1. $[Q \neq N_a, N_s]$.

$$\dot{\tilde{V}}(Q) = \begin{cases} \dot{Q}, & Q > \bar{q}; \\ -\dot{Q}, & Q < \bar{q}. \end{cases} \tag{18}$$

2. $[Q = N_a \neq N_s]$. In this case, the flow is on a point of discontinuity, N_a; thus, the set value map is defined as the convex combination of the surrounding smooth ODEs.

$$\dot{\tilde{V}}(Q) = \begin{cases} \psi\dot{Q}^{LH} + (1-\psi)\dot{Q}^{HH}, \psi \in [0,1], & Q > \bar{q} \text{ and } Q > N_s; \\ \psi\dot{Q}^{LL} + (1-\psi)\dot{Q}^{HL}, \psi \in [0,1], & Q > \bar{q} \text{ and } Q < N_s; \\ -\psi\dot{Q}^{LH} - (1-\psi)\dot{Q}^{HH}, \psi \in [0,1], & Q < \bar{q} \text{ and } Q > N_s; \\ -\psi\dot{Q}^{LL} - (1-\psi)\dot{Q}^{HL}, \psi \in [0,1], & Q < \bar{q} \text{ and } Q < N_s. \end{cases} \tag{19}$$

3. $[Q = N_s \neq N_a]$. In this case, the flow is on a different point of discontinuity, N_s. We take a similar approach to what we did before;

$$\dot{\tilde{V}}(Q) = \begin{cases} \psi\dot{Q}^{LH} + (1-\psi)\dot{Q}^{LL}, \psi \in [0,1], & Q > \bar{q} \text{ and } Q > N_a; \\ \psi\dot{Q}^{HH} + (1-\psi)\dot{Q}^{HL}, \psi \in [0,1], & Q > \bar{q} \text{ and } Q < N_a; \\ -\psi\dot{Q}^{LH} - (1-\psi)\dot{Q}^{LL}, \psi \in [0,1], & Q < \bar{q} \text{ and } Q > N_a; \\ -\psi\dot{Q}^{HH} - (1-\psi)\dot{Q}^{HL}, \psi \in [0,1], & Q < \bar{q} \text{ and } Q < N_a. \end{cases} \tag{20}$$

4. $[Q = N_s = N_a]$. In this case, the flow is on the (only) point of discontinuity, $N_a = N_s$.

$$\dot{\tilde{V}}(Q) = \begin{cases} \psi\dot{Q}^{LH} + (1-\psi)\dot{Q}^{HL}, \psi \in [0,1], & Q > \bar{q}; \\ -\psi\dot{Q}^{LH} - (1-\psi)\dot{Q}^{HL}, \psi \in [0,1], & Q < \bar{q}. \end{cases} \tag{21}$$

ψ is simply a parameter to generate the convex combination of smooth ODEs. In order to prove global asymptotic stability, we must have $\dot{\tilde{V}}(Q) < 0$ for all $Q \geq 0$, $Q \neq \bar{q}$ and all $\psi \in [0,1]$. Due to the amount of algebra involved in this proof, we only include here the proof for Case 1, ACF ($N_a < N_s$), while noting the proofs for Case 2, SCF, and Case 3, SASC, will follow similarly. In this proof, it will be helpful to recall that by Assumption 1, $N > q^{HL} = \lambda_H/\mu_L$. Also, we do not need to consider the fourth case, $Q = N_a = N_s$, because we are currently examining the case where $N_a < N_s$.

Case 1.1 $q^{HL} \leq N_a$***:*** In this case, the equilibrium is $\bar{q} = q^{HL}$. We need to examine the three cases i. $Q \neq N_a, N_s$, ii. $Q = N_a$, and iii. $Q = N_s$. There are a number of subcases to consider within each case:

i. $[Q \neq N_a, N_s]$

(a) $Q > \bar{q} = q^{HL}$.

$$\begin{aligned} \dot{\tilde{V}}(Q) &= \dot{Q} = 1_{\{Q<N_a\}}\lambda_H + 1_{\{Q\geq N_a\}}\lambda_L - (1_{\{Q<N_s\}}\mu_L + 1_{\{Q\geq N_s\}}\mu_H)(Q \wedge N) \\ &\leq \lambda_H - \mu_L(Q \wedge N) < \lambda_H - \mu_L\bar{q} = \lambda_H - \mu_L q^{HL} = 0 \end{aligned}$$

(b) $Q < \bar{q} = q^{HL}$.

$$\dot{\tilde{V}}(Q) = -\dot{Q} = -\lambda_H + \mu_L(Q \wedge N) < -\lambda_H + \mu_L\bar{q} = -\lambda_H + \mu_L q^{HL} = 0$$

ii. $[Q = N_a \neq N_s]$ We want to show that for all $\psi \in [0,1]$, $\dot{\tilde{V}}(Q) < 0$:

(a) $Q > \bar{q}$ and $Q > N_s$. This case cannot occur because $Q = N_a < N_s$.

(b) $Q > \bar{q}$ and $Q < N_s$.

$$\begin{aligned} \dot{\tilde{V}}(Q) &= \psi\dot{Q}^{LL} + (1-\psi)\dot{Q}^{HL} = \psi[\lambda_L - \mu_L(Q \wedge N)] + (1-\psi)[\lambda_H - \mu_L(Q \wedge N)] \\ &\leq \lambda_H - \mu_L(Q \wedge N) < \lambda_H - \mu_L\bar{q} = \lambda_H - \mu_L q^{HL} = 0, \ \forall \psi \in [0,1] \end{aligned}$$

(c) $Q < \bar{q}$ and $Q > N_s$. This case cannot occur because $Q = N_a < N_s$.

(d) $Q < \bar{q}$ and $Q < N_s$. This case cannot occur because $Q < \bar{q} = q^{HL} \leq N_a$.

iii. $[Q = N_s \neq N_a]$ We want to show that for all $\psi \in [0,1]$, $\dot{\tilde{V}}(Q) < 0$:

(a) $Q > \bar{q}$ and $Q > N_a$.

$$\begin{aligned}\dot{\tilde{V}}(Q) &= \psi\dot{Q}^{LH} + (1-\psi)\dot{Q}^{LL} = \psi[\lambda_L - \mu_H(Q \wedge N)] + (1-\psi)[\lambda_L - \mu_L(Q \wedge N)] \\ &\leq \lambda_L - \mu_L(Q \wedge N) < \lambda_H - \mu_L(Q \wedge N) < \lambda_H - \mu_L q^{HL} = 0, \ \forall \psi \in [0,1]\end{aligned}$$

(b) $Q > \bar{q}$ and $Q < N_a$. This case cannot occur because $Q = N_s > N_a$.

(c) $Q < \bar{q}$ and $Q > N_a$. This case cannot occur because $Q < \bar{q} = q^{HL} \leq N_a$.

(d) $Q < \bar{q}$ and $Q < N_a$. This case cannot occur because $Q = N_s > N_a$.

Case 1.2 $q^{LL} \leq N_a \leq q^{HL}$***:*** In this case, the equilibrium is $\bar{q} = N_a$. We need to examine the two cases $Q \neq N_a, N_s$ and $Q = N_s$[4]. There are a number of subcases to consider within each of our two cases:

i. $[Q \neq N_a, N_s]$

(a) $Q > \bar{q} = N_a$.

$$\dot{\tilde{V}}(Q) = \dot{Q} = \lambda_L - \left(1_{\{Q<N_s\}}\mu_L + 1_{\{Q\geq N_s\}}\mu_H\right)(Q \wedge N) \leq \lambda_L - \mu_L(Q \wedge N) < \lambda_L - \mu_L q^{LL} = 0$$

(b) $Q < \bar{q} = N_a < N_s$.

$$\dot{\tilde{V}}(Q) = -\dot{Q} = -\lambda_H + \mu_L(Q \wedge N) < -\lambda_H + \mu_L\bar{q} \leq -\lambda_H + \mu_L q^{HL} = 0$$

ii. $[Q = N_s \neq N_a]$ We want to show that for all $\psi \in [0,1]$, $\dot{\tilde{V}}(Q) < 0$:

(a) $Q > \bar{q}$ and $Q > N_a$.

$$\begin{aligned}\dot{\tilde{V}}(Q) &= \psi\dot{Q}^{LH} + (1-\psi)\dot{Q}^{LL} = \psi[\lambda_L - \mu_H(Q \wedge N)] + (1-\psi)[\lambda_L - \mu_L(Q \wedge N)] \\ &\leq \lambda_L - \mu_L(Q \wedge N) < \lambda_L - \mu_L\bar{q} \leq \lambda_L - \mu_L q^{LL} = 0, \ \forall \psi \in [0,1]\end{aligned}$$

(b) $Q > \bar{q}$ and $Q < N_a$. This case cannot occur because $\bar{q} = N_a$.

(c) $Q < \bar{q}$ and $Q > N_a$. This case cannot occur because $\bar{q} = N_a$.

(d) $Q < \bar{q}$ and $Q < N_a$. This case cannot occur because $Q = N_s > N_a$.

Case 1.3 $N_a \leq q^{LL} \leq N_s$***:*** In this case, the equilibrium is $\bar{q} = q^{LL}$. We need to examine the three cases i. $Q \neq N_a, N_s$, ii. $Q = N_a$, and iii. $Q = N_s$. There are a number of subcases to consider within each case:

i. $[Q \neq N_a, N_s]$

(a) $Q > \bar{q} = q^{LL} \geq N_a$.

$$\dot{\tilde{V}}(Q) = \dot{Q} = \lambda_L - \left(1_{\{Q<N_s\}}\mu_L + 1_{\{Q\geq N_s\}}\mu_H\right)(Q \wedge N) \leq \lambda_L - \mu_L(Q \wedge N) < \lambda_L - \mu_L q^{LL} = 0$$

(b) $Q < \bar{q} = q^{LL} \leq N_s$.

$$\dot{\tilde{V}}(Q) = -\dot{Q} = -1_{\{Q<N_a\}}\lambda_H - 1_{\{Q\geq N_a\}}\lambda_L + \mu_L(Q \wedge N) \leq -\lambda_L + \mu_L(Q \wedge N) < -\lambda_L + \mu_L q^{LL} = 0$$

[4] We do not need to consider the second case because N_a is our equilibrium and our Lyapunov function is equal to 0 when $Q = N_a$.

ii. $[Q = N_a \neq N_s]$ We want to show that for all $\psi \in [0,1]$, $\dot{\tilde{V}}(Q) < 0$:

(a) $Q > \bar{q}$ and $Q > N_s$. This case cannot occur because $Q = N_a < N_s$.

(b) $Q > \bar{q}$ and $Q < N_s$. This case cannot occur because $Q = N_a \leq q^{LL} = \bar{q}$.

(c) $Q < \bar{q}$ and $Q > N_s$. This case cannot occur because $Q = N_a < N_s$.

(d) $Q < \bar{q}$ and $Q < N_s$.

$$\begin{aligned}\dot{\tilde{V}}(Q) &= -\psi\dot{Q}^{LL} - (1-\psi)\dot{Q}^{HL} = -\psi[\lambda_L - \mu_L(Q \wedge N)] - (1-\psi)[\lambda_H - \mu_L(Q \wedge N)] \\ &\leq -\lambda_L + \mu_L(Q \wedge N) < -\lambda_L + \mu_L\bar{q} = -\lambda_L + \mu_L q^{LL} = 0, \ \forall \psi \in [0,1]\end{aligned}$$

iii. $[Q = N_s \neq N_a]$ We want to show that for all $\psi \in [0,1]$, $\dot{\tilde{V}}(Q) < 0$:

(a) $Q > \bar{q}$ and $Q > N_a$.

$$\begin{aligned}\dot{\tilde{V}}(Q) &= \psi\dot{Q}^{LH} + (1-\psi)\dot{Q}^{LL} = \psi[\lambda_L - \mu_H(Q \wedge N)] + (1-\psi)[\lambda_L - \mu_L(Q \wedge N)] \\ &\leq \lambda_L - \mu_L(Q \wedge N) < \lambda_L - \mu_L q^{LL} = 0, \ \forall \psi \in [0,1]\end{aligned}$$

(b) $Q > \bar{q}$ and $Q < N_a$. This case cannot occur because $Q = N_s > N_a$.

(c) $Q < \bar{q}$ and $Q > N_a$. This case cannot occur because $Q = N_s \geq q^{LL} = \bar{q}$.

(d) $Q < \bar{q}$ and $Q < N_a$. This case cannot occur because $Q = N_s > N_a$.

Case 1.4 $q^{LH} \leq N_s \leq q^{LL}$***:*** In this case, the equilibrium is $\bar{q} = N_s$. We need to examine the two cases i. $Q \neq N_a, N_s$ and ii. $Q = N_a$[5]. There are a number of subcases to consider within each of our two cases:

i. $[Q \neq N_a, N_s]$

(a) $Q > \bar{q} = N_s > N_a$.

$$\dot{\tilde{V}}(Q) = \dot{Q} = \lambda_L - \mu_H(Q \wedge N) < \lambda_L - \mu_H\bar{q} \leq \lambda_L - \mu_H q^{LH} = 0$$

(b) $Q < \bar{q} = N_s$.

$$\begin{aligned}\dot{\tilde{V}}(Q) &= -\dot{Q} = -1_{\{Q<N_a\}}\lambda_H - 1_{\{Q\geq N_a\}}\lambda_L + \mu_L(Q \wedge N) \\ &\leq -\lambda_L + \mu_L(Q \wedge N) < -\lambda_L + \mu_L\bar{q} \leq -\lambda_L + \mu_L q^{LL} = 0\end{aligned}$$

ii. $[Q = N_a \neq N_s]$ We want to show that for all $\psi \in [0,1]$, $\dot{\tilde{V}}(Q) < 0$:

(a) $Q > \bar{q}$ and $Q > N_s$. This case cannot occur because $Q = N_a < N_s$.

(b) $Q > \bar{q}$ and $Q < N_s$. This case cannot occur because $Q > \bar{q} = N_s$.

(c) $Q < \bar{q}$ and $Q > N_s$. This case cannot occur because $Q = N_a < N_s$.

(d) $Q < \bar{q}$ and $Q < N_s$.

$$\begin{aligned}\dot{\tilde{V}}(Q) &= -\psi\dot{Q}^{LL} - (1-\psi)\dot{Q}^{HL} = -\psi[\lambda_L - \mu_L(Q \wedge N)] - (1-\psi)[\lambda_H - \mu_L(Q \wedge N)] \\ &\leq -\lambda_L + \mu_L(Q \wedge N) < -\lambda_L + \mu_L\bar{q} \leq -\lambda_L + \mu_L q^{LL} = 0, \ \forall \psi \in [0,1]\end{aligned}$$

Case 1.5 $N_s \leq q^{LH}$***:*** In this case, the equilibrium is $\bar{q} = q^{LH}$. We need to examine the three cases i. $Q \neq N_a, N_s$, ii. $Q = N_a$, and iii. $Q = N_s$. There are a number of subcases to consider within each case:

[5] We do not need to consider the third case because N_s is our equilibrium and our Lyapunov function is equal to 0 when $Q = N_s$.

i. $[Q \neq N_a, N_s]$

(a) $Q > \bar{q} \geq N_s > N_a$.

$$\dot{\tilde{V}}(Q) = \dot{Q} = \lambda_L - \mu_H(Q \wedge N) < \lambda_L - \mu_H \bar{q} = \lambda_L - \mu_H q^{LH} = 0$$

(b) $Q < \bar{q}$.

$$\begin{aligned}\dot{\tilde{V}}(Q) &= -\dot{Q} = -1_{\{Q<N_a\}}\lambda_H - 1_{\{Q\geq N_a\}}\lambda_L + (1_{\{Q<N_s\}}\mu_L + 1_{\{Q\geq N_s\}}\mu_H)(Q \wedge N) \\ &\leq -\lambda_L + \mu_H(Q \wedge N) < -\lambda_L + \mu_H \bar{q} = \lambda_L - \mu_H q^{LH} = 0\end{aligned}$$

ii. $[Q = N_a \neq N_s]$ We want to show that for all $\psi \in [0,1]$, $\dot{\tilde{V}}(Q) < 0$:

(a) $Q > \bar{q}$ and $Q > N_s$. This case cannot occur because $Q = N_a < N_s$.

(b) $Q > \bar{q}$ and $Q < N_s$. This case cannot occur because $Q = N_a < N_s \leq q^{LH} = \bar{q}$.

(c) $Q < \bar{q}$ and $Q > N_s$. This case cannot occur because $Q = N_a < N_s$.

(d) $Q < \bar{q}$ and $Q < N_s$.

$$\begin{aligned}\dot{\tilde{V}}(Q) &= -\psi\dot{Q}^{LL} - (1-\psi)\dot{Q}^{HL} = -\psi[\lambda_L - \mu_L(Q \wedge N)] - (1-\psi)[\lambda_H - \mu_L(Q \wedge N)] \\ &\leq -\lambda_L + \mu_L(Q \wedge N) < -\lambda_L + \mu_L q^{LH} < -\lambda_L + \mu_L q^{LL} = 0,\ \forall \psi \in [0,1]\end{aligned}$$

iii. $[Q = N_s \neq N_a]$ We want to show that for all $\psi \in [0,1]$, $\dot{\tilde{V}}(Q) < 0$:

(a) $Q > \bar{q}$ and $Q > N_a$.

$$\begin{aligned}\dot{\tilde{V}}(Q) &= \psi\dot{Q}^{LH} + (1-\psi)\dot{Q}^{LL} = \psi[\lambda_L - \mu_H(Q \wedge N)] + (1-\psi)[\lambda_L - \mu_L(Q \wedge N)] \\ &\leq \lambda_L - \mu_L(Q \wedge N) < \lambda_L - \mu_L q^{LL} = 0,\ \forall \psi \in [0,1]\end{aligned}$$

(b) $Q > \bar{q}$ and $Q < N_a$. This case cannot occur because $\bar{q} = q^{LH} \geq N_s > N_a$.

(c) $Q < \bar{q}$ and $Q > N_a$. This case cannot occur because $\bar{q} = q^{LH} \geq N_s = Q$.

(d) $Q < \bar{q}$ and $Q < N_a$. This case cannot occur because $\bar{q} = q^{LH} \geq N_s = Q$.

This concludes the proof for the global stability of Case 1. □