Optimal Scheduling of Proactive Care with Patient Deterioration

Yue Hu, Carri W. Chan, Jing Dong
Decision, Risk, and Operations, Columbia Business School
yhu22@gsb.columbia.edu, cwchan@columbia.edu, jing.dong@gsb.columbia.edu

Healthcare is a limited resource environment where scarce capacity is often reserved for the most severe patients. However, there has been a growing interest in the use of preventive care to provide treatment to patients early on, before they deteriorate. On one hand, providing care for patients when they are less critical could mean that fewer resources are needed to return them to a healthy, stable state. On the other hand, utilizing limited capacity for patients who may never need care in the future takes the capacity away from other more critical patients who need it now. To understand this tension, we propose a multi-server queueing model with two patient classes: moderate and urgent. A moderate patient who does not receive treatment may recover and leave or may deteriorate and become an urgent patient. In this setting, we characterize how moderate and urgent patients should be prioritized for care when proactive care for moderate patients is an option. We identify a parameter, the *modified* $c\mu/\theta$*-index*, which plays an important role in determining the optimal scheduling policy. This index lends itself to an intuitive interpretation of how to balance holding costs, service time, abandonment, and degradation of patient class.

Key words: Proactive Care, Mutli-class Queue, Optimal Control, Equilibrium, Transient Performance Analysis

1. Introduction

In recent years, with the advancement of predictive analytics and data availability, considerable efforts have been made to develop and utilize predictive tools in healthcare. For example, predictive models have been created to evaluate the risk of ICU admission (Churpek et al. 2014), hospital acquired infection (Chang et al. 2011), Cardiovascular events (Rumsfeld et al. 2016), and various other adversarial patient deterioration. From the operations perspective, this brings the opportunity of developing approaches to provide effective proactive care and, potentially, improve system performance (Hu et al. 2018). In particular, there are well-documented evidence that delayed treatment can lead to worse medical outcomes such as longer length of stay or higher mortality rate (Chan et al. 2008, Chalfin et al. 2007, Chan et al. 2016). Proactive care, with the help of the predictive models, can help reduce treatment delays and improve patient outcomes. That said, recent systematic reviews of early-warning-based alarms in the medical literature have demonstrated that they only marginally improve outcomes while substantially increasing the workload for physicians and nurses (Alam et al. 2014). The key reason for this is that existing predictive models are often not very accurate in identifying patients who will actually deteriorate (Yoon et al. 2016).

Isolating the potential impact of proactive care is not straightforward. On one hand, providing care for patients when they are less critical could mean that fewer resources are needed to return them to a healthy, stable state. This has the potential benefit of reducing the overall workload of the system. On the other hand, utilizing limited capacity for patients who are less critical may take capacity away from other more critical patients whose care needs are more urgent. Moreover, some of these less critical patients may become stable without ever needing the critical care. Thus, providing proactive care to them may end up generating more workload for the system. In this paper, to develop a better understanding of the key tradeoffs in proactive care, we propose a multi-class queueing system that explicitly models patients' deterioration and recovering behavior, and study the optimal scheduling policy for proactive care based on the model. Our analysis also provides insights on how the accuracy of the predictive information affects the prioritization of proactive care services.

While proactive service has long been considered in manufacturing settings where preventative maintenance effectively reduces the demand for future repair services (McCall 1965, Pierskalla and Voelker 1976), in healthcare operations management, there are very few works on proactive care (see Section 1.1 for a detailed review of some related work). Our modeling approach aims to provide a systematic way to capture the key tradeoffs in the limited resource environment: the potential benefit of curing patients with less resources versus the potential cost of delaying treatment for the more critical patients and generating more overall workload to the system.

We conduct analysis on both the long-run average performance and the transient performance, with the focus on developing structural insights on the optimal scheduling policy. The long-run average performance analysis provides guidance on scheduling proactive care when the system is in its normal state of operation. We also highlight the value of transient performance analysis and transient optimal control. This is because healthcare systems often operate in a highly non-stationary environment. A surge in demand due to random shocks such as disease outbreaks or mass casualty events can bring the system far from its normal state of operation. It is thus important to understand what is the most cost effective way to bring the system back to the normal workload. Our analysis provides insights into the merits of proactive care. Importantly, we are able to characterize settings where proactive care can be beneficial and others where it is better to focus all resources on the sickest patients. Our main contributions can be summarized as follows.

Queueing model with patient deterioration. We propose a Markovian multi-server queue with two patient classes: urgent and moderate. The key feature we incorporate is that a moderate patient who does not receive timely treatment may recover and leave or may deteriorate and

become an urgent class patient. If we assume there is a classifier (e.g. an early warning system) that classifies potentially risky patients into the moderate class, then the proportion of moderate patients who will actually deteriorate into the urgent class measures the true positive rate of the classifier. Our analysis, which builds on a deterministic fluid approximation of this queueing model, provides insights on how different model parameters affect the optimal scheduling policy for proactive care.

Equilibrium analysis. To minimize the long-run average cost for the fluid model, we show that the decision to prioritize the urgent class versus the moderate class is governed by what we refer to as the modified $c\mu/\theta$-rule. In particular, the modified $c\mu/\theta$-rule accounts for the deterioration dynamics in addition to the holding cost, service rate and abandonment rate. The exact expression of the modified $c\mu/\theta$-index lends itself to a very intuitive interpretation of which parameters – pre or post degradation – impact the costs. We also establish the long-run regularity under strict priority rules by characterizing the corresponding equilibrium system behavior. Interestingly, when strict priority is given to the urgent class, we identify cases where a bi-stability can arise. Bi-stability refers to the phenomenon where the system can converge to one of two equilibria, depending on its initial condition. Although one of the equilibria leads to good system performance, the other can result in very high costs.

Transient optimal control. To minimize the cumulative transient cost (until reaching the equilibrium point) for the fluid model, we show that the optimal policy may switch priority depending on the interplay between two indexes: the $c\mu$-index and the modified $c\mu/\theta$-index. In particular, it is optimal to schedule according to the modified $c\mu/\theta$-rule when the system state is far away from the equilibrium, and follow the $c\mu$-rule when the state gets close to the equilibrium. Furthermore, if the same class is prioritized by both the $c\mu$-rule and the modified $c\mu/\theta$-rule, then it is optimal to assign strict priority to this class throughout the transient time horizon. On the other hand, if one class is prioritized near the equilibrium and the other is prioritized far away from the equilibrium, then the optimal scheduling policy switches priority at most once along the trajectory. After characterizing the structure of the optimal scheduling policy, calculating the optimal policy curve where priority switches can be done efficiently. We then conduct numerical sensitivity analysis on the policy curve and quantify the effect of the 'prediction accuracy' on the optimal scheduling policy.

Our transient analysis also provides a paradigm for solving transient control problems in queues. In particular, the analysis can be summarized by three steps: i) Approximate the transient dynamics using a proper fluid model; ii) Derive the structure of the optimal scheduling policy for the fluid model. As the fluid model is a deterministic dynamical system, this step is done utilizing

Pontryagin's Mminimum Principle and special techniques to deal with state constraints; iii) Based on the structure of the optimal policy, solve a simpler version of the optimal control problem, i.e. solve for the optimal policy curve.

The rest of the paper is organized as follows. We conclude this section with a brief review of related literature (Section 1.1). The model and detailed problem formulation are introduced in Section 2. We derive the optimal scheduling policy to minimize the long-run average cost in Section 3, and the optimal scheduling policy to minimize the cumulative transient cost until reaching the equilibrium point in Section 4. Some generalizations to multi-class systems (more than two classes) are discussed in Section 5. Lastly, we provide some concluding remarks in Section 6.

1.1. Related Literature

Our work is mainly related to three streams of literature. From the problem context, our problem is related to proactive care for healthcare operations management, and scheduling in multi-class queues, especially queues with dynamic class types. From the methodology perspective, our work is related to transient queueing control. In what follows, we briefly review related works in these areas.

Proactive Service. Most existing literature on proactive care focuses on the optimal screening strategy. For example, Özekici and Pliska (1991) study the optimal scheduling of inspection in the context of screening for cancerous tumors. They take false positives into account but not the limited resource environment, i.e. they do not consider the externality each patient places on other patients. Örmeci et al. (2015) study the optimal scheduling of screening where the screening service shares resources with the more urgent diagnostic service. They model the benefit of screening through its effect on improving the 'environment'. Sun et al. (2017) study whether to perform triage under austere conditions, where triage occupies scarce resources but can provide more information on how to prioritize patients. Hu et al. (2018) take an empirical approach to examine the cost and benefit of proactively transferring 'risky' patients to the ICU. There are also works modeling proactive care as providers having advance information about customers' future service needs but do not model patient deterioration as we do. Examples include Xu and Chan (2016), Yom-Tov et al. (2018), Delana et al. (2019) and Cheng et al. (2019). Our work complements this literature by providing a general modeling framework that takes several key aspects of proactive care into account. These aspects include a limited resource environment, patient deterioration, patient recovery, different service needs and costs of waiting for different classes of patients. We also derive structural insights on the optimal scheduling policy for proactive care.

Optimal scheduling of multi-class queues. Our modeling approach falls into the category of multi-class queues. There is a growing literature on optimal scheduling of multi-class queues; see, for example, Mandelbaum and Stolyar (2004), Harrison and Zeevi (2004), Stolyar et al. (2004), and Puha and Ward (2019) for a recent review of works on scheduling multi-class queues with impatient customers. Due to the linear structure in system dynamics, in some cases, a simple index-based policy can be shown to be optimal. For example, the $c\mu$-rule is shown to be optimal for single server queue without abandonment (Cox and Smith 1991). The $c\mu/\theta$-rule is shown to be asymptotically optimal for multi-class queues with exponential patience time distribution in the many-server overloaded regime (Atar et al. 2011). We also note that due to the prohibitively large state-space and policy-space for these problems, approximation techniques are often employed to solve the corresponding Markov decision process and develop structural insights on the optimal policy, (e.g., Van Mieghem (1995), Tezcan and Dai (2010), Gurvich and Whitt (2010)).

The most relevant multi-class queueing models to ours are queues with dynamic class types. Sharing similar motivation to our work, Akan et al. (2012) models the wait list for donated organs as a multi-class overloaded queue. Disease evolution is captured by allowing patients to switch between different classes representing different health levels. Xie et al. (2017) conduct performance analysis for systems where delayed customers may renege the current queue and transfer to a higher-priority class. Cao and Xie (2016) derive the optimal scheduling policy for a single-server two-class model with holding and transferring costs. A modified $c\mu$-rule is shown to be optimal under some parameter regime. Down and Lewis (2010) study an N-model in which customers from the class with flexible servers ('low-priority') can be upgraded to the one with dedicated servers ('high-priority'). The $c\mu$-rule is proved to be optimal for scheduling flexible workers under certain conditions. Most of these works rely on exact or numerical analysis of the corresponding Markov decision process, where the analysis can become prohibitively challenging when more complexities are added to the model. In this paper, we adopt a fluid approximation approach, which borrows insights from the conventional heavy-traffic asymptotic analysis under the fluid scaling (Whitt 2002).

Transient Queueing Control. Analyzing transient queueing dynamics is often very challenging, not to mention optimizing over different control policies. Only a limited set of numerical and approximation techniques have been developed for the performance analysis of transient queues. These include inverting Laplace transforms (Abate and Whitt 1988, 2006), heavy-traffic asymptotics (Honnappa et al. 2015), etc. Our study uses the fluid approximation and employs tools from the optimal control theory for dynamical systems (Hartl et al. 1995) to derive the optimal transient

scheduling policy. The most relevant works to ours are Larrañaga et al. (2013) and Larrañaga (2015), where they consider a multi-class single-server queue with abandonment but no degradation. Aiming to minimize the cumulative transient holding cost for the fluid approximation, the authors show that the optimal policy may switch priority depending on the interplay between the $c\mu$-index and the $c\mu/\theta$-index. We note that adding the degradation component is a highly non-trivial extension due to the more complicated boundary behavior (when the state constraints are binding). Moreover, the optimal trajectories in our case cannot be characterized in closed form. We highlight that the analysis laid out in Section 4 substantially extends the framework for transient optimal control with state constraints; this approach may shed insights for other queueing control problems.

2. The Model

To explore the potential benefits of proactive care, we propose a Markovian two-class multi-server queueing system as depicted in Figure 1. Patients (customers/jobs) are defined by their need for service. We refer to Class 1 as the *urgent class*: those who need service with certainty. Focusing resource allocation to just these patients is a common approach in the health services literature. In this work, we also consider a *moderate class* (Class 2): those who currently do not need service, but are at risk of becoming urgent and in need of service. These Class 2 patients may deteriorate to Class 1 or they may 'self cure' and leave the system. *Proactive Care* (preventive care), i.e. providing service to Class 2 patients, can prevent Class 2 patients from becoming Class 1 patients.

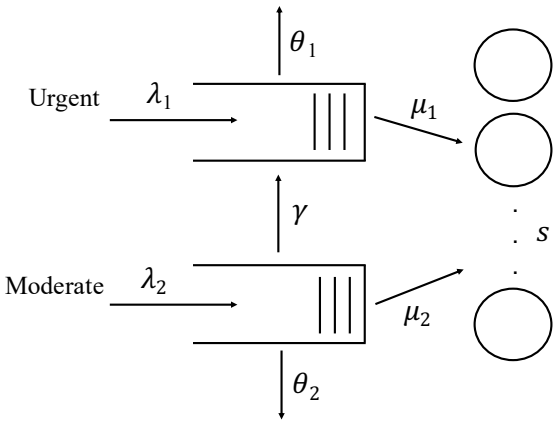
In our motivating healthcare context, Class 1 patients may correspond to patients who are physiologically unstable and in need of care in an Intensive Care Unit (ICU), while Class 2 patients may correspond to those on the general medical ward who are *at risk* of deteriorating. Those who are on the general medical ward, but are known to have no risk of needing ICU care, are outside of our modeling framework. Many patients on the general medical ward will never need ICU care, while others may decompensate and be transferred up to the ICU. With improving accuracy of early warning systems, proactive ICU admission *before* a patient is truly critical is becoming a reality (see, e.g., Hu et al. (2018)). What remains is to understand when and how such care should be utilized.

We consider a system with s identical servers (e.g. beds, doctors, and/or nurses), i.e. they offer the same quality of service. Class i patients, $i = 1, 2$, arrive to the system according to a time-homogeneous Poisson process with rate λ_i. Class 1 patients have independent and identically distributed service requirements following an exponential distribution with mean $1/\mu_1$. If a Class

1 patient waits too long in the (infinite-buffer) queue, it will 'abandon'. Its patience time is exponentially distributed with mean $1/\theta_1$. For Class 1 patients, one can interpret this abandonment as an undesirable event that could correspond to the patients being placed in an off-service unit, being transferred to another hospital, or even dying.

Class 2 patients can either be proactively served (i.e. before transitioning to Class 1), 'abandon' the system, or deteriorate into Class 1. Should the system administrator choose to provide proactive care to a Class 2 patient, its service time is exponentially distributed with mean $1/\mu_2$. Alternatively, if the Class 2 patient is still in the queue after an exponentially distributed patience time with mean $1/\theta_2$, it will 'abandon'. For these moderate patients, the abandonment can be a good thing where the patient is no longer at risk for deterioration, i.e. the patient self-cures. Finally, a Class 2 patient will deteriorate into the urgent type (Class 1) according to an exponentially distributed clock with rate γ while waiting in queue.

Figure 1 Two-class queue



We make a few assumptions on the model parameters to capture the healthcare setting we are interested in analyzing. To reflect the fact that Class 1 patients are more severe, we assume that Class 1 patients have a longer average service requirement than Class 2 patients, i.e., $\mu_1 < \mu_2$.

We next provide a useful interpretation for $\phi := \gamma/(\theta_2 + \gamma)$. Note that if no proactive care is provided to Class 2 patients, $\gamma/(\theta_2+\gamma)$ of them will deteriorate into the urgent class. Suppose Class 2 patients are identified via a classifier that determines patients who are 'at risk' of deteriorating (e.g. Escobar et al. (2012)), then $\gamma/(\theta_2+\gamma)$ can be interpreted as the true positive rate of this classifier. That is, it measures the accuracy of the classifier. If we know with certainty that Class 2 patients will eventually deteriorate into Class 1 patients, then $\theta_2 \to 0$ and $\gamma/(\theta_2+\gamma) \to 1$.

To understand the key tradeoff we are trying to capture with this model, we start by discussing the extreme case where $\theta_1 = 0$. In this case, if no service is provided to Class 2 patients, each Class

2 patient generates an average workload of $\gamma/(\mu_1(\theta_2+\gamma))$ to the system. This is because $\gamma/(\theta_2+\gamma)$ of the Class 2 patients will deteriorate into Class 1 and all Class 1 patients must be served. On the other hand, if we can provide proactive care to all Class 2 patients, then each Class 2 patient will generate an average workload of $1/\mu_2$. The magnitude of $\gamma/(\theta_2+\gamma)$ impacts whether it may be more or less beneficial, from a workload perspective, to provide preventive care to Class 2 patients. Of course, the actual problem we are facing is more complicated than minimizing the system workload. In particular, the different waiting, abandonment and/or deterioration costs incurred by the two classes can also have a substantial impact on the optimal scheduling policy.

Let $X_i(t)$ denote the number of Class i patients in the system at time t, $t \geq 0$. We denote by $Z_i(t)$ the number of servers assigned to Class i patients, and by $Q_i(t)$ the queue length of Class i at time t. Clearly, $Z_1(t) + Z_2(t) \leq s$ and $X_i(t) - Z_i(t) = Q_i(t) \geq 0$ for $i = 1, 2$. We also write $X(t) = (X_1(t), X_2(t))$, $Z(t) = (Z_1(t), Z_2(t))$, and $Q(t) = (Q_1(t), Q(t))$. Note that the state of the system at time t can be described by $(X(t), Q(t))$. A scheduling policy Π is defined as a rule for allocating servers to customers, i.e. Z_i's are the control variables. We consider Markovian policies under which the server allocations are made based on the current state of (X, Q) only. In particular, the policy is non-anticipating. Under this class of scheduling policies, $\{(X(t), Q(t)) : t \geq 0\}$ forms a Markov process.

As the processes $\{(X(t), Q(t)) : t \geq 0\}$ actually depends on the scheduling policy Π, we can more explicitly mark the dependence by writing the stochastic process as $\{(X^\Pi(t), Q^\Pi(t)) : t \geq 0\}$. We also denote $R_i^\Pi(t)$ as the cumulative number of the patients that have abandoned Class i queue by time t, and $\Gamma^\Pi(t)$ as the cumulative number of Class 2 patients that have degraded into Class 1 by time t. For what follows, we shall drop the superscript Π when it can be understood by context.

We incur costs for all patients who wait, urgent patients who abandon, and/or moderate patients who deteriorate. In particular, we denote h_i as the holding cost per unit time for Class i patients waiting in queue. In addition, each abandonment of Class 1 patient is associated with a fixed cost r_1, and each transition of Class 2 patient into Class 1 has a fixed cost ν. Our goal is to minimize the aggregated cost incurred, for example,

$$\mathbb{E}\left[\int_0^T h_1 Q_1(t) + h_2 Q_2(t) dt + r_1 R_1(T) + \nu \Gamma(T)\right] \tag{1}$$

Note that under the Markovian modeling assumption, we have

$$\mathbb{E}\left[R_1(T)\right] = \theta_i \mathbb{E}\left[\int_0^T Q_1(t) dt\right] \text{ and } \mathbb{E}\left[\Gamma(T)\right] = \gamma \mathbb{E}\left[\int_0^T Q_2(s) ds\right].$$

Thus, (1) can be equivalently written as

$$\mathbb{E}\left[\int_0^T c_1 Q_1(t) + c_2 Q_2(t) dt\right], \text{ where } c_1 = h_1 + r_1\theta_1 \text{ and } c_2 = h_2 + \nu\gamma.$$

This implies that we can incorporate the abandonment cost and the deterioration cost into the holding cost. In what follows, we shall use c_1 and c_2 to denote the 'generalized' holding cost.

In this paper, we focus on two cost measures. One is the long-run average cost; the other is the accumulated transient cost. The two cost formulations have different focuses and are both very relevant in practice. The long-run average cost formulation mainly concerns minimizing the cost when the system is in its 'normal' state of operation. When random shocks (e.g. a disease outbreak or a mass casualty event) bring the system far from its 'normal' state of operation, the transient cost formulation focuses on minimizing the cost incurred to bring the system back to normal. More precisely, the **long-run average cost minimization problem** is

$$\min_{\Pi} \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\int_0^T \left(c_1 Q_1^{\Pi}(t) + c_2 Q_2^{\Pi}(t)\right) dt\right]. \tag{S1}$$

For the **accumulated transient cost minimization problem**, we define

$$\mathcal{T}(\epsilon) := \inf\left\{t \geq 0 : Q_1(t) + Q_2(t) \leq \epsilon\right\},$$

for some fixed ϵ. That is, $\mathcal{T}(\epsilon)$ is the time until the total queue length is sufficiently small. We assume that we have ample capacity s such that $\mathbb{E}[\mathcal{T}(\epsilon)] < \infty$ for any fixed initial state $(X(0), Q(0)) = (x_0, q_0)$. Then the transient optimization problem can be written as

$$\min_{\Pi} \mathbb{E}\left[\int_0^{\mathcal{T}(\epsilon)} \left(c_1 Q_1^{\Pi}(t) + c_2 Q_2^{\Pi}(t)\right) dt\right]. \tag{S2}$$

These cost minimization problems are Markov decision processes (MDP). Due to the large (infinite) state-space and policy-space, they are prohibitively hard to solve from a computational standpoint. Even if we solve it numerically, limited insights about the optimal policy can be gained. Various approximation techniques have been developed in the literature to solve large-scale MDPs. With the goal of gaining structural insights on the optimal scheduling policy, we employ a fluid approximation approach; similar method has been used in, for example, Perry and Whitt (2009).

2.1. The Fluid Model

To construct the fluid model, we replace the stochastic arrival, service, abandonment and deterioration processes by their corresponding deterministic flow rates. We use the lowercase q to denote the

fluid queue length process, and a fluid scheduling policy π specifies the service capacity allocation process (z_1, z_2). Under π, the fluid dynamics takes the form:

$$\begin{aligned} dq_1(t) &= \lambda_1 - z_1(t)\mu_1 - \theta_1 q_1(t) + \gamma q_2(t) \\ dq_2(t) &= \lambda_2 - z_2(t)\mu_2 - \theta_2 q_2(t) - \gamma q_2(t). \end{aligned} \tag{2}$$

Let $\mathscr{F}$ denote the set of fluid admissible scheduling policies. We say that a policy belongs to $\mathscr{F}$ if the server allocation only depends on the current state of the system (Markovian), and it satisfies the following constraints:

$$\begin{aligned} & z_i(t) \geq 0, \quad i = 1, 2, \ t \geq 0 \\ & z_1(t) + z_2(t) \leq s, \quad t \geq 0 \\ & dq_i(t) \geq 0 \text{ whenever } q_i(t) = 0, \quad i = 1, 2, \ t \geq 0. \end{aligned} \tag{3}$$

The first and second constraints in (3) require that a non-negative amount of service capacity is assigned to each class, and the total amount of allocated resource does not exceed service capacity. The third constraint guarantees that the resulting queue length process q_i is non-negative for all $t \geq 0$. Note that the queue length process $\{q(t) : t \geq 0\}$ actually depends on the scheduling policy π. We can more explicitly mark the dependence by writing it as $\{q^\pi(t) : t \geq 0\}$. To keep the notation concise, we shall drop the superscript when it can be understood from context.

We comment that the fluid dynamics capture the mean dynamics of the stochastic system well, as we will demonstrate in Sections 3.3 and 4.4. In addition, this type of fluid model often arises in the literature as the functional law of large numbers limit for a sequence of properly scaled stochastic systems under the conventional heavy traffic limit (Iglehart and Whitt 1970, Reed and Ward 2008). In this limiting regime, we scale up the arrival rate and the service rate while we scale down the space (Alternatively, we can scale up time while we scale down the abandonment rate and the space). The number of servers is held fixed*. The conventional heavy traffic scaling works well when approximating systems with a relative small number of servers, and is thus very relevant for healthcare applications. Note that in many healthcare systems, e.g. ICUs and EDs, the number of servers is typically not very large, e.g. $5 - 20$.

2.2. Problem Formulation

In this section, we introduce the fluid counterparts of the stochastic cost minimization problems.

Fluid long-run average cost optimization problem:

$$\min_{\pi \in \mathscr{F}} \limsup_{T \to \infty} \frac{1}{T} \int_0^T \left(c_1 q_1^\pi(t) + c_2 q_2^\pi(t) \right) dt \tag{F1}$$

* In particular, we do not scale up the number of servers as in the many-server heavy traffic regime.

For the transient optimization problem, let $\tau := \inf\{t \geq 0 : q_1(t) + q_2(t) = 0\}$, which is the first time when both fluid queues reduce to 0. We assume that there is ample capacity s such that for any $q(0) = q_0$, $\tau < \infty$ (The precise condition will be specified in Section 4).

Fluid transient optimization problem:

$$\min_{\pi \in \mathscr{F}} \int_0^\tau (c_1 q_1^\pi(t) + c_2 q_2^\pi(t))\, dt, \tag{F2}$$

Our analysis relies on understanding the long-run regularity of the fluid model. We thus provide the following definitions.

Definition 1 (Equilibrium) *Consider the autonomous dynamical system $dq(t) = f(q(t))$ with $q(0) = q_0$. Suppose f has an equilibrium point q_e, i.e. $f(q_e) = 0$. Let $||\cdot||$ be the Euclidean norm in $\mathbb{R}_2$. Then*

(1) q_e is **locally asymptotically stable** *if there exists $\delta > 0$, such that if $||q(0) - q_e|| < \delta$, then $\lim_{t\to\infty} ||q(t) - q_e|| = 0$.*

(2) q_e is **globally asymptotically stable** *if for any initial condition $q(0)$, $\lim_{t\to\infty} ||q(t) - q_e|| = 0$.*

We shall start by solving the long-run average cost minimization problem (F2) in Section 3. The analysis also provides valuable insights on the long-run regularity of the fluid dynamical systems. We then solve the transient cost minimization problem (F2) in Section 4.

3. Optimal Long-Run Scheduling Policy

In this section, we solve the fluid long-run average cost minimization problem. To ensure system stability for any arrival and service rates, we impose the following assumptions on the abandonment and deterioration rates.

Assumption 1 *$\theta_1 > 0$ and $\theta_2 + \gamma > 0$.*

The cost-minimization problem can be explicitly written as

$$\begin{aligned}
\min_{z_1, z_2} \quad & \limsup_{T\to\infty} \frac{1}{T} \int_0^T (c_1 q_1(t) + c_2 q_2(t))\, dt \\
s.t. \quad & dq_1(t) = \lambda_1 - \mu_1 z_1(t) - \theta_1 q_1(t) + \gamma q_2(t) \\
& dq_2(t) = \lambda_2 - \mu_2 z_2(t) - (\theta_2 + \gamma) q_2(t) \\
& z_1(t) + z_2(t) \leq s, \quad t \geq 0 \\
& z_1(t), z_2(t), q_1(t), q_2(t) \geq 0, \quad t \geq 0.
\end{aligned}$$

This is an infinite dimensional linear program (LP). We first make a few observations that allow us to reformulate the problem as a finite dimensional LP. These observations will be made rigorous

in Theorem 1. If the fluid dynamical system converges to an equilibrium point as $t \to \infty$, then minimizing the long-run average cost can be reformulated as finding the optimal equilibrium point. In particular, we have the following alternative problem formulation.

$$\begin{aligned} \min_{z_1^e, z_2^e} \quad & c_1 q_1^e + c_2 q_2^e \\ s.t. \quad & \lambda_1 + \gamma q_2^e = \mu_1 z_1^e + \theta_1 q_1^e \\ & \lambda_2 = \mu_2 z_2^e + (\theta_2 + \gamma) q_2^e \\ & z_1^e + z_2^2 \leq s \\ & z_1^e, z_2^e, q_1^e, q_2^e \geq 0. \end{aligned} \tag{4}$$

Note that the first two constraints in (4) characterize the equilibrium points: rate-in equals rate-out. By a rearrangement of (4), we have an equivalent optimization problem:

$$\begin{aligned} \max_{z_1^e, z_2^e} \quad & \frac{c_1}{\theta_1}\mu_1 z_1^e + \left(\frac{c_2}{\theta_2 + \gamma}\mu_2 + \frac{\frac{\gamma}{(\theta_2+\gamma)}c_1}{\theta_1}\mu_2 \right) z_2^e \\ s.t. \quad & z_1^e + z_2^e \leq s \\ & \lambda_1 - \mu_1 z_1^e + \frac{\gamma}{\theta_2 + \gamma}(\lambda_2 - \mu_2 z_2^e) \geq 0 \\ & \lambda_2 - \mu_2 z_2^e \geq 0 \\ & z_1^e, z_2^e \geq 0. \end{aligned} \tag{5}$$

It is straightforward to see that the optimal solution to (5) is to assign the maximum feasible value to the z_i^e with a larger weight in the objective function. Motivated by this observation, we define the modified $c\mu/\theta$-index as follows: **The modified $c\mu/\theta$ index for Class 1** is $c_1\mu_1/\theta_1$, and **the modified $c\mu/\theta$ index for Class 2** is

$$\frac{c_2}{\theta_2 + \gamma}\mu_2 + \frac{\frac{\gamma}{(\theta_2+\gamma)}c_1}{\theta_1}\mu_2. \tag{6}$$

In interpreting the index (6), we note that the first term corresponds to the standard $c\mu/\theta$ from the literature (Atar et al. 2010), since $\theta_2 + \gamma$ is the rate at which Class 2 jobs leave the queue via abandonment or deterioration. The $c\mu/\theta$ index has the nice interpretation which captures the cost rate, c, the service rate, μ, and the abandonment rate θ. Interestingly, the second term can be similarly viewed when considering the Class 2 patients who may deteriorate. In expectation, these patients induce a cost rate of $\gamma c_1/(\theta_2 + \gamma)$, since the probability of deteriorating to Class 1 is $\gamma/(\theta_2 + \gamma)$. They abandon at rate θ_1, but can be 'served' while as Class 2 at rate μ_2.

Formally, we have the following theorem characterizing the optimal scheduling policy.

Theorem 1 *Under Assumption 1, giving strict priority to the class with a higher modified $c\mu/\theta$-index minimizes the long-run average cost* (F1)*. That is, if*

$$\frac{c_2}{\theta_2+\gamma}\mu_2+\frac{\frac{\gamma}{(\theta_2+\gamma)}c_1}{\theta_1}\mu_2\le\frac{c_1}{\theta_1}\mu_1,$$

then it is optimal to give strict priority to Class 1. Otherwise, it is optimal to give strict priority to Class 2.

To prove Theorem 1, we need to first check if the fluid dynamical system converges to the desired equilibrium point under the strict priority rule implied by the modified $c\mu/\theta$-index. These convergence analyses are interesting in their own right, as they reveal important characteristics of the system dynamics. In Section 3.1, we study the fluid dynamics under strict priority to Class 1, and in Section 3.2, we study the fluid dynamics under strict priority to Class 2. The goal is to establish the long-run regularity of the fluid model.

3.1. Strict Priority to Class 1 Patients

Under strict priority rule to Class 1 patients, when $q_1(t)>0$, we will assign all the capacity to Class 1. When $q_1(t)=0$, we will assign to Class 1 the minimum amount of capacity to maintain its queue at zero if there is enough capacity; otherwise, we will assign all the capacity to Class 1. In particular, the system dynamics are characterized as follows:

(i) If $q_1(t)>0$,

$$dq_1(t)=\lambda_1-\mu_1 s-\theta_1 q_1(t)+\gamma q_2(t),\quad dq_2(t)=\lambda_2-\theta_2 q_2(t)-\gamma q_2(t); \tag{7}$$

(ii) If $q_1(t)=0, q_2(t)>0$,

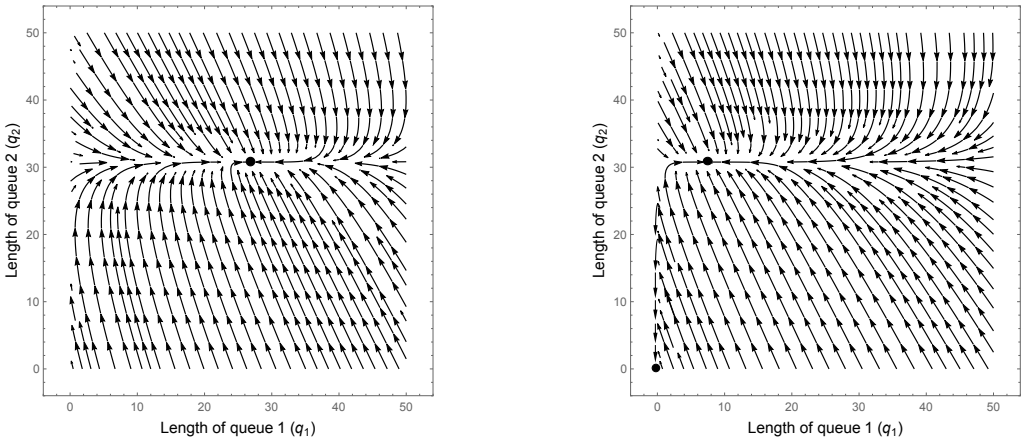
$$\begin{aligned} dq_1(t)&=\lambda_1-\mu_1\left(\frac{\lambda_1+\gamma q_2(t)}{\mu_1}\wedge s\right)+\gamma q_2(t),\\ dq_2(t)&=\lambda_2-\mu_2\left(s-\frac{\lambda_1+\gamma q_2(t)}{\mu_1}\right)^+-\theta_2 q_2(t)-\gamma q_2(t); \end{aligned} \tag{8}$$

(iii) If $q_1(t)=0, q_2(t)=0$,

$$dq_1(t)=\lambda_1-\mu_1\left(\frac{\lambda_1}{\mu_1}\wedge s\right),\quad dq_2(t)=\lambda_2-\mu_2\left(\left(s-\frac{\lambda_1}{\mu_1}\right)^+\wedge\frac{\lambda_2}{\mu_2}\right). \tag{9}$$

Figure 2 provides two examples of vector fields for the fluid model with different parameters. We observe that the fluid model enjoys a certain long-run regularity in the sense that the fluid trajectory will converge to an asymptotically stable equilibrium point represented by a black dot. This is well-expected, as we have positive abandonment rates. However, as can be seen in Figure 2(b), there can be more than one equilibrium point. In this case, which point the fluid trajectory converges to depends on its initial condition. Using a Lyapunov argument, Theorem 2 characterizes the long-run regularity of the fluid dynamical system under strict priority to Class 1.

Figure 2 **Globally asymptotically stable equilibrium (a) V.S. locally asymptotically stable equilibrium (b)**
((a):$\lambda_1 = 10, \lambda_2 = 20, \mu_1 = 1.5, \mu_2 = 3, \gamma = 0.25, \theta_1 = 0.1, \theta_2 = 0.4$
(b):$\lambda_1 = 10, \lambda_2 = 20, \mu_1 = 1, \mu_2 = 3, \gamma = 0.25, \theta_1 = 0.1, \theta_2 = 0.4$**)**



(a) Case I: $s = 10$ (b) Case II: $s = 17$

Theorem 2 *Under Assumption 1, for the dynamical system* (7) - (9)*,*

Case I. *When* $\mu_1 > \frac{\gamma}{\theta_2+\gamma}\mu_2$*,*

Ia If $\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} \leq s$**,** *the system has a globally asymptotically stable equilibrium at*

$$q_1^e = 0, \quad q_2^e = 0.$$

Ib If $\frac{\lambda_1}{\mu_1} + \frac{\gamma}{\theta_2+\gamma}\frac{\lambda_2}{\mu_1} \leq s < \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2}$**,** *the system has a globally asymptotically stable equilibrium at*

$$q_1^e = 0, \quad q_2^e = \frac{\mu_1\mu_2\left(\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} - s\right)}{(\theta_2+\gamma)\mu_1 - \gamma\mu_2} > 0.$$

Ic If $s < \frac{\lambda_1}{\mu_1} + \frac{\gamma}{\theta_2+\gamma}\frac{\lambda_2}{\mu_1}$**.** *the system has a globally asymptotically stable equilibrium at*

$$q_1^e = \frac{\lambda_1 + \frac{\gamma}{\theta_2+\gamma}\lambda_2 - s\mu_1}{\theta_1} > 0, \quad q_2^e = \frac{\lambda_2}{\theta_2+\gamma} > 0.$$

Case II. *When* $\mu_1 < \frac{\gamma}{\theta_2+\gamma}\mu_2$*,*

IIa If $\frac{\lambda_1}{\mu_1} + \frac{\gamma}{\theta_2+\gamma}\frac{\lambda_2}{\mu_1} < s$**,** *the system has a globally asymptotically stable equilibrium at*

$$q_1^e = 0, \quad q_2^e = 0.$$

IIb If $\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} < s \leq \frac{\lambda_1}{\mu_1} + \frac{\gamma}{\theta_2+\gamma}\frac{\lambda_2}{\mu_1}$**,** *the system has two locally asymptotically stable equilibria*

$$q_{11}^e = 0, \quad q_{21}^e = 0; \quad \textit{and} \quad q_{12}^e = \frac{\lambda_1 + \frac{\gamma}{\theta_2+\gamma}\lambda_2 - s\mu_1}{\theta_1} \geq 0, \quad q_{22}^e = \frac{\lambda_2}{\theta_2+\gamma} > 0.$$

IIc If $s = \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2}$**,** *the system has an equilibrium at* $(q_{11}^e, q_{21}^e) = (0,0)$ *and a locally asymptotically stable equilibrium at*

$$q_{12}^e = \frac{\lambda_1 + \frac{\gamma}{\theta_2+\gamma}\lambda_2 - s\mu_1}{\theta_1} > 0, \quad q_{22}^e = \frac{\lambda_2}{\theta_2+\gamma}.$$

IId If $s < \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2}$**,** *the system has a globally asymptotically stable equilibrium at*

$$q_1^e = \frac{\lambda_1 + \frac{\gamma}{\theta_2+\gamma}\lambda_2 - s\mu_1}{\theta_1} > 0, \quad q_2^e = \frac{\lambda_2}{\theta_2+\gamma}.$$

Remark 1 *We note that when $\mu_1 = \gamma\mu_2/(\theta_2+\gamma)$, the system can have uncountably many equilibrium points. In particular, for $s = \lambda/\mu_1 + \lambda_2/\mu_2$, any (q_1^e, q_2^e) satisfying $q_1^e = 0$ and $(\lambda_1 + \gamma q_2^e)/\mu_1 < s$ is an equilibrium point. We do not consider this parameter regime, i.e. $\mu_1 = \gamma\mu_2/(\theta_2+\gamma)$, in this paper.*

We make a few observations from Theorem 2. First note that when $\mu_1 < \frac{\gamma}{\theta_2+\gamma}\mu_2$ (Case II), the modified $c\mu/\theta$-index of Class 2 is larger than that of Class 1, irrespective of c_2. From the workload perspective, Case II indicates that it is more efficient to serve the Class 2 patients before they deteriorate, i.e.

$$\frac{1}{\mu_2} < \frac{\gamma}{\gamma+\theta}\frac{1}{\mu_1}.$$

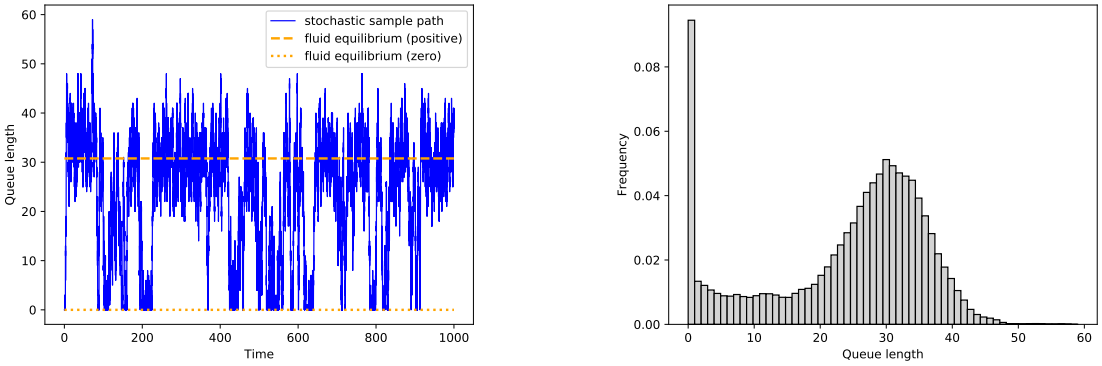
In this case, if we prioritize Class 1 patients, the system can have two equilibrium points when the service capacity is in the critical region $\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} < s \leq \frac{\lambda_1}{\mu_1} + \frac{\gamma}{\theta_2+\gamma}\frac{\lambda_2}{\mu_1}$ (Case IIb). This phenomenon is referred to as bi-stability. The bi-stability arises because if we delay treatment for Class 2 patients, they will end up generating more workload on average when they deteriorate into Class 1 patients. In Case IIb, even though we have enough capacity to serve both classes when the care is provided in a timely manner, i.e. $\lambda_1/\mu_1 + \lambda_2/\mu_2 < s$, we do not have enough capacity to serve all the patients when treatments for Class 2 are delayed, i.e.

$$s \leq \frac{\lambda_1}{\mu_1} + \frac{\gamma}{\theta_2+\gamma}\frac{\lambda_2}{\mu_1}.$$

In the fluid system, when bi-stability arises, the system can converge to either one of the equilibrium points depending its initial value. This implies that for the corresponding stochastic system, the queue length process will fluctuate around one equilibrium point for a while before transitioning to the region around the other equilibrium point. In particular, Figure 3(a) shows a typical sample path of the stochastic queue length process, i.e. we plot $Q_2(t)$ for $t \in [0, 1000]$. When we plot the histogram of the stochastic queue length process (Figure 3(b)), we observe that it follows bi-modal distribution where the two peaks are around the two fluid equilibria.

We also note that in Case IIb, one of the equilibrium points leads to zero holding cost while the other equilibrium point has positive queue length for at least one class. Ideally, we want to avoid the 'bad' equilibrium point. However, when strict priority is given to Class 1, we have no control over which equilibrium point the stochastic system will fluctuate around. To avoid the bad equilibrium point, we either have to switch priority to Class 2 or impose some alternative intervention, such as admission control, when the system is moving towards the bad equilibrium point.

Figure 3 **Bi-stability in the stochastic system**
$(s=17, \lambda_1=10, \lambda_2=20, \mu_1=1, \mu_2=3, \gamma=0.25, \theta_1=0.1, \theta_2=0.4, x_1(0)=6, x_2(0)=6)$



(a) Sample path of $Q_2(t)$ (b) Histogram of $Q_2(t)$

3.2. Strict Priority to Class 2 Patients

Under strict priority rule to Class 2 patients, when $q_2(t) > 0$, we will assign all capacity to Class 2. When $q_2(t) = 0$, we will assign to Class 2 the minimum amount of capacity to maintain its queue at zero if there is enough capacity; otherwise, we will assign all the capacity to Class 2. In particular, the system dynamics are characterized as follows:

(i) If $q_2(t) > 0$,

$$dq_1(t) = \lambda_1 - \theta_1 q_1(t) + \gamma q_2(t), \quad dq_2(t) = \lambda_2 - \mu_2 s - \theta_2 q_2(t) - \gamma q_2(t); \tag{10}$$

(ii) If $q_2(t) = 0, q_1(t) > 0$,

$$dq_1(t) = \lambda_1 - \mu_1 \left(s - \frac{\lambda_2}{\mu_2}\right)^+ - \theta_1 q_1(t), \quad dq_2(t) = \lambda_2 - \mu_2 \left(\frac{\lambda_2}{\mu_2} \wedge s\right); \tag{11}$$

(iii) If $q_1(t) = 0, q_2(t) = 0$,

$$dq_1(t) = \lambda_1 - \mu_1 \left(\left(s - \frac{\lambda_2}{\mu_2}\right)^+ \wedge \frac{\lambda_1}{\mu_1}\right), \quad dq_2(t) = \lambda_2 - \mu_2 \left(\frac{\lambda_2}{\mu_2} \wedge s\right). \tag{12}$$

Following a similar Lyapunov argument as in Theorem 2, we can establish the following long-run regularity result for the fluid model under strict priority to Class 2.

Theorem 3 *Under Assumption 1, for the dynamical system* (10) - (12),

Case A. If $\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} \le s$**,** *the system has a globally asymptotically stable equilibrium at*

$$q_1^e = 0, \quad q_2^e = 0.$$

Case B. If $\frac{\lambda_2}{\mu_2} \le s < \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2}$**,** *the system has a globally asymptotically stable equilibrium at*

$$q_1^e = \frac{\mu_1}{\theta_1}\left(\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} - s\right) > 0, \quad q_2^e = 0.$$

Case C. If $s < \frac{\lambda_2}{\mu_2}$**,** *the system has a globally asymptotically stable equilibrium at*

$$q_1^e = \frac{\lambda_1 + \frac{\gamma}{\theta_2+\gamma}(\lambda_2 - s\mu_2)}{\theta_1} > 0, \quad q_2^e = \frac{\lambda_2 - s\mu_2}{\theta_2 + \gamma} > 0.$$

Note that under strict priority to Class 2, there is always a globally asymptotic stable equilibrium; there is never a bi-stability, even in the case where $\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} < s \leq \frac{\lambda_1}{\mu_1} + \frac{\gamma}{\theta_2+\gamma}\frac{\lambda_2}{\mu_1}$. This is because giving strict priority to Class 2 will maximize the number of Class 2 patients that can be served 'efficiently' at rate μ_2, rather than serving them after they deteriorate to Class 1. In contrast, when giving strict priority to Class 1, the ability to serve Class 2 patients at the efficient rate depends on when all Class 1 jobs are completed and this can further be determined by how many patients are in the system at time 0. Thus, the initial state of the system has a large impact on which equilibrium the system will converge to when giving strict priority to Class 1, but not when giving strict priority to Class 2.

We next compare the equilibrium points in Theorem 3 to those in Theorem 2. Denote P_1 as strict priority to Class 1, and P_2 as strict priority to Class 2. We also define the nominal load, L_N, and the deterioration-based load, L_D, as:

$$L_N := \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} \quad \text{and} \quad L_D := \frac{\lambda_1}{\mu_1} + \frac{\lambda_2\gamma}{\mu_1(\theta_2+\gamma)}.$$

When there are abundant resources, we can achieve zero queues in equilibrium. Specifically, when $\mu_1 > \gamma\mu_2/(\theta_2+\gamma)$, if $L_N < s$, P_1 and P_2 both have zero queues in equilibrium. When $\mu_1 < \gamma\mu_2/(\theta_2+\gamma)$, if $L_D < s$, P_1 and P_2 both have zero queues in equilibrium. In these parameter regions, P_1 and P_2 are both optimal. A parameter region of particular interests is when $\mu_1 < \gamma\mu_2/(\theta_2+\gamma)$ and $L_N < s < L_D$. In this case, P_2 leads to zero queues in equilibrium while P_1 leads to bi-stability. This implies that we would prefer P_2 in this parameter region.

When capacity is scarce, the two priority rules can lead to very different equilibrium queues. P_1 always leads to a larger Class 2 queue than P_2. However, P_1 may not always lead to a smaller Class 1 queue. In particular, when $s < L_D$ and $\mu_1 < \gamma\mu_2/(\theta_2+\gamma)$, P_2 will lead to both a smaller Class 1 queue and a smaller Class 2 queue than P_1 in equilibrium.

3.3. Numerical experiments

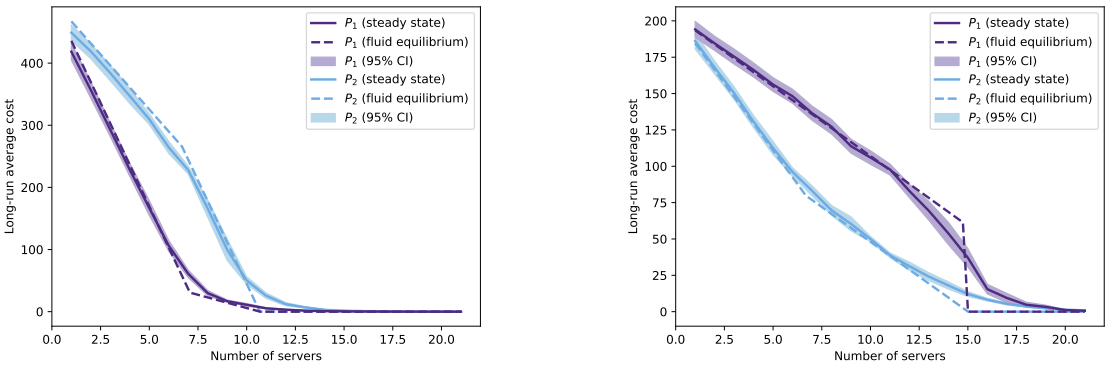
In this section, we conduct some numerical experiments to compare the two priority rules. This serves as a sanity check for Theorem 1. More importantly, we also compare the long-run average fluid cost to the long-run average cost of the corresponding stochastic system.

We start with the case where $\mu_1 > \gamma\mu_2/(\theta_2+\gamma)$. Figure 4 plots the long-run average costs for different values of s. The fluid costs are plotted in dashed lines while the costs for the stochastic

system are plotted in solid lines. As the long-run average costs for the stochastic systems are estimated using simulation, we also provide the corresponding 95% confidence interval for our estimates. Figure 4(a) illustrates the scenario where the modified $c\mu/\theta$-index suggests prioritizing Class 1 patients, while Figure 4(b) has the modified $c\mu/\theta$-index suggesting prioritizing Class 2 patients. We first note that the long-run average fluid cost approximates the long-run average cost of the stochastic system reasonably well. Second, we observe that following the modified $c\mu/\theta$-rule – prioritizing the class with a larger modified $c\mu/\theta$-index – indeed leads to a lower cost. Thus, while the cost of fluid system may deviate from that of the stochastic systems, the resulting policy recommendations are consistent. Lastly, we observe that when s is large enough, P_1 and P_2 lead to very similar costs, i.e. almost 0 costs.

We also consider the case where $\mu_1 < \gamma\mu_2/(\theta_2+\gamma)$. In this case, the modified $c\mu/\theta$-index suggests prioritizing Class 2 patients. Figure 5 plots the long-run average costs for different values of s. We observe that, as expected, P_2 leads to a lower cost for both the fluid system and the corresponding stochastic system. We also observe that for some values of s, P_1 leads to two different long-run average fluid costs, which correspond to the two different equilibria in Theorem 2 Case IIb. In this case, the long-run average cost of the corresponding stochastic system is approximately a weighted average of the two fluid equilibrium costs.

Figure 4 **Optimal scheduling policy in Case I with** $\mu_1 > \frac{\gamma}{\gamma+\theta_2}\mu_2$
((a): $\lambda_1 = 10, \lambda_2 = 20, \mu_1 = 2.5, \mu_2 = 3, \gamma = 0.25, \theta_1 = 0.15, \theta_2 = 0.4, c_1 = 4, c_2 = 1$
(b): $\lambda_1 = 10, \lambda_2 = 20, \mu_1 = 1.2, \mu_2 = 3, \gamma = 0.25, \theta_1 = 0.5, \theta_2 = 0.4, c_1 = 4, c_2 = 2$**)**



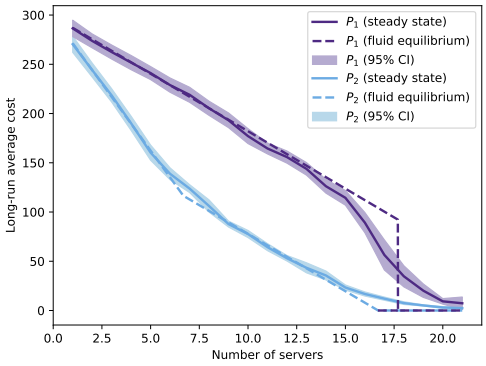
(a) $\frac{c_1}{\theta_1}\mu_1 > \frac{c_2}{\theta_2+\gamma}\mu_2 + \frac{\frac{\gamma}{(\theta_2+\gamma)}c_1}{\theta_1}\mu_2$

(b) $\frac{c_1}{\theta_1}\mu_1 < \frac{c_2}{\theta_2+\gamma}\mu_2 + \frac{\frac{\gamma}{(\theta_2+\gamma)}c_1}{\theta_1}\mu_2$

4. Optimal Transient Scheduling Policy

Healthcare systems often operate in highly non-stationary environments. Random shocks like disease outbreaks or mass casualty events can push the system far from its 'normal', i.e. equilibrium,

Figure 5 **Optimal scheduling policy in Case II with $\mu_1 < \frac{\gamma}{\gamma+\theta_2}\mu_2$**
($\lambda_1 = 10, \lambda_2 = 20, \mu_1 = 1, \mu_2 = 3, \gamma = 0.25, \theta_1 = 0.3, \theta_2 = 0.4, c_1 = 3.5, c_2 = 3$)



behavior. When such a congestion event happens, the key question we wish to address is how to bring the system back to its normal state of operation in the most cost-effective way. In this section, we study the transient optimal control problem (F2) to address this problem. In particular, we derive the optimal scheduling policy to help the system recover from demand shocks.

To develop insights into how to manage our system after a large shock, we study the transient analysis of the system where we assume $q(0) = q_0 > 0$, but we have abundant capacity to bring the fluid queues to zero in a finite amount of time under some admissible control. In particular, we make the following assumption on the capacity s:

Assumption 2 $s > \lambda_1/\mu_1 + \lambda_2/\mu_2$.

Recall that $\tau = \inf\{t \geq 0 : q_1(t) + q_2(t) = 0\}$ is the first time both of the fluid queues are emptied. Based on the analysis in Section 2.1, Assumption 2 implies that there exists a scheduling policy π, under which, for any $q(0) = q_0 > 0$, $\tau < \infty$.

We next introduce a few more notations to simplify the presentation of the problem. From the fluid dynamics (2), we define $f(q,z) = (f_1(q,z), f_2(q,z))$ where $f_1(q,z) = \lambda_1 - z_1\mu_1 - \theta_1 q_1 + \gamma q_2$ and $f_2(q,z) = \lambda_2 - z_2\mu_2 - \theta_2 q_2 - \gamma q_2$. From the constraints on the admissible controls (3), we define $g(q) = (g_1(q), g_2(q))$, where $g_i(q) = -q_i$, for $i = 1, 2$, and $h(z) = (h_1(z), h_2(z), h_3(z))$, where $h_1(z) = -z_1$, $h_2(z) = -z_2$ and $h_3(z) = z_1 + z_2 - s$. We also define $F(q) = c_1 q_1 + c_2 q_2$. Then the transient optimal control problem can be explicitly written as:

$$\begin{aligned} \min_{z} \quad & \int_0^\tau F\left(q(t)\right) dt, \\ s.t. \quad & dq(t) = f\left(q(t), z(t)\right), \\ & g(q(t)) \leq 0, \\ & h(z(t)) \leq 0. \end{aligned} \tag{F2$'$}$$

In the optimal control theory literature, optimization problems of the form (F2′) are referred to as *optimal control problems with state constraints*. Despite a rich body of literature in optimal control, control problems with state constraints are, in general, very difficult to solve as they impose boundaries on the state space of the dynamical system (Trélat 2012). While some results can be derived in special cases with certain structural properties, there is no systematic way to deal with these problems; we refer to the survey paper Hartl et al. (1995) for an overview. The challenge brought by state constraints are further outlined in Section 4.1 in the specific setting of our problem.

We combine several techniques from optimal control theory to derive the optimal transient control. Our solution strategy is to first derive the structure of the optimal scheduling policy for our model with degradation. In particular, as we shall explain in Theorem 4, the optimal scheduling policy switches priority at most once and can be characterized by two simple index rules. Then solving for the optimal scheduling problem reduces to finding the policy curve that governs where in space the switch in priority rule happens. We provide a closed form characterization of the policy curve in Proposition 4 in the case where the switch is from strict priority to Class 1 to strict priority to Class 2. For the case where the switch is from strict priority to Class 2 to strict priority to Class 1, we provide a numerical construction of the policy curve. We comment that due to the complexity introduced by patient degradation, we are not able to characterize the policy curve in closed form in the later case. We shall provide more discussions about this in Section 4.3.3.

The next theorem characterizes the structure of the transient optimal scheduling policy. Before we present the actual theorem, we first introduce two index rules. We define the $c\mu$-rule as a policy that prioritizes the class with a higher $c_i\mu_i$ value, i.e. the $c\mu$-index. Similarly, the modified $c\mu/\theta$-rule is a policy that prioritizes the class with a higher modified $c\mu/\theta$-index, as defined in Theorem 1. Note that the policy is to be specified for each time point $t \in [0,\tau]$ or each state $q \in \mathbb{R}^2_+$. In particular, for Markovian controls, a scheduling policy can be viewed as a mapping from the 'current' state of the system $q(t) = (q_1(t), q_2(t))$ to the allocation of the service capacity $z(t) = (z_1(t), z_2(t))$.

Theorem 4 *Under Assumptions 1 and 2, for the transient optimal control problem* (F2′)*:*

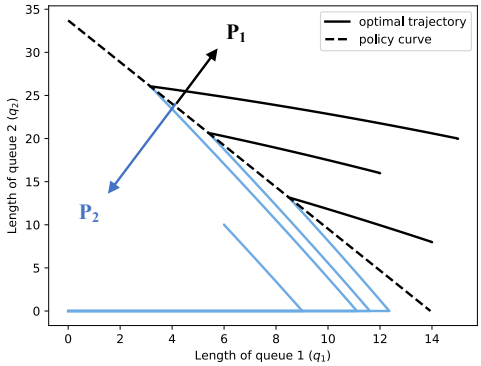
I. If the $c\mu$-rule and the modified $c\mu/\theta$-rule both prioritize class i, $i = 1,2$, then the strict priority rule to class i is optimal for any $t \in [0,\tau]$.

II. If the $c\mu$-rule prioritizes class i but the modified $c\mu/\theta$-rule prioritizes class j, for $i \neq j, i, j = 1,2$, then there exist positive real numbers ϵ and M such that it is optimal to prioritize class i when $q_1(t) + q_2(t) < \epsilon$ and prioritize class j when $q_1(t) + q_2(t) > M$. Furthermore, the optimal scheduling policy switches priority at most once over the transient time horizon $[0,\tau]$.

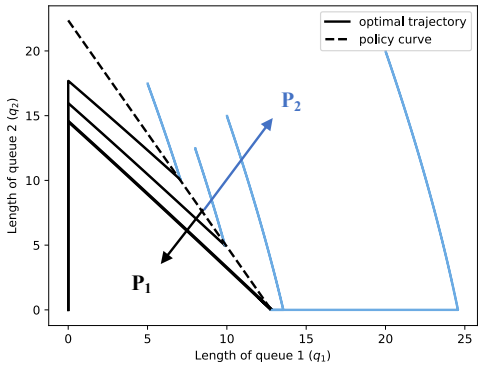
Based on Theorem 4, if the $c\mu$-rule and the modified $c\mu/\theta$-rule agree with each other, it is optimal to give strict priority to the class with a higher $c\mu$-index (and correspondingly higher modified $c\mu/\theta$-index) for any $q \in \mathbb{R}^2_+$. If the two index rules do not agree, we will follow the $c\mu$-rule when we are close enough to the equilibrium point, $(0,0)$; when we are far from the equilibrium point, we should follow the modified $c\mu/\theta$-rule. Moreover, in this case, we switch priority at most once; the time the switch occurs depends on the value of q_0. This indicates that there exists a *policy curve* $\{q : u(q) = 0\}$, where we switch from the modified $c\mu/\theta$-rule to the $c\mu$-rule.

The remaining task is to characterize the policy curve. In Figure 6, we provide numerical examples of the optimal trajectory of the queue length process. Figure 6(a) shows the case where the modified $c\mu/\theta$-rule prioritizes Class 1 while the $c\mu$-rule prioritizes Class 2. We plot four optimal fluid trajectories starting from different initial conditions. We also plot the corresponding policy curve (dashed line), which we construct by interpolating the switching points into a line. Figure 6(b) shows that case where the modified $c\mu/\theta$-rule prioritizes Class 2 while the $c\mu$-rule prioritizes Class 1. Similar to Figure 6(a), we plot four optimal fluid trajectories starting from different initial conditions and the corresponding policy curve. We will provide more discussions about the policy curve in Sections 4.3.3 and 4.4.1.

Figure 6 **Optimal transient queue length trajectory**
((a): $\lambda_1 = 10, \lambda_2 = 20, \mu_1 = 1.5, \mu_2 = 3, \gamma = 0.2, \theta_1 = 0.1, \theta_2 = 0.4, s = 17, c_1 = 5, c_2 = 3$
(b): $\lambda_1 = 10, \lambda_2 = 20, \mu_1 = 1, \mu_2 = 2.5, \gamma = 0.4, \theta_1 = 0.1, \theta_2 = 0.2, s = 26, c_1 = 5, c_2 = 1$**)**



(a) The $c\mu$-rule: P_2, the modified $c\mu/\theta$-rule: P_1



(b) The $c\mu$-rule: P_1, the modified $c\mu/\theta$-rule: P_2

Remark 2 *Even though Theorem 4 is stated under Assumption 1, following the same line of analysis, we can show that if* $\theta_1 = \theta_2 = \gamma = 0$*, we can recover the well-known optimality of the* $c\mu$*-rule throughout the transient time horizon (see Corollary 1 in Appendix C). If* $\gamma = 0$ *but* $\theta_1, \theta_2 > 0$*, then we should follow the* $c\mu$*-rule when we are close to the origin and the ordinary* $c\mu/\theta$*-rule*

when we are far from the origin (see Corollary 2 in Appendix C). A result similar to Corollary 2, i.e. for a two-class queue with abandonment but no deterioration, is established in Larrañaga (2015). Nevertheless, the approach utilized in Larrañaga (2015) for the special case is not directly generalizable to our setting with degradation.

We next provide the general strategy of proving Theorem 4. It includes three main parts. We first provide some formal definitions to describe the boundary behavior and rule out some 'irregular' boundary behaviors such as chattering in Section 4.1. We then establish the optimal scheduling policy when $q_1 + q_2 < \epsilon$ for ϵ sufficiently small in Section 4.2. This is done by solving the optimal control problem directly. Lastly, we establish the optimal scheduling policy for the rest of the state space in Section 4.3. This is done by utilizing Pontryagin's Minimum Principle.

4.1. Boundary Behavior

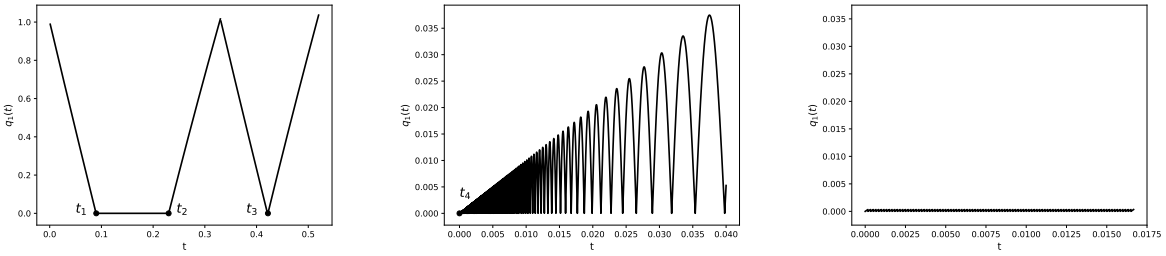
The main challenge in dealing with an optimal control problem of the form (F2′) is to characterize the system behavior on the boundary where the state constraints hold tight. In our case, the state constraint $g(q(t)) = -q(t) \leq 0$ requires the queue length process $q(t)$ to stay non-negative for all $t \in [0, \tau]$.

To characterize the boundary behavior, we would ideally like to identify when the trajectory enters the boundary and when it exits the boundary. In particular, we would like to characterize the time points t_k's when $g_i(q(t_k)) = 0$ for some $i = 1, 2$, but for any $\delta > 0$, there exists $t \in (t_k - \delta, t_k + \delta)$ such that $g_i(q(t)) > 0$. An important class of points of this type is known as the *junction time* (Hartl et al. 1995). We next provide some formal definitions to characterize the junction times. An interval $\mathcal{I} := [t_1, t_2] \subseteq [0, \tau]$ (or $[t_1, t_2)$, $(t_1, t_2]$, (t_1, t_2)) is called an *interior arc* if $g(q(t)) < 0$ holds for all $t \in \mathcal{I}$. Correspondingly, an interval $\mathcal{I} := [t_1, t_2] \subseteq [0, \tau]$ (or $[t_1, t_2)$, $(t_1, t_2]$, (t_1, t_2)) is called a *boundary arc* if $g_i(q(t)) = 0$, for some $i = 1, 2$, holds for all $t \in \mathcal{I}$. A time instant t_1 is called an *entry time* if an interior interval ends at and a boundary interval starts at t_1. A time instant t_2 is called an *exit time* if a boundary interval ends and an interior interval starts at t_2. Furthermore, if the trajectory of $q_i(t)$, $i = 1, 2$, only 'touches' the boundary at time t_3, i.e., $q_i(t_3) = 0$, but there exists $\delta > 0$ such that $q_i(t) > 0$ for any $t \in (t_3 - \delta, t_3 + \delta)$ and $t \neq t_3$, then t_3 is called a *contact point*. Entry, exit, and contact times taken together are called *junction times*. Figure 7(a) provides a pictorial illustration of different types of junction times for $q_1(t)$. In particular, t_1, t_2, and t_3 in Figure 7(a) are an entry, exit, and contact point respectively. In addition, the interval $[t_1, t_2]$ is a boundary arc, and the interval $[0, t_1)$ is an interior arc.

Not all boundary trajectories can be characterized by the junction times. A class of boundary behaviors that is often hard to deal with is known as chattering, which happens when the trajectory

$q_i(t)$ oscillates between zero and positive values infinitely fast. Specifically, a time instant t_4 is said to be a *chattering point* of the state trajectory q_i, if $q_i(t_4) = 0$, and for any $\delta > 0$ there exists s' and $s'' \in (t_4 - \delta, t_4 + \delta)$ such that $q_i(s') > 0$ and $q_i(s'') = 0$. In addition, an interval is said to be a *chattering interval* if any sub-interval of it contains at least one chattering point. Figure 7(b) provides an example where the state trajectory has a chattering point t_4, and Figure 7(c) provides an example of a chattering interval.

Figure 7 Different types of junction times and chattering behavior



(a) Entry/exit/contact point (b) Chattering point (c) Chattering interval

Chattering behavior can arise in many different optimal control problems. One classical example is Fuller's problem (Fuller 1963). Noticeably, for non-constrained linear control problems with compact polyhedral control space, it has been shown that there always exists an optimal solution that switches finitely many times among the vertices of the control polyhedron; see, for example, Chapter 2.8 in Schättler and Ledzewicz (2012). However, the pathological situation of chattering has not been ruled out for linear systems with state constraints, which is the case of our problem (F2′). We overcome the difficulty here by showing that for (F2′), it is without loss of optimality to consider trajectories without chattering points or chattering intervals.

Lemma 1 *For the transient optimal control problem* (F2′)*, it is without loss of optimality to consider state trajectories without chattering behavior.*

4.2. The $c\mu$-Rule Near the Origin

When the state is close enough to the origin $(0, 0)$, which is also an equilibrium point for the fluid system under Assumption 2 and an appropriate control, we establish that the $c\mu$-rule is optimal.

Proposition 1 *Under Assumptions 1 and 2, for the transient optimal control problem* (F2′)*, if* $q_1(t), q_2(t) \in [0, \epsilon)$*, with* $\epsilon > 0$ *sufficiently small, the* $c\mu$*-rule is optimal on the transient time interval* $[t, \tau]$.

The result in Proposition 1 is derived based on the observation that when the queue length is sufficiently small, the dominant effect for the system dynamic comes from service completions which has an order ϵ effect. The effect of abandonment and degradation is only second-order, namely, order ϵ^2. This can be seen clearly when we directly solve the ODE to specify the trajectories under different scheduling rules. Focusing on service completion only, $c_i\mu_i$ is the rate at which we can reduce the holding cost per unit time and per unit capacity allocated to serving Class i jobs, $i = 1, 2$. Intuitively, in order to reduce holding cost as fast as possible, the class with a larger $c\mu$-index should be prioritized.

4.3. The Optimal Policy for the Rest of the State Space

When the states are far away from the origin, we have to take abandonment and degradation into account, and these substantially complicate the analysis. To develop structural insights in this region, we utilize a necessary characterization for the optimal solution to the control problem, which is known as Pontryagin's Minimum Principle (Hartl et al. 1995).

To understand the underlying mechanism, we first note that if we view the optimal control problem (F2′) as an infinite dimensional linear program, then we can write down its dual problem and study the optimal primal-dual structure. There are two classes of 'dual variables'. One is referred to as the *adjoint vectors* (also known as the *co-state vectors*), which are the 'dual variables' for the fluid dynamics, i.e. $dq(t) = f(q(t), z(t))$. The other is called the *Lagrangian multipliers*, which are the 'dual variables' for the state constraints, i.e. $g(q(t)) \le 0$, and the pure-control constraints i.e. $h(z(t)) \le 0$. More precisely, let $p \in \mathbb{R}^2$ denote the adjoint vector, and $\eta \in \mathbb{R}^2$ and $\xi \in \mathbb{R}^3$ denote the Lagrangian multipliers for the state and control constraints, respectively. The *Hamiltonian* $H : \mathbb{R}^2 \times \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ of the system can be defined as:

$$\begin{aligned} H(q(t), z(t), p(t)) &:= p(t)^T f\left(q(t), z(t)\right) + F\left(q(t)\right) \\ &= p_1(t)dq_1(t) + p_2(t)dq_2(t) + c_1q_1(t) + c_2q_2(t) \\ &= p_1(t)\left(\lambda_1 - \mu_1 z_1(t) - \theta_1 q_1(t) + \gamma q_2(t)\right) + p_2(t)\left(\lambda_2 - \mu_2 z_2(t) - (\theta_2 + \gamma)q_2(t)\right) \\ &\quad + c_1q_1(t) + c_2q_2(t). \end{aligned}$$

The *augmented Hamiltonian* $L : \mathbb{R}^2 \times \mathbb{R}^2 \times \mathbb{R}^2 \times \mathbb{R}^2 \times \mathbb{R}^3 \to \mathbb{R}$ is defined as

$$\begin{aligned} L(q(t), z(t), p(t), \eta(t), \xi(t)) &:= H(q(t), z(t), p(t)) + \eta(t)^T g(q(t)) + \xi(t)^T h(z(t)) \\ &= p_1(t)\left(\lambda_1 - \mu_1 z_1(t) - \theta_1 q_1(t) + \gamma q_2(t)\right) \\ &\quad + p_2(t)\left(\lambda_2 - \mu_2 z_2(t) - (\theta_2 + \gamma)q_2(t)\right) + c_1q_1(t) + c_2q_2(t) - \eta_1(t)q_1(t) \\ &\quad - \eta_2(t)q_2(t) + \xi_1(t)(z_1(t) + z_2(t)) - \xi_2(t)z_1(t) - \xi_3(t)z_2(t). \end{aligned}$$

Pontryagin's Minimum Principle states a number of necessary conditions which the optimal solution to the optimal control problem (F2′) satisfies. The actual Theorem can be found in Appendix B.3. Here we provide a brief overview of the conditions.

1) *Ordinary Differential Equation condition* (ODE) specifies the dynamics of the 'optimal primal trajectory' $q^*(t)$, i.e.,

$$q^*(0) = q_0, \quad dq^*(t) = f\left(q^*(t), z^*(t)\right). \tag{ODE}$$

2) *Adjoint Vector condition* (ADJ) specifies the dynamics of the 'optimal dual trajectory' $p^*(t)$. In particular, we have

$$dp_1^*(t) = \theta_1 p_1^*(t) - c_1 + \eta_1^*(t), \quad dp_2^*(t) = -\gamma p_1^*(t) + (\theta_2 + \gamma) p_2^*(t) - c_2 + \eta_2^*(t). \tag{ADJ}$$

In general, we cannot fully characterize $p^*(t)$ due the fact that $p_i^*(0)$ and $\eta_i^*(t)$ are 'unspecified', i.e. we cannot fully specify their values or dynamics based on the necessary conditions.

3) *Minimization condition* (M) characterizes the optimal control $z^*(t)$ as a minimizer of the Hamiltonian, i.e.

$$H(q^*(t), z^*(t), p^*(t)) = \min_z \{H(q^*(t), z(t), p^*(t))\}. \tag{M}$$

As $H(q^*(t), z(t), p^*(t))$ is linear in $z(t)$, it is easy to see from (M) that the optimal control strictly prioritizes one class at any given time. In particular, as $z_1^*(t) + z_2^*(t) = s$ for $t \in [0, \tau^*]$, we can write $z_1^*(t) = s - z_2^*(t)$. Then, we define

$$\psi(t) := \frac{\partial H(q^*(t), s - z_2(t), z_2(t), p^*(t))}{\partial z_2} = \mu_1 p_1^*(t) - \mu_2 p_2^*(t). \tag{13}$$

$\psi(t)$ is referred to as the *switching curve*, because the sign of $\psi(t)$ determines which class we should give priority to. Note that to minimize H, when $\psi(t) > 0$, priority should be given to Class 1 at time t, i.e.,

$$z_1^*(t) = \begin{cases} s & \text{if } q_1^*(t) > 0 \\ \min\left\{s, \frac{\lambda_1 + \gamma q_2^*(t)}{\mu_1}\right\} & \text{if } q_1^*(t) = 0, \end{cases} \quad \text{and} \quad z_2^*(t) = s - z_1^*(t). \tag{14}$$

When $\psi(t) < 0$, priority should be given to Class 2, i.e.

$$z_1^*(t) = s - z_2^*(t), \quad \text{and} \quad z_2^*(t) = \begin{cases} s & \text{if } q_2^*(t) > 0 \\ \min\left\{s, \frac{\lambda_2}{\mu_2}\right\} & \text{if } q_2^*(t) = 0. \end{cases} \tag{15}$$

However, when $\psi(t) = 0$, the optimal control is undetermined. We also note that $\psi(t)$ can be fully characterized by $p_i^*(t)$'s. Thus, analyzing the structure of the optimal dual trajectory $p^*(t)$ can reveal important information about the optimal scheduling policy $z^*(t)$.

4) For optimal control problems with state constraints, if F, f, g, h do not depend on t explicitly, *Hamiltonian condition* (H) requires that $H(q^*(t), z^*(t), p^*(t))$ is a constant for all $t \in [0, \tau^*]$. Further,

if the problem has a fixed termination state but free termination time, as in our case, then the constant is equal to zero (Cristiani and Martinon 2010). In particular, we have

$$H(q^*(t), z^*(t), p^*(t)) = 0. \tag{H}$$

5) *Transversality condition* (T) requires that

$$-\mu_1 p_1^*(t) + \xi_1^*(t) - \xi_2^*(t) = 0, \quad -\mu_2 p_2^*(t) + \xi_1^*(t) - \xi_3^*(t) = 0. \tag{T}$$

6) *Complementarity condition* (C) requires that

C1) $\eta_1^*(t) = 0$ if $q_1^*(t) > 0$; $\eta_1^*(t) \geq 0$ if $q_1^*(t) = 0$.

C2) $\eta_2^*(t) = 0$ if $q_2^*(t) > 0$; $\eta_2^*(t) \geq 0$ if $q_2^*(t) = 0$.

C3) $\xi_1^*(t) = 0$ if $z_1^*(t) + z_2^*(t) < s$; $\xi_1^*(t) \geq 0$ if $z_1^*(t) + z_2^*(t) = s$.

C4) $\xi_2^*(t) = 0$ if $z_1^*(t) > 0$; $\xi_2^*(t) \geq 0$ if $z_1^*(t) = 0$.

C5) $\xi_3^*(t) = 0$ if $z_2^*(t) > 0$; $\xi_3^*(t) \geq 0$ if $z_2^*(t) = 0$.

7) *Jump condition* (J) characterizes the potential discontinuity of the adjoint vector $p^*(t)$ and the Hamiltonian $H(q^*(t), z^*(t), p^*(t))$ at junction times or in the boundary arcs. Specifically, For any time β in a boundary arc or a junction time, the adjoint vector $p^*(t)$ and the Hamiltonian $H(q^*(t), z^*(t), p^*(t))$ may have a discontinuity, but they must satisfy the following jump conditions:

$$\begin{aligned} (J1): p^*(\beta-) &= p^*(\beta+) + \omega_1^*(\beta)\nabla_q g_1(q^*(\beta)) + \omega_2^*(\beta)\nabla_q g_2(q^*(\beta)) \\ (J2): H(q^*(\beta-), z(\beta-), p^*(\beta-)) &= H(q^*(\beta+), z(\beta+), p^*(\beta+)) - \omega_1^*(\beta)\nabla_t g_1(q^*(\beta)) \\ &\quad - \omega_2^*(\beta)\nabla_t g_2(q^*(\beta)) \\ (J3): \omega^*(\beta) &\geq 0, \quad \omega^*(\beta)^T g\left(q^*(\beta)\right) = 0, \end{aligned} \tag{J}$$

where $\nabla_x g$ denote the derivative of g with respect to x.

From the discussion of the necessary conditions, we highlight that if we can characterize the switching curve $\psi(t)$, then we will be able to unfold the corresponding optimal policies. However, this is a highly nontrivial task, as we are not able to fully characterize $p^*(t)$.

4.3.1. The Modified $c\mu/\theta$-Rule Far from the Origin We now derive several key properties of the switching curve $\psi(t)$ from Pontryagin's Minimum Principle. These properties together allow us to establish the optimal scheduling policy when the states are large (far from the origin).

The first property characterize the switching curve on the boundary arc.

Lemma 2 *Let $[t_1, t_2]$ be a boundary arc along the optimal state trajectory with entry point t_1 and exit point t_2. For any $t \in (t_1, t_2)$, the switching curve $\psi(t) = 0$.*

The second property establish the continuity of the switching curve.

Lemma 3 *The switching curve $\psi(t)$ is continuous over $[0,\tau^*]$.*

Assume there exists an optimal control to problem (F2′) under which the state trajectory only has a finite number of junction points. Let N denote the total number of entry and contact points in the optimal state trajectory $q_1^*(t)$ and $q_2^*(t)$. These N entry or contact points are ordered and denoted by τ_j, $j=1,...,N$. In particular, τ_1 is the first time when one of the queues gets emptied from the initial queue length q_0. τ_N is the last time before τ^* when one of the queues gets emptied. Naturally, the queue that gets emptied at time τ_N is maintained at zero until the other queue reaches zero at time τ^*. From Lemmas 2 and 3, we know that $\psi(\tau_j)=0$ for entry/exit point τ_j. To this end, we examine the switching curve backward in time from each entry point and derive the following characterization of the switching curve.

Lemma 4 *For any entry and contact point $\tau_j, j=1,...,N$, there exists an interval $(0,\alpha_j), 0<\alpha_j<\tau_j$, such that for $t\in(0,\alpha_j)$, the backward switching curve $\psi(\tau_j-t)$ takes the form*

$$\psi(\tau_j-t)=A_1(\tau_j)e^{-\theta_1 t}+A_2(\tau_j)e^{-(\theta_2+\gamma)t}+\frac{c_1}{\theta_1}\mu_1-\left(\frac{c_2}{\theta_2+\gamma}\mu_2+\frac{\frac{\gamma}{(\theta_2+\gamma)}c_1}{\theta_1}\mu_2\right),$$

where $A_1(\tau_j), A_2(\tau_j)$ are constants that depend on the values of τ_j and $p^*(\tau_j)$.

Following Lemma 4, we define the *pseudo switching curve* backward from τ_j as

$$D^{\tau_j}(t):=A_1(\tau_j)e^{-\theta_1 t}+A_2(\tau_j)e^{-(\theta_2+\gamma)t}+\frac{c_1}{\theta_1}\mu_1-\left(\frac{c_2}{\theta_2+\gamma}\mu_2+\frac{\frac{\gamma}{(\theta_2+\gamma)}c_1}{\theta_1}\mu_2\right),\quad j=1,...,N. \tag{16}$$

In particular, the pseudo switching curve removes the constraint that $t\in(0,\alpha_j)$ from Lemma 4 and it agrees with the switching curve $\psi(\tau_j-t)$ as long as the multipliers $\eta_1^*(\tau_j-t)$ and $\eta_2^*(\tau_j-t)$ stay at zero. However, if one of the multipliers becomes strictly positive at some time s, i.e., $\eta_i^*(\tau_j-s)>0$ for some $i=1,2$, the switching curve $\psi(\tau_j-t)$ may deviate from the pseudo switching curve for $t\geq s$.

The significance of Lemma 4 is that even though the constants $A_1(\tau_j)$ and $A_2(\tau_j)$ are unspecified, there are only a very few possibilities for the shape of $D^{\tau_j}(t)$, and thus for the part of $\psi(\tau_j-t)$ that coincides with $D^{\tau_j}(t)$. Now, consider the first (forward in time) entry point τ_1. By the definition of τ_1, both classes have strictly positive queues before τ_1, so the multipliers $\eta_1^*(\tau_1-t)$ and $\eta_2^*(\tau_1-t)$ are zero for all $t\in(0,\tau_1]$. In this case, the backward switching curve $\psi(\tau_1-t)$ and the pseudo switching curve $D^{\tau_1}(t)$ coincide over the interval $t\in(0,\tau_1]$. Note that for $t>\tau_1$, the queue length

trajectory is beyond its initialization, and thus $\psi(\tau_1 - t)$ is not defined. On the other hand, the pseudo switching curve $D^{\tau_1}(t)$, as a function of t, is well-defined for all $t \geq 0$. Sending t to infinity in the pseudo switching curve $D^{\tau_1}(t)$, we get

$$\lim_{t\to\infty} D^{\tau_1}(t) = \frac{c_1}{\theta_1}\mu_1 - \left(\frac{c_2}{\theta_2+\gamma}\mu_2 + \frac{\frac{\gamma}{(\theta_2+\gamma)}c_1}{\theta_1}\mu_2\right). \tag{17}$$

The sign of the right-hand-side of (17) is governed by the modified $c\mu/\theta$-index, which is positive if the modified $c\mu/\theta$-index for Class 1 is larger. It is important to correctly interpret the limit in (17) for the backward switching curve $\psi(\tau_1 - t)$. Because $\psi(\tau_1 - t)$ only couples with $D^{\tau_1}(t)$ on $(0, \tau_1]$ and is not defined for $t > \tau_1$, one may hypothesize that if the initial queue lengths, q_0, are large enough, then τ_1, the amount of time needed to empty one of the queues, is also large, and we might be able to send t large enough that the sign of $\psi(\tau_1 - t)$ will be governed by the modified $c\mu/\theta$-index. However, we also need to note that the constants $A_1(\tau_1)$ and $A_2(\tau_1)$ depend on q_0 through τ_1 and $p^*(\tau_1)$. We thus need to rigorously establish that $A_1(\tau_1)$ and $A_2(\tau_1)$ are properly bounded. Putting all these analysis together, we are able to establish the following result.

Proposition 2 *Under Assumptions 1 and 2, for the transient optimal control problem (F2′), there exists a positive real number M such that when $q_1(t) + q_2(t) > M$, the modified $c\mu/\theta$-rule is optimal at t.*

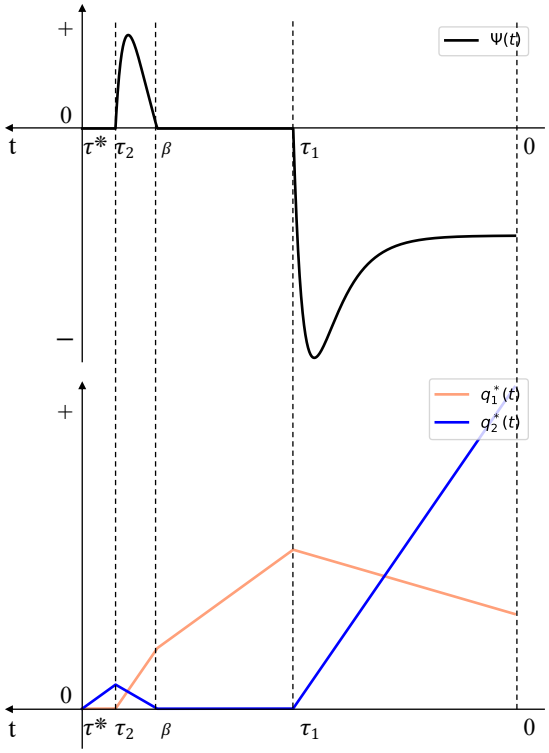
4.3.2. Number of Priority Switches Propositions 1 and 2 imply that the $c\mu$-rule is optimal when the states are close enough to the origin, and the modified $c\mu/\theta$-rule is optimal when the states are far away from the origin. We now specify what happens in between these two extreme regions. By analyzing possible shapes of the switching curve characterized in Lemma 4, we are able to establish the following proposition.

Proposition 3 *Under Assumptions 1 and 2, for the transient optimal control problem* (F2′)*, if the $c\mu$-rule and the modified $c\mu/\theta$-rule prioritize the same class, the optimal transient scheduling policy does not switch priority. If the two index rule prioritize different classes, the optimal transient scheduling policy switches priority at most once over the transient time horizon* $[0, \tau^*]$.

Figure 8 illustrates the interaction between the switching curve and the transient system dynamics. In particular, we plot the switching curve $\psi(t)$ and the corresponding optimal state trajectory $q^*(t)$ for $t \in [0, \tau^*]$ backward in time. In this example, over the initial time interval $[0, \tau_1)$, $\psi(t)$ is negative, so strict priority is given to Class 2 (the modified $c\mu/\theta$-rule). Class 2 queue empties at time τ_1 and is given priority to be maintained at zero over the interval $[\tau_1, \beta)$. Immediately after

β, the switching curve becomes strictly positive and priority is switched to Class 1 (the $c\mu$-rule). Note that Class 1 queue decreases and the Class 2 queue increases from over $[\beta, \tau_2)$. Priority is kept at Class 1 on $[\beta, \tau^*]$.

Figure 8 **Example backward switching curve and state trajectory**



4.3.3. The Policy Curve From Theorem 4, the optimal scheduling rule switches priority at most once. This implies there exists a policy curve $\mathcal{P}$ that divides the state space and governs where the priority switches. Note, this curve is distinct from, but intimately related to, the switching curve, $\psi(t)$. Suppose the $c\mu$-rule prioritizes Class 2 and the modified $c\mu/\theta$-rule prioritizes Class 1. By utilizing Hamiltonian condition (H), we are able to characterize the policy curve for switching from P_1 to P_2 explicitly. Namely, if the states are initialized 'above' $\mathcal{P}$, then the server prioritizes Class 1 until $q(t) \in \mathcal{P}$ at some t. From time t onwards, the server prioritizes Class 2 until the system is emptied at τ.

Proposition 4 *Under Assumptions 1 and 2, for the transient optimal control problem* (F2′)*, if* $c_1\mu_1 < c_2\mu_2$ *but* $\frac{c_1}{\theta_1}\mu_1 > \left(\frac{c_2}{\theta_2+\gamma}\mu_2 + \frac{\frac{\gamma}{(\theta_2+\gamma)}c_1}{\theta_1}\mu_2\right)$*, the policy curve* $\mathcal{P}$ *for switching from* P_1 *to* P_2 *is given by*

$$\mathcal{P} := \left\{(\mathbf{a_1}, \mathbf{a_2}) \in \mathbb{R}^2_+ : \frac{1}{\mu_2}\left(\frac{c_1(\lambda_1\mu_1 + (\lambda_1 - s\mu_1)\mu_2)}{\theta_1} + \frac{B_1(\mathbf{a_2})B_2(\mathbf{a_1}, \mathbf{a_2})}{B_3(\mathbf{a_1}, \mathbf{a_2})}\right) = 0\right\}, \tag{18}$$

where

$$B_1(\mathbf{a_2}) := (c_1(-\mathbf{a_2}(\theta_2+\gamma)+\lambda_2)\mu_1 + c_2\mathbf{a_2}\theta_1\mu_2 + c_1(\mathbf{a_2}\gamma+\lambda_1-s\mu_1)\mu_2)$$

$$B_2(\mathbf{a_1},\mathbf{a_2}) := \Bigg(-\mu_2(\mathbf{a_2}\gamma\theta_1 + \mathbf{a_1}\theta_1(\gamma-\theta_1+\theta_2) - \gamma\lambda_1 + \theta_1\lambda_1 - \theta_2\lambda_1 - \gamma\lambda_2 + s\gamma\mu_2)$$
$$+(\lambda_2 - s\mu_2)((\gamma-\theta_1+\theta_2)\mu_1 - \gamma\mu_2)\left(1+\frac{\mathbf{a_2}(\theta_2+\gamma)}{-\lambda_2+s\mu_2}\right)^{\frac{\theta_1}{\theta_2+\gamma}}\Bigg)$$

$$B_3(\mathbf{a_1},\mathbf{a_2}) := (\gamma-\theta_1+\theta_2)\left(\theta_1(\mathbf{a_2}(\theta_2+\gamma)-\lambda_2)\mu_1 + \theta_1(-\mathbf{a_2}\gamma+\mathbf{a_1}\theta_1-\lambda_1+s\mu_1)\mu_2\right).$$

If $c_1\mu_1 > c_2\mu_2$ but $\frac{c_1}{\theta_1}\mu_1 < \left(\frac{c_2}{\theta_2+\gamma}\mu_2 + \frac{\frac{\gamma}{(\theta_2+\gamma)}c_1}{\theta_1}\mu_2\right)$, the policy curve for switching from P_2 to P_1 cannot be characterized in closed form. This is due the degradation dynamics. In particular, we lack information of the Lagrange multiplier $\eta_1^*(t)$ on the boundary arc when $q_1^*(t)=0$. Due to the degradation, $\eta_1^*(t)$ not only affects $p_1^*(t)$ but also $p_2^*(t)$ through $p_1^*(t)$, see (ADJ). As such, the condition that $H(q^*(t),z^*(t),p^*(t))=0$ is not enough to pin down the value of policy curve. Note that this is not the case in Proposition 4, because on the boundary arc when $q_2^*(t)=0$, $\eta_1^*(t)$ affects $p_2^*(t)$ only. See Appendix B.7 for a more detailed discussion.

To characterize the policy curve for switching from P_2 to P_1, we propose the following numerical scheme:

Step 1. Construct n optimal trajectories $q^*(t)$ starting from different initial conditions that are far from the origin. This can be done by solving a discretized version of (F2′). We can also fix the time horizon to be a large enough constant value, T. This leads to a finite-dimensional linear program. Record the n corresponding switching points.

Step 2. Fit the best curve that goes through the n switching points.

In Section 4.4.1, we conduct extensive numerical experiments on $\mathcal{P}$ for switching from P_2 to P_1. The curve appears to be close to linear. Thus, we suggest setting n to be around 5, setting the discretization step size to be around $0.1\mu_1$, and fitting the best line to the n switching points.

4.4. Numerical Experiments

In this section, we conduct numerical experiments for the optimal transient scheduling problem. We first provide further analysis on the policy curve in Section 4.4.1. The goal is to study how the policy curve changes with model parameters. We also compare the transient fluid trajectory to the corresponding stochastic process in Section 4.4.2.

4.4.1. Sensitivity Analysis on the Policy Curve. To complement the result in Proposition 4, we restrict to the case where the $c\mu$-rule prioritizes the urgent class (Class 1) and the modified $c\mu/\theta$-rule prioritizes the moderate class (Class 2). This is also the more interesting case in practice,

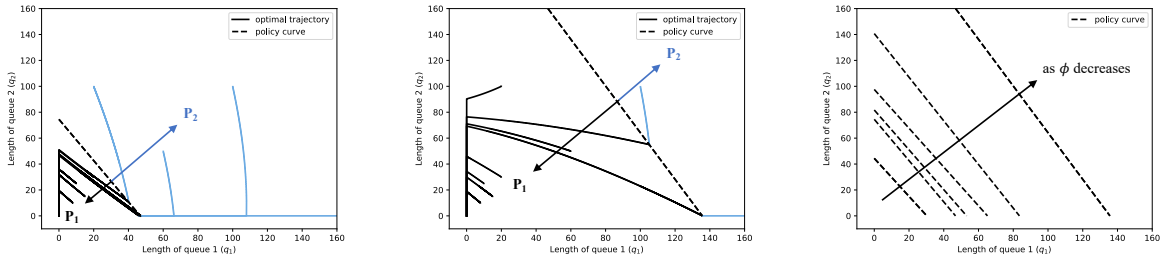
as the more urgent class is likely to have priority under 'normal' operations, i.e. near the equilibrium. Specifically, is it reasonable to assume that in practice, the urgent class has a much higher holding cost, thus $c_1\mu_1$ is likely to be larger than $c_2\mu_2$. Proactive care is attractive if it has the potential of reducing the overall workload, i.e. $\mu_1 < \frac{\gamma}{\theta+\gamma}\mu_2$. In this case, $\frac{c_1}{\theta_1}\mu_1 < \left(\frac{c_2}{\theta_2+\gamma}\mu_2 + \frac{\frac{\gamma}{(\theta_2+\gamma)}c_1}{\theta_1}\mu_2\right)$.

We first conduct sensitivity analysis on the value $\phi := \gamma/(\theta_2+\gamma)$. As discussed in Section 2, θ_2 can be interpreted as the self-curing rate, and ϕ can be interpreted as the true positive rate of some classifier, e.g. an early warning system, for proactive care. In what follows, we vary the value of ϕ from 0.6 to 0.4 in increments of size -0.05. Since γ and θ_2 both affect the value of ϕ, we first keep γ fixed and vary the value of θ_2 (Figure 9). Then, we keep θ_2 fixed and vary the value of γ (Figure 10). Figures 9 and 10 illustrate how the policy curve changes as ϕ decreases. The policy curve (plotted in dashed line) is interpolated from the switching epochs of the optimal trajectories. Plotting the policy curves together, Figures 9(c) and 10(c) demonstrate that the policy curve contracts inwards as ϕ increases. As the prediction power of the classifier improves, the region in which the optimal scheduling policy prioritizes moderate patients increases. When the classifier is perfect, i.e. $\phi = 1$ (achieved with $\theta_2 = 0$ in Figure 9(c)), the size of the region where it is optimal to prioritize Class 1 is minimized. The fact that the area of this minimal region is non-trivial suggests that the 'ϵ-neighborhood' around the origin where the $c\mu$-rule is optimal can be substantial. On the other hand, as the prediction power of the classifier declines, i.e. ϕ decreases, a phase transition in terms of the prioritization occurs. In particular, there exists a threshold ϕ_0 such that when $\phi \geq \phi_0$, we will give priority to Class 2 when we are far from the origin and give priority to Class 1 when we get close enough to the origin. Once $\phi < \phi_0$, the policy curve 'vanishes'; namely, we should give strict priority to Class 1 for all states. In the cases in Figures 9 and 10, the values of ϕ_0 are approximately 0.33 and 0.35 respectively. We note that given the complex expression of the policy curve in Proposition 4 when switching in the opposite direction (from P_1 to P_2), it is expected that the policy curve will vary under different parameters, even if they have the same ϕ value – i.e., Figure 9 versus Figure 10.

Similar to the above sensitivity analysis on the policy curve with respect to ϕ via θ_2 or γ, we also conduct numerical experiments for different values of s and θ_1. Our results indicate that the policy curve expands outwards as s increases, and contracts inwards as θ_1 decreases. Detailed figures for the sensitivity analysis for s and θ_1 are provided in Appendix D.

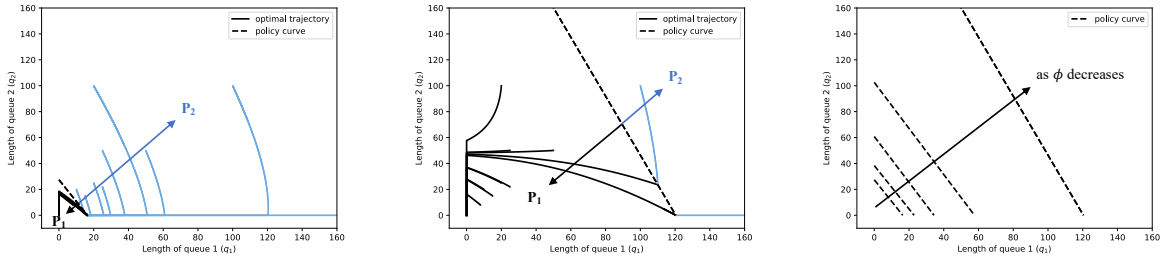
4.4.2. Stochastic Queue Length Process We have demonstrated in Section 3.3 that the fluid model and the stochastic system are close to each other in a long-run average sense. Indeed, the proximity between the fluid and the stochastic queue length process also holds in the transient

Figure 9 **Sensitivity analysis of the policy curve with respect to $\gamma/(\theta_2+\gamma)$ with γ fixed** **($\lambda_1=10,\lambda_2=20,\mu_1=1,\mu_2=2.5,\theta_1=0.1,s=26,c_1=5,c_2=1$)**



(a) $\phi=0.6,\gamma=0.1,\theta_2=0.07$ (b) $\phi=0.4,\gamma=0.1,\theta_2=0.15$ (c) Decreasing $\phi=$ 1, 0.6, 0.55, 0.5, 0.45, 0.4

Figure 10 **Sensitivity analysis of the policy curve with respect to $\gamma/(\theta_2+\gamma)$ with θ_2 fixed** **($\lambda_1=10,\lambda_2=20,\mu_1=1,\mu_2=2.5,\theta_1=0.1,s=26,c_1=5,c_2=1$)**



(a) $\phi=0.6,\gamma=0.38,\theta_2=0.25$ (b) $\phi=0.4,\gamma=0.17,\theta_2=0.25$ (c) Decreasing $\phi=0.6, 0.55, 0.5, 0.45, 0.4$

regime. As mentioned in Section 2.1, the fluid model can arise as a functional law of large numbers limit for a sequence of properly scaled stochastic systems in the conventional heavy traffic regime. In what follows, we elaborate on the scaling and conduct numerical comparisons between the fluid trajectory and scaled stochastic sample path.

Consider a sequence of stochastic systems indexed by n. For Class i in the n^{th} system, $i=1,2$, $n\in\mathbb{N}_+$, the arrival and service rates satisfy $\lambda_i^n:=\lambda_i n$, $\mu_i^n:=\mu_i n$. Moreover, we scale down space by considering the fluid-scaled queue length process $\bar{Q}_i^n(\cdot):=Q_i^n(\cdot)/n$ for the n-th stochastic system. Given the initial fluid queue length q_0, the n^{th} stochastic system has initial queue length $Q^n(0):=\lceil q_0 n\rceil$.

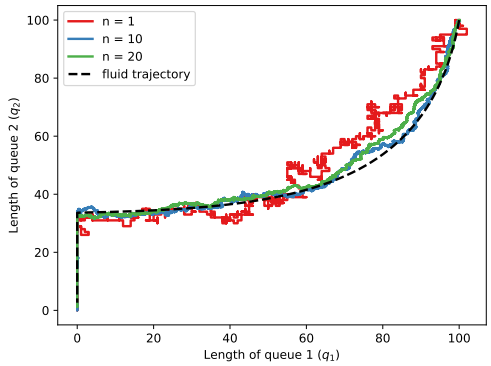
In each stochastic system, priority is assigned with non-preemptive service. In the case where the $c\mu$ rule prioritizes Class 2 and the modified $c\mu/\theta$-rule prioritizes Class 1, the policy curve for the n-th stochastic system is modified from the optimal fluid policy curve $\mathcal{P}$ characterized in Proposition 4. For the other case where the $c\mu$ rule prioritizes Class 1 and the modified $c\mu/\theta$-rule prioritizes Class 2, we impose an approximating linear policy curve $\mathcal{P}$ which is the best linear fit

to the optimal switching epochs found numerically. In either case, for the n^{th} stochastic system, a switch in priority will happen at time t if $\bar{Q}^n(t) \in \mathcal{P}^n$, where

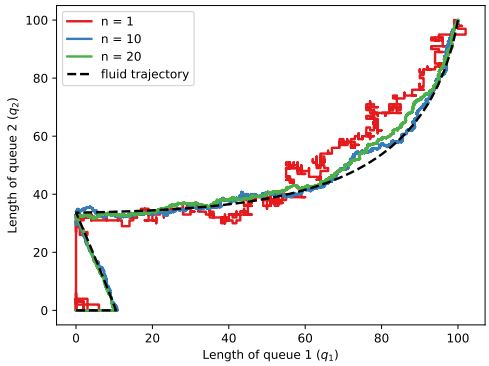
$$\mathcal{P}^n := \left\{ (\bar{Q}_1^n, \bar{Q}_2^n) : \bar{Q}_1^n \in [q_1 - 1/n, q_1 + 1/n], \bar{Q}_2^n \in [q_2 - 1/n, q_2 + 1/n], (q_1, q_2) \in \mathcal{P} \right\}.$$

Figure 11 compares the fluid trajectory with a simulated sample path for the stochastic system for different values of n. We observe from the plots that for a relatively small scaling parameter, e.g. $n = 10$, the stochastic sample path is already well approximated by the fluid model. Furthermore, if we plot the trajectory of the average queue length over multiple sample paths of the stochastic system, the behavior of the 'average trajectory' mimics the fluid model even more closely.

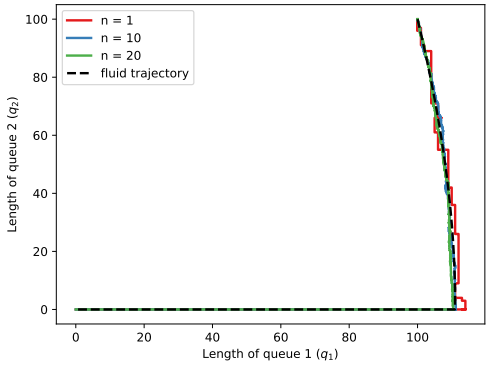
Figure 11 **Comparison of the transient fluid trajectory and the stochastic sample path**
((a)(b): $\lambda_1 = 10, \lambda_2 = 20, \mu_1 = 1.5, \gamma = 0.2, \theta_1 = 0.1, \theta_2 = 0.4, s = 17, c_1 = 5, c_2 = 3$
(c)(d): $\lambda_1 = 10, \lambda_2 = 20, \mu_1 = 1, \mu_2 = 2.5, \gamma = 0.4, \theta_1 = 0.1, \theta_2 = 0.2, s = 26, c_1 = 5, c_2 = 1$**)**



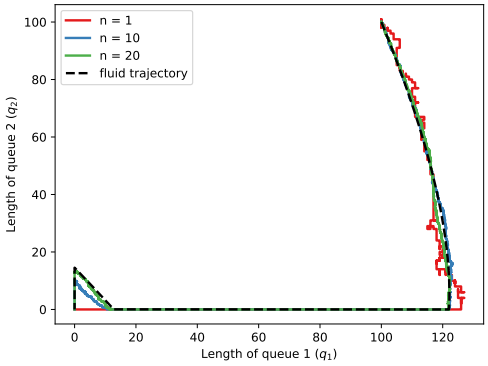
(a) P_1 ($\mu_2 = 2$)



(b) P_1 to P_2 ($\mu_2 = 3$)



(c) P_2 ($\mu_2 = 6$)

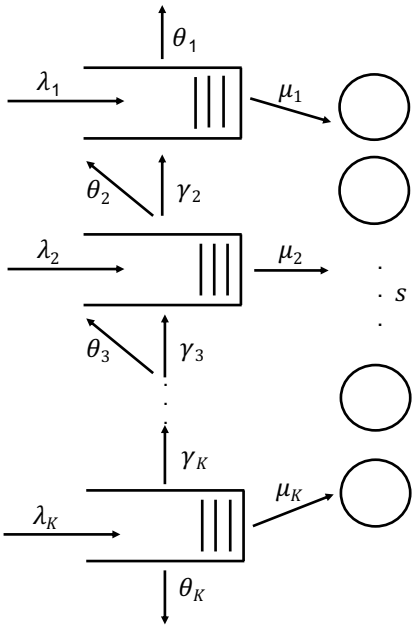


(d) P_2 to P_1 ($\mu_2 = 2.5$)

5. Generalization to Multi-Class System

Thus far, our analysis has focused on a two-class system. We now provide a heuristic extension of the optimal scheduling policy to minimize the long-run average cost and the transient cost for multi-class multi-server queueing networks with K patient classes as depicted in Figure 12 . The patient classes are associated with different urgency levels, with Class 1 being the most urgent and Class K the least. Class i, $i = 1, ..., K$, is associated with its arrival rate λ_i, service rate μ_i, abandonment rate θ_i and cost rate c_i. To capture patient deterioration, delayed Class-i patients, $i = 2, ..., K$, degrade into class $i-1$ at rate γ_i.

Figure 12 Multi-class queueing network



5.1. Long-Run Average Analysis

Following similar lines of analysis as in Section 3, we first optimize over the set of equilibrium points. In particular, we have the following linear program:

$$\begin{aligned} \min_{\{z_i^e,\, i=1,\ldots,K\}} \quad & \sum_{i=1}^{K} c_i q_i^e \\ s.t. \quad & \lambda_i - \mu_i z_i^e - q_i^e(\gamma_i + \theta_i) + \gamma_{i+1} q_{i+1}^e = 0, \quad i = 1, ..., K \\ & \sum_{i=1}^{K} z_i^e \le s \\ & z_i^e,\, q_i^e \ge 0, \quad i = 1, ..., K \end{aligned} \tag{19}$$

where $\gamma_1, \gamma_{K+1} \equiv 0$.

Rearranging the terms in (19), it is straightforward to see that the optimal solution to (19) is a generalization of the modified $c\mu/\theta$-rule. Namely, the solution assigns the maximum value to the z_i^e with a larger modified $c\mu/\theta$ index denoted by r_i, where

$$r_i := \mu_i \sum_{j=1}^{i} \frac{c_j}{\gamma_j + \theta_j} \prod_{k=j+1}^{i} \frac{\gamma_k}{\gamma_k + \theta_k},$$

with $\prod_{k=i+1}^{i} \frac{\gamma_k}{\gamma_k+\theta_k} \equiv 1$. Note that for the most urgent class (Class 1), the index r_1 is equal to $c_1\mu_1/\theta_1$, which coincides with the regular $c\mu/\theta$-index. For the other relatively moderate classes (Class $i = 2, ..., K$), $r_i > c_i\mu_i/\theta_i$ because it accounts for the degradation from the current class to more urgent/severe classes.

To establish the optimality of the modified $c\mu/\theta$-rule for the long-run average cost, we also need to verify that the optimal equilibrium point can be achieved under the modified $c\mu/\theta$-rule. This requires extending the Lyapunov analysis to establish the asymptotic stability of the equilibrium points under the corresponding priority rules. We note that this task will become very tedious, especially for a large number of patient classes, K.

5.2. Transient Analysis

Similar to the long-run equilibrium analysis, the transient analysis for the two-queue system can also be partially generalized to multi-class systems.

First, based on the insights from the two-class case, when the states are arbitrarily close to the origin, the effect of degradation and abandonment on the system dynamics is only second-order. Using a Taylor expansion argument, we can show that there exists $\epsilon > 0$ small enough, such that the $c\mu$-rule is optimal in the ϵ-neighborhood around the origin – i.e., when $q_i(t) \in [0, \epsilon)$ for $i = 1, 2, \ldots, K$.

Second, by analyzing Pontryagin's Minimum Principle for the multi-class case, we see that at any time t, the optimal policy prioritizes the class with the largest $p_i^*(t)\mu_i$ value, where p_i^* is the optimal adjoint vector. Let τ_1 be the first time after initialization when one of the queues gets emptied. Using a similar backward construction as in Lemma 4, we can characterize $p_i^*(\tau_1 - t)$ and show that $\lim_{t\to\infty} p_i^*(\tau_1 - t)\mu_i = r_i$, where r_i is the modified $c\mu/\theta$-index for Class i. This suggests that the modified $c\mu/\theta$-rule might be optimal when the states are far enough from the origin. However, we emphasize that this is only a heuristic argument due to similar complications encountered in Section 4.3.1. Namely, since the time argument t in the backward adjoint vector $p^*(\tau_1 - t)$ cannot exceed τ_1, rigorously establishing such a result requires a highly non-trivial derivation.

Lastly, the optimal scheduling policy for regions between the ϵ-neighborhood of the origin and the far from the origin region remain unclear. Though Pontryagin's Minimum Principle suggests

that the optimal control assigns strict priority to one class at any given time, it is not necessarily true that it switches priority at most once along the trajectory, as in the case of a two-class system. We perform extensive numerical experiments for a three-class model by solving the corresponding LP using discretization. The LP approximation confirms that the optimal solution follows the modified $c\mu/\theta$-rule when the state q is sufficiently far from the origin, and the $c\mu$-rule near the origin. In many problem instances, the optimal scheduling policy also switches priority rule at most once. However, there are also cases where the optimal scheduling policy switches priority more than once, and it follows neither the generalized modified $c\mu/\theta$-rule nor the $c\mu$-rule during part of the transient horizon.

We find that applying a *one-switch* policy, where we switch priority at most once, and follow the the modified $c\mu/\theta$ when the system state is far from the equilibrium and the $c\mu$-rule when the state is close to the equilibrium, turns out to have reasonably good performance. Figure 13 demonstrates two problem instances where the modified $c\mu/\theta$-rule and the $c\mu$-rule prioritize in the order of Classes $3, 2, 1$ and Classes $1, 2, 3$ respectively. In both instances, the optimal LP solution may, under certain initial conditions, prioritize Class 2 over part of the transient horizon. Nevertheless, the optimality gap of the one-switch policy is fairly small, while applying the modified $c\mu/\theta$-rule or the $c\mu$-rule throughout can sometimes lead to very large optimality gaps. In general, we expect the one-switch policy is a reasonable heuristic policy when the modified $c\mu/\theta$-index and the $c\mu$-index are relatively aligned.

Figure 13 **Optimality gap of different policies (percentage gap to the optimal LP approximation)**
((1): $\lambda_1 = 10, \lambda_2 = 20, \lambda_3 = 30, \mu_1 = 1, \mu_2 = 3, \mu_3 = 6, \theta_1 = 0.1, \theta_2 = 0, \theta_3 = 0.2, \gamma_2 = 0.3, \gamma_3 = 0.3, s = 50, c_1 = 50, c_2 = 3, c_3 = 1$, modified $c\mu/\theta$-index $= \{250, 780, 948\}, c\mu$-index $= \{50, 9, 6\}$
(2): $\lambda_1 = 10, \lambda_2 = 20, \lambda_3 = 30, \mu_1 = 5, \mu_2 = 5.1, \mu_3 = 8, \theta_1 = 0.2, \theta_2 = 0.1, \theta_3 = 0.2, \gamma_2 = 0.1, \gamma_3 = 0.2, s = 30, c_1 = 15, c_2 = 10, c_3 = 6$, modified $c\mu/\theta$-index $= \{375, 446, 470\}, c\mu$-index $= \{75, 51, 48\}$**)**

	Instance 1			Instance 2		
Initialization	One-switch	Modified $c\mu/\theta$	$c\mu$	One-switch	Modified $c\mu/\theta$	$c\mu$
5	**0.00%**	113.59%	**0.00%**	**0.00%**	27.87%	**0.00%**
10	**0.00%**	101.63%	**0.00%**	**0.00%**	27.37%	**0.00%**
50	4.61%	29.87%	4.61%	**0.00%**	22.20%	**0.00%**
100	11.18%	15.74%	116.75%	**0.00%**	17.84%	**0.00%**
250	6.45%	7.38%	122.05%	**0.00%**	10.23%	**0.00%**
500	3.02%	3.34%	78.43%	0.02%	4.59%	0.02%
1000	0.75%	0.88%	46.40%	0.02%	1.61%	0.91%

6. Discussion and Conclusion

In this work, we have proposed a novel multi-class queueing model to capture patient deterioration and explore the potential benefits of proactive care. We identify an important metric, the modified $c\mu/\theta$-index, which plays a critical role in specifying the optimal policy. Moreover, we find that the modified $c\mu/\theta$-index lends itself to a very intuitive interpretation. In particular, as in the case of the standard $c\mu/\theta$-index, it balances the relative importance of holding costs, service times, and abandonment rates. However, the standard index is augmented by an important additional term that accounts for patient degradation. Importantly, this additional term is characterized by the efficiency of the predictive model as measured the probability of deterioration, $\gamma/(\theta_2+\gamma)$, as well as the efficiency of preventative care as measured by the service rate of moderate patients, μ_2.

Our analysis of the transient control problems also provides important insights into the management of such systems. Specifically, when the system is near the equilibrium, cost minimization is driven by service completions. Hence, following the $c\mu$-rule is optimal. However, when the system is initialized with a very large backlog, it is essential to account for the abandonment and degradation dynamics, as captured by the modified $c\mu/\theta$-rule.

Our model and analysis is intended to provide insights into how preventative care should be utilized. In order to facilitate analysis to derive these insights, we proposed a parsimonious model of patient deterioration and preventative care. While this work is motivated by an important practical application, it is also general enough to provide insights into other service systems where customer degradation and proactive service exist. If one wants to develop prescriptive insights, e.g. to specify for a hospital exactly when to utilized preventative care, a high fidelity simulation model which incorporates features such as time-varying arrival rates, congestion dependent service-times, etc. could be valuable. While we take a common approach in the literature of defining holding and abandonment costs for different patient types (e.g. (Akan et al. 2012, Sun et al. 2017)), an important empirical question is estimating these – and other system parameters – from real data.

Our paradigm for solving the transient control problem may provide insights to the management of other queueing systems. Utilizing Pontryagin's Minimum Principle, we are able to fully characterize the structure of the optimal scheduling policy for the two-class system as well as gaining insights when there are more than two classes. However, using this approach to fully characterize the policy curves when there are more than two classes appears to be intractable.

References

Abate, J., W. Whitt. 1988. Transient behavior of the m/m/1 queue via laplace transforms. *Advances in Applied Probability* **20**(1) 145–178.

Abate, J., W. Whitt. 2006. A unified framework for numerically inverting laplace transforms. *INFORMS Journal on Computing* **18**(4) 408–421.

Akan, M., O. Alagoz, B. Ata, F. S. Erenay, A. Said. 2012. A broader view of designing the liver allocation system. *Operations research* **60**(4) 757–770.

Alam, N., E. Hobbelink, A. van Tienhoven, P. van de Ven, E. Jansma, P. Nanayakkara. 2014. The impact of the use of the early warning score (EWS) on patient outcomes: a systematic review. *Resuscitation* **85**(5) 587–594.

Atar, R., C. Giat, N. Shimkin. 2010. The cμ/θ rule for many-server queues with abandonment. *Operations Research* **58**(5) 1427–1439.

Atar, R., C. Giat, N. Shimkin. 2011. On the asymptotic optimality of the cμ/θ rule under ergodic cost. *Queueing Systems* **67**(2) 127–144.

Cao, P., J. Xie. 2016. Optimal control of a multiclass queueing system when customers can change types. *Queueing Systems* **82**(3-4) 285–313.

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**(6) 1477–1483.

Chan, C. W., V. F. Farias, G. J. Escobar. 2016. The impact of delays on service times in the intensive care unit. *Management Science* **63**(7) 2049–2072.

Chan, P. S., H. M. Krumholz, G. Nichol, B. K. Nallamothu, A. H. A. N. R. of Cardiopulmonary Resuscitation Investigators. 2008. Delayed time to defibrillation after in-hospital cardiac arrest. *New England Journal of Medicine* **358**(1) 9–17.

Chang, Y., M. Yeh, Y. Li, C. Hsu, M. Hsu, W. Chiu. 2011. Predicting hospital-acquired infections by scoring system with simple parameters. *PLoS One* **6**(8).

Cheng, G., J. Xie, Z. Zheng. 2019. Optimal stopping for medical treatment with predictive information. *Available at SSRN 3397530* .

Churpek, M., T. Yuen, S. Y. Park, D. Edelson. 2014. Using electronic health record data to develop and validate a prediction model for adverse outcomes on the wards. *Critical care medicine* **42**(4) 841.

Cox, D. R., W. Smith. 1991. *Queues*, vol. 2. CRC Press.

Cristiani, E., P. Martinon. 2010. Initialization of the shooting method via the hamilton-jacobi-bellman approach. *Journal of Optimization Theory and Applications* **146**(2) 321–346.

Delana, K., N. Savva, T. Tezcan. 2019. Proactive customer service: operational benefits and economic frictions. Forthcoming in Manufacturing & Service Operations Management.

Down, D. G., M. E. Lewis. 2010. The n-network model with upgrades. *Probability in the Engineering and Informational Sciences* **24**(2) 171–200.

Escobar, G. J., J. C. LaGuardia, B. J. Turk, A. Ragins, P. Kipnis, D. Draper. 2012. Early detection of impending physiologic deterioration among patients who are not in intensive care: Development of predictive models using data from an automated electronic medical record. *Journal of Hospital Medicine* **7**(5) 388–395.

Frankowska, H. 2010. Optimal control under state constraints. *Proceedings of the International Congress of Mathematicians 2010 (ICM 2010) (In 4 Volumes) Vol. I: Plenary Lectures and Ceremonies Vols. II–IV: Invited Lectures*. World Scientific, 2915–2942.

Fuller, A. 1963. Study of an optimum non-linear control system. *International Journal of Electronics* **15**(1) 63–71.

Grass, D., J. P. Caulkins, G. Feichtinger, G. Tragler, D. A. Behrens. 2008. Optimal control of nonlinear processes. *Berlino: Springer* .

Gurvich, I., W. Whitt. 2010. Service-level differentiation in many-server service systems via queue-ratio routing. *Manufacturing and Service Operations Management* **58**(2) 237–253.

Harrison, J. M., A. Zeevi. 2004. Dynamic scheduling of a multiclass queue in the halfin-whitt heavy traffic regime. *Operations Research* **52**(2) 243–257.

Hartl, R. F., S. P. Sethi, R. G. Vickson. 1995. A survey of the maximum principles for optimal control problems with state constraints. *SIAM review* **37**(2) 181–218.

Honnappa, H., J. Jain, A. Ward. 2015. A queueing model with independent arrivals, and its fluid and diffusion limits. *Queueing Systems* **80**(1-2) 71–103.

Hu, W., C. W. Chan, J. R. Zubizarreta, G. J. Escobar. 2018. An examination of early transfers to the icu based on a physiologic risk score. *Manufacturing & Service Operations Management* **20**(3) 531–549.

Iglehart, D. L., W. Whitt. 1970. Multiple channel queues in heavy traffic. ii: sequences, networks, and batches. *Advances in Applied Probability* **2**(2) 355–369.

Larrañaga, M. 2015. Dynamic control of stochastic and fluid resource-sharing systems. Ph.D. thesis.

Larrañaga, M., U. Ayesta, I. M. Verloop. 2013. Dynamic fluid-based scheduling in a multi-class abandonment queue. *Performance Evaluation* **70**(10) 841–858.

Mandelbaum, A., A. Stolyar. 2004. Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized cμ-rule. *Operations Research* **52**(6) 836–855.

McCall, J. J. 1965. Maintenance policies for stochastically failing equipment: a survey. *Management science* **11**(5) 493–524.

Örmeci, E. L., E. D. Güneş, D. Kunduzcu. 2015. A modeling framework for control of preventive services. *Manufacturing & Service Operations Management* **18**(2) 227–244.

Özekici, S., S. R. Pliska. 1991. Optimal scheduling of inspections: A delayed markov model with false positives and negatives. *Operations Research* **39**(2) 261–273.

Perry, O., W. Whitt. 2009. Responding to unexpected overloads in large-scale service systems. *Management Science* **55**(8) 1353–1367.

Pierskalla, W. P., J. A. Voelker. 1976. A survey of maintenance models: the control and surveillance of deteriorating systems. *Naval Research Logistics Quarterly* **23**(3) 353–388.

Puha, A., A. Ward. 2019. Tutorial paper: Scheduling an overloaded multiclass many-server queue with impatient customers. Working paper.

Reed, J. E., A. Ward. 2008. Approximating the GI/GI/1+GI queue with a nonlinear drift diffusion: Hazard rate scaling in heavy traffic. *Mathematics of Operations Research* **33**(3) 606–644.

Rumsfeld, J., K. Joynt, T. Maddox. 2016. Big data analytics to improve cardiovascular care: promise and challenges. *Nature Reviews Cardiology* **13** 350–359.

Schättler, H., U. Ledzewicz. 2012. *Geometric optimal control: theory, methods and examples*, vol. 38. Springer Science & Business Media.

Sethi, S. P., G. L. Thompson. 2000. *Optimal control theory: Applications to management science and economics*. Springer.

Stolyar, A. L., et al. 2004. Maxweight scheduling in a generalized switch: State space collapse and workload minimization in heavy traffic. *The Annals of Applied Probability* **14**(1) 1–53.

Sun, Z., N. T. Argon, S. Ziya. 2017. Patient triage and prioritization under austere conditions. *Management Science* **64**(10) 4471–4489.

Tezcan, T., J. Dai. 2010. Dynamic control of n-systems with many servers: Asymptotic optimality of a static priority policy in heavy traffic. *Operations Research* **58**(1) 94–110.

Trélat, E. 2012. Optimal control and applications to aerospace: some results and challenges. *Journal of Optimization Theory and Applications* **154**(3) 713–758.

Van Mieghem, J. A. 1995. Dynamic scheduling with convex delay costs: The generalized c— mu rule. *The Annals of Applied Probability* 809–833.

Whitt, W. 2002. *Stochastic-Process Limits*. Springer, New York.

Xie, J., T. Zhu, A.-K. Chao, S. Wang. 2017. Performance analysis of service systems with priority upgrades. *Annals of Operations Research* **253**(1) 683–705.

Xu, K., C. W. Chan. 2016. Using future information to reduce waiting times in the emergency department via diversion. *Manufacturing & Service Operations Management* **18**(3) 314–331.

Yom-Tov, G. B., Y. Xie, L. Yedidsion. 2018. An invitation control policy for proactive service systems: Balancing efficiency, value and service level. Tech. rep., working paper. ec1.

Yoon, J., A. Alaa, S. Hu, M. Schaar. 2016. Forecasticu: a prognostic decision support system for timely prediction of intensive care unit admission. *International Conference on Machine Learning*. 1680–1689.

Appendix A: Proofs in Section 3

A.1. Proof of Theorem 1

PROOF: Consider the equivalent LP formulation (5) of the long-run average optimization problem (F1). For any given set of parameters, we first solve the LP (5) to obtain an optimal solution (z_1^{e*}, z_2^{e*}) which represents the optimal long-run average service capacity allocated to Class 1 and 2. The LP can be solved in closed form to specify the optimal solution (z_1^{e*}, z_2^{e*}) by noting that the optimal solution must be on one of the vertices of the feasible region. We then show that (z_1^{e*}, z_2^{e*}) and the corresponding (q_1^{e*}, q_2^{e*}) is the globally asymptotically stable equilibrium under the modified $c\mu/\theta$-rule, which corresponds to P_1 or P_2 depending on which class has a higher modified $c\mu/\theta$-index. This step is based on the stability analysis for P_1 and P_2, which are given in Theorems 2 and 3. Following similar parameter regimes examined in the stability analysis, we divide the analysis here into different cases. For each case, the three tables below list the optimal LP solution (z_1^{e*}, z_2^{e*}) and the corresponding static control under which (z_1^{e*}, z_2^{e*}) is a globally asymptotically stable equilibrium for allocation.

Case 1. $\mu_1 > \frac{\gamma}{\theta_2+\gamma}\mu_2$, and $\frac{\lambda_2}{\mu_2} \le \frac{\lambda_1}{\mu_1} + \frac{\gamma}{\theta_2+\gamma}\frac{\lambda_2}{\mu_1}$.

	Modified $c\mu/\theta$-rule prioritizes Class 1		Modified $c\mu/\theta$-rule prioritizes Class 2	
	(z_1^{e*}, z_2^{e*})	Control	(z_1^{e*}, z_2^{e*})	Control
$\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} \le s$	$\left(\frac{\lambda_1}{\mu_1}, \frac{\lambda_2}{\mu_2}\right)$	P_1, P_2	$\left(\frac{\lambda_1}{\mu_1}, \frac{\lambda_2}{\mu_2}\right)$	P_1, P_2
$\frac{\lambda_1}{\mu_1} + \frac{\gamma}{\theta_2+\gamma}\frac{\lambda_2}{\mu_1} \le s < \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2}$	$\left(\frac{\lambda_1\theta_2+\gamma(\lambda_1+\lambda_2-s\mu_2)}{(\theta_2+\gamma)\mu_1-\gamma\mu_2}, s-\frac{\lambda_1\theta_2+\gamma(\lambda_1+\lambda_2-s\mu_2)}{(\theta_2+\gamma)\mu_1-\gamma\mu_2}\right)$	P_1	$\left(s-\frac{\lambda_2}{\mu_2}, \frac{\lambda_2}{\mu_2}\right)$	P_2
$\frac{\lambda_2}{\mu_2} \le s < \frac{\lambda_1}{\mu_1} + \frac{\gamma}{\theta_2+\gamma}\frac{\lambda_2}{\mu_1}$	$(s, 0)$	P_1	$\left(s-\frac{\lambda_2}{\mu_2}, \frac{\lambda_2}{\mu_2}\right)$	P_2
$s < \frac{\lambda_2}{\mu_2}$	$(s, 0)$	P_1	$(0, s)$	P_2

Case 2. $\mu_1 > \frac{\gamma}{\theta_2+\gamma}\mu_2$, and $\frac{\lambda_2}{\mu_2} > \frac{\lambda_1}{\mu_1} + \frac{\gamma}{\theta_2+\gamma}\frac{\lambda_2}{\mu_1}$.

	Modified $c\mu/\theta$-rule prioritizes Class 1		Modified $c\mu/\theta$-rule prioritizes Class 2	
	(z_1^{e*}, z_2^{e*})	Control	(z_1^{e*}, z_2^{e*})	Control
$\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} \le s$	$\left(\frac{\lambda_1}{\mu_1}, \frac{\lambda_2}{\mu_2}\right)$	P_1, P_2	$\left(\frac{\lambda_1}{\mu_1}, \frac{\lambda_2}{\mu_2}\right)$	P_1, P_2
$\frac{\lambda_2}{\mu_2} \le s < \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2}$	$\left(\frac{\lambda_1\theta_2+\gamma(\lambda_1+\lambda_2-s\mu_2)}{(\theta_2+\gamma)\mu_1-\gamma\mu_2}, s-\frac{\lambda_1\theta_2+\gamma(\lambda_1+\lambda_2-s\mu_2)}{(\theta_2+\gamma)\mu_1-\gamma\mu_2}\right)$	P_1	$\left(s-\frac{\lambda_2}{\mu_2}, \frac{\lambda_2}{\mu_2}\right)$	P_2
$\frac{\lambda_1}{\mu_1} + \frac{\gamma}{\theta_2+\gamma}\frac{\lambda_2}{\mu_1} \le s < \frac{\lambda_2}{\mu_2}$	$\left(\frac{\lambda_1\theta_2+\gamma(\lambda_1+\lambda_2-s\mu_2)}{(\theta_2+\gamma)\mu_1-\gamma\mu_2}, s-\frac{\lambda_1\theta_2+\gamma(\lambda_1+\lambda_2-s\mu_2)}{(\theta_2+\gamma)\mu_1-\gamma\mu_2}\right)$	P_1	$(0, s)$	P_2
$s < \frac{\lambda_1}{\mu_1} + \frac{\gamma}{\theta_2+\gamma}\frac{\lambda_2}{\mu_1}$	$(s, 0)$	P_1	$(0, s)$	P_2

Case 3. $\mu_1 < \frac{\gamma}{\theta_2+\gamma}\mu_2$. In this case, the modified $c\mu/\theta$-rule prioritizes Class 2.

	(z_1^{e*}, z_2^{e*})	Control
$\frac{\lambda_1}{\mu_1} + \frac{\gamma}{\theta_2+\gamma}\frac{\lambda_2}{\mu_1} < s$	$\left(\frac{\lambda_1}{\mu_1}, \frac{\lambda_2}{\mu_2}\right)$	P_1, P_2
$\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} \le s \le \frac{\lambda_1}{\mu_1} + \frac{\gamma}{\theta_2+\gamma}\frac{\lambda_2}{\mu_1}$	$\left(\frac{\lambda_1}{\mu_1}, \frac{\lambda_2}{\mu_2}\right)$	P_2
$\frac{\lambda_2}{\mu_2} \le s < \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2}$	$\left(s - \frac{\lambda_2}{\mu_2}, \frac{\lambda_2}{\mu_2}\right)$	P_2
$s < \frac{\lambda_2}{\mu_2}$	$(0, s)$	P_2

Q.E.D.

A.2. Proof of Theorem 2

The stability analysis for P_1 in Theorem 2 divides the parameter regime into six cases. In each case, we construct a Lyapunov function to establish the asymptotic stability. As the proof for each case follows exactly the same line of analysis, we only present the proof for Case Ia which has a globally asymptotically stable equilibrium and Case IIb which has two locally asymptotically stable equilibria. The Lyapunov function we utilize to prove each case differs; they are summarized in the table below.

	Lyapunov function		Lyapunov function
Case Ia (see proof)	$\frac{1}{\mu_1}\lvert q_1 - q_1^e\rvert + \frac{1}{\mu_2}\lvert q_2 - q_2^e\rvert$	Case IIa	$\lvert q_1 - q_1^e\rvert + \frac{\gamma}{\theta_2+\gamma}\lvert q_2 - q_2^e\rvert$
Case Ib	$\frac{1}{\mu_1}\lvert q_1 - q_1^e\rvert + \frac{1}{\mu_2}\lvert q_2 - q_2^e\rvert$	Case IIb (see proof)	Local equilibrium $(0,0)$: $\frac{1}{\mu_1}\lvert q_1 - q_1^e\rvert + \frac{1}{\mu_2}\lvert q_2 - q_2^e\rvert$ Local equilibrium $\left(\frac{\lambda_1 + \frac{\gamma}{\theta_2+\gamma}\lambda_2 - s\mu_1}{\theta_1}, \frac{\lambda_2}{\theta_2+\gamma}\right)$: $\lvert q_1 - q_1^e\rvert + \lvert q_2 - q_2^e\rvert$
Case Ic	$\lvert q_1 - q_1^e\rvert + \lvert q_2 - q_2^e\rvert$	Case IIc and IId	$\lvert q_1 - q_1^e\rvert + \lvert q_2 - q_2^e\rvert$

PROOF: Let V denote the Lyapunov function we constructed. To prove the asymptotic stability of an equilibrium point q^e, we need to verify that 1) $V(q^e) = 0$ and $V(q) \to \infty$ as $||q|| \to \infty$; 2) $\nabla_q V(q)^T f(q) < 0$ for $q \neq q^e$. In the case of local stability, the second condition is checked locally with q restricted to be in some neighborhood of q_e, i.e. $0 < ||q - q^e|| < \delta$ for some $\delta > 0$. As 1) is straightforward from our definition of the Lyapunov function, we focus on verifying 2) only.

Case I. $\frac{\gamma}{\theta_2+\gamma}\mu_2 < \mu_1$**,** i.e., $\frac{\gamma}{\mu_1} - \frac{\theta_2+\gamma}{\mu_2} < 0$.

Ia. If $\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} \le s$**,** consider Lyapunov function of the form

$$V(q) = \frac{1}{\mu_1}|q_1 - q_1^e| + \frac{1}{\mu_2}|q_2 - q_2^e|,$$

where (q_1^e, q_2^e) is the corresponding equilibrium point $(0,0)$.

(i) If $q_1(t) > 0$,

$$dq_1(t) = \lambda_1 - \mu_1 s - \theta_1 q_1(t) + \gamma q_2(t)$$
$$dq_2(t) = \lambda_2 - \theta_2 q_2(t) - \gamma q_2(t).$$

$$\begin{aligned}\nabla_q V(q)^T f(q) &= \frac{1}{\mu_1}(\lambda_1 - \mu_1 s - \theta_1 q_1(t) + \gamma q_2(t)) + \frac{1}{\mu_2}(\lambda_2 - \theta_2 q_2(t) - \gamma q_2(t)) \\ &= \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} - s - \frac{\theta_1}{\mu_1} q_1(t) + \left(\frac{\gamma}{\mu_1} - \frac{\theta_2+\gamma}{\mu_2}\right) q_2(t) \\ &< 0,\end{aligned}$$

where the last inequality follows from the facts that $\lambda_1/\mu_1 + \lambda_2/\mu_2 \le s$, $q_1(t) > 0$, $\gamma/\mu_1 - (\theta_2+\gamma)/\mu_2 < 0$, and $q_2(t) \ge 0$.

(ii) If $q_1(t) = 0$, $q_2(t) > 0$,

(iia) if $\frac{\lambda_1 + \gamma q_2(t)}{\mu_1} \geq s$,

$$dq_1(t) = \lambda_1 - \mu_1 s + \gamma q_2(t)$$
$$dq_2(t) = \lambda_2 - \theta_2 q_2(t) - \gamma q_2(t).$$

$$\begin{aligned}\nabla_q V(q)^T f(q) &= \frac{1}{\mu_1}(\lambda_1 - \mu_1 s + \gamma q_2(t)) + \frac{1}{\mu_2}(\lambda_2 - \theta_2 q_2(t) - \gamma q_2(t)) \\ &= \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} - s + \left(\frac{\gamma}{\mu_1} - \frac{\theta_2 + \gamma}{\mu_2}\right) q_2(t) \\ &< 0,\end{aligned}$$

where the last inequality follows from the following facts: $\lambda_1/\mu_1 + \lambda_2/\mu_2 \leq s$, $\gamma/\mu_1 - (\theta_2 + \gamma)/\mu_2 < 0$, and $q_2(t) > 0$.

(iib) If $\frac{\lambda_1 + \gamma q_2(t)}{\mu_1} < s$,

$$dq_1(t) = \lambda_1 - \mu_1 \left(\frac{\lambda_1 + \gamma q_2(t)}{\mu_1}\right) + \gamma q_2(t) = 0$$
$$dq_2(t) = \lambda_2 - \mu_2 \left(s - \frac{\lambda_1 + \gamma q_2(t)}{\mu_1}\right) - \theta_2 q_2(t) - \gamma q_2(t).$$

$$\begin{aligned}\nabla_q V(q)^T f(q) &= \frac{1}{\mu_2}\left(\lambda_2 - \mu_2 \left(s - \frac{\lambda_1 + \gamma q_2(t)}{\mu_1}\right) - \theta_2 q_2(t) - \gamma q_2(t)\right) \\ &= \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} - s + \left(\frac{\gamma}{\mu_1} - \frac{\theta_2 + \gamma}{\mu_2}\right) q_2(t) \\ &< 0,\end{aligned}$$

where the last inequality follows from the following facts: $\lambda_1/\mu_1 + \lambda_2/\mu_2 \leq s$, $\gamma/\mu_1 - (\theta_2 + \gamma)/\mu_2 < 0$, and $q_2(t) > 0$. Hence, in Case Ia, the system has a globally asymptotic equilibrium at $(0, 0)$.

Case II. $\frac{\gamma}{\theta_2 + \gamma}\mu_2 > \mu_1$**,** i.e., $\frac{\gamma}{\mu_1} - \frac{\theta_2 + \gamma}{\mu_2} > 0$.

IIb. If $\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} < s \leq \frac{\lambda_1}{\mu_1} + \frac{\gamma}{\theta_2 + \gamma}\frac{\lambda_2}{\mu_1}$**,** to check for local stability, it is sufficient to find a Lyapunov function V that satisfies $\nabla_q V(q)^T f(q) < 0$ in an open neighborhood of the equilibrium point. We construct different Lyapunov functions for different equilibrium points.

(A) Local stability of $(q_1^e, q_2^e) = (0, 0)$: Consider Lyapunov function of the form

$$V(q) = \frac{1}{\mu_1}|q_1 - q_1^e| + \frac{1}{\mu_2}|q_2 - q_1^e|.$$

Let $0 < \epsilon < (s\mu_1 - \lambda_1)/\gamma$ be such that

$$\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} - s + \left(\frac{\gamma}{\mu_1} - \frac{\theta_2 + \gamma}{\mu_2}\right)\epsilon < 0.$$

We know such ϵ exists because $\lambda_1/\mu_1 + \lambda_2/\mu_2 - s < 0$ and $\gamma/\mu_1 - (\theta_2 + \gamma)/\mu_2 > 0$. Consider states (q_1, q_2) with $q_2 < \epsilon$.

(i) If $q_1(t) > 0$,

$$dq_1(t) = \lambda_1 - \mu_1 s - \theta_1 q_1(t) + \gamma q_2(t)$$
$$dq_2(t) = \lambda_2 - \theta_2 q_2(t) - \gamma q_2(t)$$

$$\begin{aligned}\nabla_q V(q)^T f(q) &= \frac{1}{\mu_1}\left(\lambda_1 - \mu_1 s - \theta_1 q_1(t) + \gamma q_2(t)\right) + \frac{1}{\mu_2}\left(\lambda_2 - \theta_2 q_2(t) - \gamma q_2(t)\right) \\ &= \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} - s + \left(\frac{\gamma}{\mu_1} - \frac{\theta_2+\gamma}{\mu_2}\right) q_2(t) - \frac{\theta_1}{\mu_1} q_1(t) \\ &< \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} - s + \left(\frac{\gamma}{\mu_1} - \frac{\theta_2+\gamma}{\mu_2}\right)\epsilon - \frac{\theta_1}{\mu_1} q_1(t) \\ &< 0.\end{aligned}$$

(ii) If $q_1(t) = 0$, $q_2(t) > 0$,

$$\begin{aligned}dq_1(t) &= \lambda_1 - \mu_1\left(\frac{\lambda_1 + \gamma q_2(t)}{\mu_1}\right) + \gamma q_2(t) = 0 \\ dq_2(t) &= \lambda_2 - \mu_2\left(s - \frac{\lambda_1 + \gamma q_2(t)}{\mu_1}\right) - \theta_2 q_2(t) - \gamma q_2(t) \\ \nabla_q V(q)^T f(q) &= \frac{1}{\mu_2}\left(\lambda_2 - \mu_2\left(s - \frac{\lambda_1 + \gamma q_2(t)}{\mu_1}\right) - \theta_2 q_2(t) - \gamma q_2(t)\right) \\ &= \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} - s + \left(\frac{\gamma}{\mu_1} - \frac{\theta_2+\gamma}{\mu_2}\right) q_2(t) \\ &< \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} - s + \left(\frac{\gamma}{\mu_1} - \frac{\theta_2+\gamma}{\mu_2}\right)\epsilon \\ &< 0.\end{aligned}$$

(B) Local stability of $(q_1^e, q_2^e) = \left(\frac{\lambda_1 + \frac{\gamma}{\theta_2+\gamma}\lambda_2 - s\mu_1}{\theta_1}, \frac{\lambda_2}{\theta_2+\gamma}\right)$: Consider Lyapunov function of the form

$$V(q) = |q_1 - q_1^e| + |q_2 - q_2^e|.$$

Consider states q such that $q_1 > 0$ and $q_2 > 0$. In this case, the dynamics of q_1 and q_2 follow

$$\begin{aligned}dq_1(t) &= \lambda_1 - \mu_1 s - \theta_1 q_1(t) + \gamma q_2(t) \\ dq_2(t) &= \lambda_2 - \theta_2 q_2(t) - \gamma q_2(t).\end{aligned}$$

(i) If $q_1(t) \geq q_1^e$, $q_2(t) \geq q_2^e$, and $(q_1(t), q_2(t)) \neq (q_1^e, q_2^e)$,

$$\begin{aligned}\nabla_q V(q)^T f(q) &= \lambda_1 - \mu_1 s - \theta_1 q_1(t) + \gamma q_2(t) + \lambda_2 - \theta_2 q_2(t) - \gamma q_2(t) \\ &= \lambda_1 - \mu_1 s - \theta_1 q_1(t) + \lambda_2 - \theta_2 q_2(t) \\ &< \lambda_1 - \mu_1 s - \theta_1 q_1^e + \lambda_2 - \theta_2 q_2^e \\ &= 0.\end{aligned}$$

(ii) If $q_1(t) \geq q_1^e$, $q_2(t) < q_2^e$,

$$\begin{aligned}\nabla_q V(q)^T f(q) &= \lambda_1 - \mu_1 s - \theta_1 q_1(t) + \gamma q_2(t) - \lambda_2 + \theta_2 q_2(t) + \gamma q_2(t) \\ &= \lambda_1 - \mu_1 s - \theta_1 q_1(t) - \lambda_2 + (2\gamma + \theta_2) q_2(t) \\ &< \lambda_1 - \mu_1 s - \theta_1 q_1^e - \lambda_2 + (2\gamma + \theta_2) q_2^e \\ &= 0.\end{aligned}$$

(iii) If $q_1(t) < q_1^e$, $q_2(t) \geq q_2^e$,

$$\begin{aligned}\nabla_q V(q)^T f(q) &= -\lambda_1 + \mu_1 s + \theta_1 q_1(t) - \gamma q_2(t) + \lambda_2 - \theta_2 q_2(t) - \gamma q_2(t) \\ &= -\lambda_1 + \mu_1 s + \theta_1 q_1(t) + \lambda_2 - (2\gamma + \theta_2) q_2(t) \\ &< -\lambda_1 + \mu_1 s + \theta_1 q_1^e + \lambda_2 - (2\gamma + \theta_2) q_2^e \\ &= 0.\end{aligned}$$

(iv) If $q_1(t) < q_1^e$, $q_2(t) < q_2^e$,

$$\begin{aligned}\nabla_q V(q)^T f(q) &= -\lambda_1 + \mu_1 s + \theta_1 q_1(t) - \gamma q_2(t) - \lambda_2 + \theta_2 q_2(t) + \gamma q_2(t) \\ &= -\lambda_1 + \mu_1 s + \theta_1 q_1(t) - \lambda_2 + \theta_2 q_2(t) \\ &< -\lambda_1 + \mu_1 s + \theta_1 q_1^e - \lambda_2 + \theta_2 q_2^e \\ &= 0.\end{aligned}$$

Hence, in Case IIb, the system has two locally asymptotic equilibria at $(0,0)$ and $\left(\frac{\lambda_1 + \frac{\gamma}{\theta_2+\gamma}\lambda_2 - s\mu_1}{\theta_1}, \frac{\lambda_2}{\theta_2+\gamma}\right)$. Q.E.D.

A.3. Proof of Theorem 3

The stability analysis for P_2 in Theorem 3 divides the parameter regime into three cases. Similar to the proof of Theorem 2 in Appendix A.2, Lyapunov functions are constructed to prove the result. Since the dynamic of q_2 evolves independently of q_1 under P_2, the analysis can be done by first proving that q_2 possesses a globally asymptotically stable equilibrium q_2^e, and then constructing another Lyapunov function to prove the global asymptotic stability of (q_1^e, q_2^e) for the joint process (q_1, q_2) when $q_2(0)$ is restricted to a neighborhood around q_2^e. As the proof for each case follows similar lines of analysis, we present the details for Case A only. The Lyapunov function we constructed for each case is listed in the table below.

	Lyapunov function for q_2	Lyapunov function for (q_1, q_2) with q_2 close to q_2^e
Case A (see proof)	$\lvert q_2 - q_2^e\rvert$	$\frac{1}{\mu_1}\lvert q_1 - q_1^e\rvert + \frac{1}{\mu_2+\epsilon}\lvert q_2 - q_2^e\rvert$, $\epsilon \geq 0$
Case B	$\lvert q_2 - q_2^e\rvert$	$\frac{1}{\mu_1}\lvert q_1 - q_1^e\rvert + \frac{1}{\mu_2}\lvert q_2 - q_2^e\rvert$
Case C	$\lvert q_2 - q_2^e\rvert$	$\lvert q_1 - q_1^e\rvert + \lvert q_2 - q_2^e\rvert$

PROOF: Let V denote the selected Lyapunov function. We shall verify that $\nabla_q V(q)^T f(q) < 0$ for $q \neq q^e$.

Case A. $\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} \leq s$. First, independently of q_1, the dynamics of q_2 are characterized by

$$dq_2(t) = \begin{cases} \lambda_2 - \mu_2 s - \theta_2 q_2(t) - \gamma q_2(t) & \text{if } q_2(t) > 0 \\ \lambda_2 - \mu_2 s & \text{if } q_2(t) = 0. \end{cases}$$

Indeed, q_2 has a globally asymptotically stable equilibrium at 0. To see this, consider Lyapunov function for q_2 of the form

$$V(q_2) = |q_2 - 0|.$$

If $q_2(t) > 0$, we have

$$\begin{aligned} dq_2(t) &= \lambda_2 - s\mu_2 - (\theta_2 + \gamma) q_2(t) \\ \nabla_q V(q_2)^T f(q_2) &= \mu_2 \left(\frac{\lambda_2}{\mu_2} - s - \frac{\gamma + \theta_2}{\mu_2} q_2(t) \right) < 0, \end{aligned}$$

where the inequality follows from the fact that $\lambda_2/\mu_2 < \lambda_1/\mu_1 + \lambda_2/\mu_2 \leq s$. Hence, q_2 has a globally asymptotically stable equilibrium at 0.

Next, we show that (q_1, q_2) jointly has a globally asymptotically stable equilibrium at $(0,0)$.

(i) If $\lambda_1/\mu_1 + \lambda_2/\mu_2 < s$ or $\gamma/\mu_1 - (\theta_2 + \gamma)/\mu_2 < 0$, consider a Lyapunov function of the form

$$V(q) = \frac{1}{\mu_1}|q_1 - 0| + \frac{1}{\mu_2}|q_2 - 0|.$$

Let $\delta > 0$ be such that

$$\left(\frac{\gamma}{\mu_1} - \frac{\theta_2+\gamma}{\mu_2}\right)\delta + \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} - s < 0.$$

This is true for all $\delta > 0$ if $\gamma/\mu_1 < (\theta_2+\gamma)/\mu_2$. If $\gamma/\mu_1 > (\theta_2+\gamma)/\mu_2$ and $\lambda_1/\mu_1 + \lambda_2/\mu_2 < s$, δ will need to be made sufficiently small. We have already shown that for any $q_2(0)$ and $\delta > 0$, there exists t_δ such that for all $t > t_\delta$, $q_2(t) < \delta$. We will focus on such t.

(ia) If $q_2(t) > 0$,

$$\begin{aligned}
dq_1(t) &= \lambda_1 - \theta_1 q_1(t) + \gamma q_2(t) \\
dq_2(t) &= \lambda_2 - \mu_2 s - \theta_2 q_2(t) - \gamma q_2(t) \\
\nabla_q V(q)^T f(q) &= \frac{1}{\mu_1}\left(\lambda_1 - \theta_1 q_1(t) + \gamma q_2(t)\right) + \frac{1}{\mu_2}\left(\lambda_2 - \mu_2 s - \theta_2 q_2(t) - \gamma q_2(t)\right) \\
&= \left(\frac{\gamma}{\mu_1} - \frac{\theta_2+\gamma}{\mu_2}\right) q_2(t) - \frac{\theta_1}{\mu_1} q_1(t) + \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} - s \\
&< \left(\frac{\gamma}{\mu_1} - \frac{\theta_2+\gamma}{\mu_2}\right)\delta - \frac{\theta_1}{\mu_1} q_1(t) + \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} - s \\
&< 0.
\end{aligned}$$

(ib) If $q_2(t) = 0$, $q_1(t) > 0$,

$$\begin{aligned}
dq_1(t) &= \lambda_1 - \mu_1(s - \lambda_2/\mu_2) - \theta_1 q_1(t) + \gamma q_2(t) \\
dq_2(t) &= \lambda_2 - \mu_2(\lambda_2/\mu_2) = 0 \\
\nabla_q V(q)^T f(q) &= \frac{1}{\mu_1}\left(\lambda_1 - \mu_1\left(s - \frac{\lambda_2}{\mu_2}\right) - \theta_1 q_1(t)\right) \\
&= -\frac{\theta_1}{\mu_1} q_1(t) + \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} - s \\
&< 0.
\end{aligned}$$

(ii) If $\lambda_1/\mu_1 + \lambda_2/\mu_2 = s$ and $\gamma/\mu_1 - (\theta_2+\gamma)/\mu_2 > 0$, consider a Lyapunov function of the form

$$V(q) = \frac{1}{\mu_1}|q_1 - 0| + \frac{1}{\mu_2+\epsilon}|q_2 - 0|,$$

where $\epsilon > 0$ is chosen such that $\lambda_1/\mu_1 + \lambda_2/(\mu_2+\epsilon) - s < 0$. Now, let $\delta > 0$ be sufficiently small such that

$$\left(\frac{\gamma}{\mu_1} - \frac{\theta_2+\gamma}{\mu_2+\epsilon}\right)\delta + \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2+\epsilon} - s < 0.$$

Similar to Case (i), we will focus on t such that $q_2(t) < \delta$.

(iia) If $q_2(t) > 0$,

$$\begin{aligned}
dq_1(t) &= \lambda_1 - \theta_1 q_1(t) + \gamma q_2(t) \\
dq_2(t) &= \lambda_2 - \mu_2 s - \theta_2 q_2(t) - \gamma q_2(t) \\
\nabla_q V(q)^T f(q) &= \frac{1}{\mu_1}\left(\lambda_1 - \theta_1 q_1(t) + \gamma q_2(t)\right) + \frac{1}{\mu_2+\epsilon}\left(\lambda_2 - \mu_2 s - \theta_2 q_2(t) - \gamma q_2(t)\right) \\
&= \left(\frac{\gamma}{\mu_1} - \frac{\theta_2+\gamma}{\mu_2+\epsilon}\right) q_2(t) - \frac{\theta_1}{\mu_1} q_1(t) + \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2+\epsilon} - s \\
&< \left(\frac{\gamma}{\mu_1} - \frac{\theta_2+\gamma}{\mu_2+\epsilon}\right)\delta - \frac{\theta_1}{\mu_1} q_1(t) + \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2+\epsilon} - s \\
&< 0.
\end{aligned}$$

(iib) If $q_2(t)=0,\ q_1(t)>0$,

$$\begin{aligned}
dq_1(t) &= \lambda_1-\mu_1(s-\lambda_2/\mu_2)-\theta_1 q_1(t)+\gamma q_2(t)\\
dq_2(t) &= \lambda_2-\mu_2(\lambda_2/\mu_2)=0\\
\nabla_q V(q)^T f(q) &= \frac{1}{\mu_1}\left(\lambda_1-\mu_1\left(s-\frac{\lambda_2}{\mu_2}\right)-\theta_1 q_1(t)\right)\\
&= -\frac{\theta_1}{\mu_1}q_1(t)+\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_2}-s\\
&< 0.
\end{aligned}$$

Hence, in Case A, the system has a globally asymptotic equilibrium at $(0,0)$. Q.E.D.

Appendix B: Proofs in Section 4

The proofs in this section are organized as follows. We start by showing that it is without loss of optimality to rule out chattering behavior (Lemma 1). We then establish the optimal scheduling policy when we are close to the equilibrium (Proposition 1). Both proofs are based on solving the state trajectory q directly. Second, we use Pontryagin's Minimum Principle and Proposition 1 to prove Proposition 3, which are then used to prove Proposition 2. Lastly, we characterize the policy curve from P_1 to P_2 (Proposition 4) using Pontryagin's Minimum Principle and Theorem 4.

B.1. Proof of Lemma 1

PROOF: We prove the lemma by showing that the cost difference between a chattering trajectory and a properly constructed trajectory without chattering is negligible.

Consider an interval $I_1:=[0,\epsilon]$ where queue 1 is initiated at zero and receives no service capacity for an $\epsilon>0$ amount of time. During this interval, a queue accumulates in queue 1. Following I_1, $I_2=(\epsilon,\epsilon+\epsilon']$ is an interval of length $\epsilon'>0$, over which queue 1 receives full service capacity s and is eventually emptied at the end of I_2. Suppose q_2 is initiated at level $q_2(0)=q_{20},\ q_{20}\in\mathbb{R}_+$. We compute the state trajectory and cost over I_1 and I_2, i.e. $[0,\epsilon+\epsilon']$.

Over the first interval I_1, the state trajectories evolve as

$$\begin{aligned}
q_1(t) &= (q_{20}\gamma+\lambda_1)\,t+o(\epsilon),\quad t\in[0,\epsilon]\\
q_2(t) &= q_{20}+(-q_{20}\gamma-q_{20}\theta_2+\lambda_2-s\mu_2)\,t+o(\epsilon),\quad t\in[0,\epsilon]\,.
\end{aligned}$$

Note that it is possible to ignore the boundary condition that $q_2(t)\geq 0$ for sufficiently small ϵ. At time ϵ, the end of time interval I_1, the length of q_1 and q_2 are

$$\begin{aligned}
q_1(\epsilon) &= (q_{20}\gamma+\lambda_1)\,\epsilon+o(\epsilon)\\
q_2(\epsilon) &= q_{20}+(-q_{20}\gamma-q_{20}\theta_2+\lambda_2-s\mu_2)\,\epsilon+o(\epsilon).
\end{aligned}$$

Using $(q_1(\epsilon),q_2(\epsilon))$ as the initial condition at the beginning of the interval I_2, we can characterize the trajectory of q_1 and q_2 over I_2 as

$$\begin{aligned}
q_1(t) &= (q_{20}\gamma\epsilon+\epsilon\lambda_1)+\big(q_{20}\gamma-q_{20}\gamma^2\epsilon-q_{20}\gamma\epsilon\theta_1-q_{20}\gamma\epsilon\theta_2+\lambda_1-\epsilon\theta_1\lambda_1+\gamma\epsilon\lambda_2\\
&\quad -s\mu_1-s\gamma\epsilon\mu_2\big)\,(t-\epsilon)+o(\epsilon),\quad t\in[\epsilon,\epsilon+\epsilon']\\
q_2(t) &= q_{20}\,(-1+(t-\epsilon)(\theta_2+\gamma))\,(-1+\epsilon(\theta_2+\gamma))+(t-(t-\epsilon)\epsilon(\theta_2+\gamma))\,\lambda_2\\
&\quad +s\epsilon\,(-1+(t-\epsilon)(\theta_2+\gamma))\,\mu_2+o(\epsilon),\quad t\in[\epsilon,\epsilon+\epsilon']\,.
\end{aligned}$$

In addition, the value of ϵ', the time it takes to empty queue 1 from initial level $q_1(\epsilon)$, is

$$\epsilon' = \frac{\epsilon\,(q_{20}\gamma + \lambda_1)}{-q_{20}\gamma - \lambda_1 + s\mu_1} + o(\epsilon).$$

The total holding cost over the two intervals I_1 and I_2, where q_1 follows the triangular-type trajectory, is given by

$$C = c_1 \int_0^{\epsilon+\epsilon'} q_1(t)dt + c_2 \int_0^{\epsilon+\epsilon'} q_2(t)dt$$

In contrast, we now consider an interval with the same length, $\epsilon + \epsilon'$, and the same initial condition $(\tilde{q}_1(0), \tilde{q}_2(0)) = (0, q_{20})$. Now, instead of having q_1 increase from zero and then decrease to zero, we assign strict priority to Class 1 and maintain $\tilde{q}_1$ at zero. The rest of the service capacity is allocated to Class 2. Similarly, we characterize the corresponding state trajectory over this interval of length $\epsilon + \epsilon'$ as

$$\tilde{q}_1(t) = 0, \quad t \in [0, \epsilon + \epsilon']$$

$$\tilde{q}_2(t) = q_{20} + (-q_{20}\gamma\mu_1 - q_{20}\theta_2\mu_1 + \lambda_2\mu_1 + q_{20}\gamma\mu_2 + \lambda_1\mu_2 - s\mu_1\mu_2)\, t/\mu_1 + o(\epsilon), \quad t \in [0, \epsilon + \epsilon'],$$

and the total holding cost as

$$\tilde{C} = c_1 \int_0^{\epsilon+\epsilon'} \tilde{q}_1(t)dt + c_2 \int_0^{\epsilon+\epsilon'} \tilde{q}_2(t)dt.$$

Comparing C and $\tilde{C}$, we get

$$\begin{aligned} C - \tilde{C} = -\frac{\epsilon^2}{2(q_{20}\gamma + \lambda_1 - s\mu_1)^2}(q_{20}\gamma + \lambda_1)\big(&- c_2\epsilon(\theta_2+\gamma)(q_{20}\gamma+\lambda_1)(q_{20}(\theta_2+\gamma) - \lambda_2) \\ &- c_2 s\,((1+\epsilon(\theta_2+\gamma))(q_{20}\gamma+\lambda_1) - s\mu_1)\,\mu_2 + c_1\big(q_{20}^2\gamma^2\epsilon(\gamma+\theta_1+\theta_2) + s\mu_1(\lambda_1 - s\mu_1) \\ &+ \epsilon\lambda_1(\theta_1\lambda_1 - \gamma\lambda_2 + s\gamma\mu_2) + q_{20}\gamma\,(\epsilon(2\theta_1+\theta_2)\lambda_1 + s\mu_1 + \gamma\epsilon(\lambda_1 - \lambda_2 + s\mu_2))\big)\big) = o(\epsilon). \end{aligned} \tag{20}$$

In addition, at the end of time $\epsilon + \epsilon'$, we have $q_1(\epsilon + \epsilon') = \tilde{q}_1(\epsilon + \epsilon') = 0$, and

$$q_2(\epsilon + \epsilon') - \tilde{q}_2(\epsilon + \epsilon') = -\frac{\epsilon^2(\theta_2+\gamma)(q_{20}\gamma+\lambda_1)(q_{20}(\theta_2+\gamma) - \lambda_2 + s\mu_2)}{q_{20}\gamma + \lambda_1 - s\mu_1} = o(\epsilon). \tag{21}$$

Importantly, (20) implies that the cost under the policy that have q_1 first increase and then decrease and the cost under strict priority rule to Class 1 which maintains q_1 at zero differ by a magnitude of $o(\epsilon)$. The same applies to the queue lengths at time $\epsilon + \epsilon'$ under the two policies by (21). Now for any interval of length L, suppose we divide it into $L/(\epsilon + \epsilon')$ small triangles (first increase and then decrease trajectories). Each has a cost difference $o(\epsilon)$ from the cost under strict priority to Class 1. Then the overall cost difference between the two policies (chattering and non-chattering) is $o(\epsilon)L/(\epsilon + \epsilon')$, which goes to zero as ϵ diminishes. Note that any chattering interval consists of infinitely many such triangular trajectories with infinitesimally small intervals over which q_1 first increases above and then decreases to zero. This implies that any admissible control policy π that yields a chattering interval where q_1 fluctuates infinitesimally around zero can be replaced by a cost-wise equivalent control $\tilde{\pi}$ that maintains q_1 at zero over the same interval and agrees with π elsewhere. The same approach applies to any chattering interval of q_2 around zero – i.e., we can show that there exists a cost-wise equivalent control under which q_2 does not chatter (stays at zero). Q.E.D.

B.2. Proof of Proposition 1

PROOF: Let $(q_1(0), q_2(0)) = (\epsilon, \epsilon)$. Since the optimal control gives strict priority to one class at any given time, for $\epsilon > 0$ sufficiently small, it is sufficient to compare the two strict priority rules; see Larrañaga (2015) for a similar observation. Under each priority rule, we characterize the fluid trajectory and calculate the cost. By comparing the costs under the strict priority rules, we note that when the system is initiated arbitrarily close to the origin, the optimal policy is to follow the $c\mu$-rule.

We first consider **strict priority to Class 1**. The time horizon is divided into two intervals with length t_1 and t_2 respectively. Class 1 first receives full service capacity and gets emptied at the end of the first interval. Over the second interval, Class 1 is maintained at zero queue and Class 2 is eventually emptied. The fluid trajectory over the first interval is characterized by

$$q_1(t) = \epsilon + (\gamma\epsilon - \epsilon\theta_1 + \lambda_1 - s\mu_1)\, t + o(\epsilon), \quad t \in [0, t_1]$$
$$q_2(t) = \epsilon + (-\gamma\epsilon - \epsilon\theta_2 + \lambda_2)\, t + o(\epsilon), \quad t \in [0, t_1],$$

and the value of t_1 is

$$t_1 = \frac{\epsilon}{s\mu_1 - \lambda_1} + o(\epsilon).$$

Taking the value of $(q_1(t_1), q_2(t_1))$ as the initial condition, the fluid trajectory over the second interval is

$$q_1(t) = 0, \quad t \in [0, t_2]$$
$$q_2(t) = \frac{\epsilon\,(\lambda_1 - \lambda_2 - s\mu_1)\,((-1 + t(\theta_2 + \gamma))\,\mu_1 - t\gamma\mu_2) - t\,(\lambda_1 - s\mu_1)\,(\lambda_2\mu_1 + (\lambda_1 - s\mu_1)\mu_2)}{\mu_1(-\lambda_1 + s\mu_1)} + o(\epsilon), \quad t \in [0, t_2],$$

and the value of t_2 is

$$t_2 = \frac{\epsilon\mu_1(-\lambda_1 + \lambda_2 + s\mu_1)}{(\lambda_1 - s\mu_1)(\lambda_2\mu_1 + (\lambda_1 - s\mu_1)\mu_2)} + o(\epsilon).$$

The cumulative holding cost under P_1 over $[0, t_1 + t_2]$ is given by

$$\begin{aligned} C^{P_1} &= c_1 \int_0^{t_1} [\epsilon + (\gamma\epsilon - \epsilon\theta_1 + \lambda_1 - s\mu_1)\, t]\, dt + c_1 \int_0^{t_1} [\epsilon + (-\gamma\epsilon - \epsilon\theta_2 + \lambda_2)\, t]\, dt \\ &\quad + c_2 \int_0^{t_2} \left[\frac{\epsilon\,(\lambda_1 - \lambda_2 - s\mu_1)\,((-1 + t(\theta_2 + \gamma))\,\mu_1 - t\gamma\mu_2) - t\,(\lambda_1 - s\mu_1)\,(\lambda_2\mu_1 + (\lambda_1 - s\mu_1)\mu_2)}{\mu_1(-\lambda_1 + s\mu_1)} \right] dt \\ &\quad + o(\epsilon^2) \\ &= -\frac{\epsilon^2\,(c_1\lambda_2\mu_1 - c_2\lambda_2\mu_2 + c_1(\lambda_1 - s\mu_1)\mu_2 + c_2\lambda_1(\mu_1 + 2\mu_2) - c_2 s\mu_1(\mu_1 + 2\mu_2))}{2(\lambda_1 - s\mu_1)(\lambda_2\mu_1 + (\lambda_1 - s\mu_1)\mu_2)} + o(\epsilon^2). \end{aligned}$$

Next, we consider **strict priority rule to Class 2**. Similar to the case under P_1, the time horizon is divided into two intervals with length t_1 and t_2 respectively. Class 2 first receives full service capacity and gets emptied over the first interval. Then, Class 2 is maintained at zero queue and Class 1 is eventually emptied over the second interval. The fluid trajectory over the first interval evolves as

$$q_1(t) = \epsilon + (\gamma\epsilon - \epsilon\theta_1 + \lambda_1)\, t + o(\epsilon), \quad t \in [0, t_1]$$
$$q_2(t) = \epsilon + (-\gamma\epsilon - \epsilon\theta_2 + \lambda_2 - s\mu_2)\, t + o(\epsilon), \quad t \in [0, t_1],$$

and t_1 has value

$$t_1 = \frac{\epsilon}{s\mu_2 - \lambda_2} + o(\epsilon).$$

Using the value of $(q_1(t_1), q_2(t_1))$ as the initial condition, the fluid trajectory over the second interval is

$$q_1(t) = \frac{\epsilon\mu_2(\lambda_1-\lambda_2+s\mu_2) - t\left(\lambda_2^2\mu_1+\lambda_2(-\epsilon\theta_1+\lambda_1-2s\mu_1)\mu_2+\mu_2(s(-\lambda_1+s\mu_1)\mu_2+\epsilon\theta_1(\lambda_1+s\mu_2))\right)}{\mu_2(-\lambda_2+s\mu_2)} + o(\epsilon), \quad t\in[0,t_2]$$

$$q_2(t) = 0, \quad t\in[0,t_2],$$

and the value of t_2 is

$$t_2 = \frac{\epsilon\mu_2(\lambda_1-\lambda_2+s\mu_2)}{(\lambda_2-s\mu_2)(\lambda_2\mu_1+(\lambda_1-s\mu_1)\mu_2)} + o(\epsilon).$$

The cumulative holding cost under P_2 over $[0, t_1+t_2]$ is given by

$$\begin{aligned}
C^{P_2} &= c_1\int_0^{t_1}\left[\epsilon+(\gamma\epsilon-\epsilon\theta_1+\lambda_1)\,t\right]dt + c_2\int_0^{t_1}\left[\epsilon+(-\gamma\epsilon-\epsilon\theta_2+\lambda_2-s\mu_2)\,t\right]dt \\
&+ c_1\int_0^{t_2}\left[\frac{\epsilon\mu_2(\lambda_1-\lambda_2+s\mu_2) - t\left(\lambda_2^2\mu_1+\lambda_2(-\epsilon\theta_1+\lambda_1-2s\mu_1)\mu_2+\mu_2(s(-\lambda_1+s\mu_1)\mu_2+\epsilon\theta_1(\lambda_1+s\mu_2))\right)}{\mu_2(-\lambda_2+s\mu_2)}\right]dt \\
&+ o(\epsilon^2) \\
&= \frac{\epsilon^2\left(-c_2(\lambda_2\mu_1+(\lambda_1-s\mu_1)\mu_2)+c_1(\lambda_1\mu_1+(2\mu_1+\mu_2)(-\lambda_2+s\mu_2))\right)}{2(\lambda_2-s\mu_2)(\lambda_2\mu_1+(\lambda_1-s\mu_1)\mu_2)} + o(\epsilon^2).
\end{aligned}$$

Comparing the total costs under P_1 and P_2, we get

$$\begin{aligned}
C^{P_1} - C^{P_2} &= \frac{\epsilon^2(c_1\mu_1-c_2\mu_2)\left(\lambda_1^2-\lambda_1(2\lambda_2+s(\mu_1-2\mu_2))+(\lambda_2+2s\mu_1)(\lambda_2-s\mu_2)\right)}{2(\lambda_1-s\mu_1)(-\lambda_2+s\mu_2)(\lambda_2\mu_1+(\lambda_1-s\mu_1)\mu_2)} + o(\epsilon^2) \\
&= \frac{\epsilon^2(c_1\mu_1-c_2\mu_2)\left(\lambda_1(\lambda_1-s\mu_1)+(\lambda_2-s\mu_2)(\lambda_2+2s\mu_1-2\lambda_1)\right)}{2(s\mu_1-\lambda_1)(s\mu_2-\lambda_2)(s\mu_1\mu_2-\lambda_1\mu_2-\lambda_2\mu_1)} + o(\epsilon^2),
\end{aligned} \tag{22}$$

Note that as $s > \lambda_1/\mu_1+\lambda_2/\mu_2$, in (22), the denominator $2(s\mu_1-\lambda_1)(s\mu_2-\lambda_2)(s\mu_1\mu_2-\lambda_1\mu_2-\lambda_2\mu_1) > 0$, and in the numerator, $(\lambda_1(\lambda_1-s\mu_1)+(\lambda_2-s\mu_2)(\lambda_2+2s\mu_1-2\lambda_1)) < 0$. Thus, for ϵ sufficiently small, $C^{P_1} - C^{P_2} < 0$ if and only if $c_1\mu_1 > c_2\mu_2$, and vice versa. This indicates that if the system is initiated sufficiently close to the origin, then the $c\mu$-rule is optimal. Q.E.D.

B.3. Pontryagin's Minimum Principle

Before we prove Propositions 2 and 3, we provide more details about Pontryagin's Minimum Principle. Consider the transient optimization problem (F2′) (also presented below).

$$\begin{aligned}
\min_{z} \quad & \int_0^{\tau} F\left(q(t)\right)dt \\
s.t. \quad & dq(t) = f\left(q(t), z(t)\right) \\
& g(q(t)) \le 0 \\
& h(z(t)) \le 0.
\end{aligned} \tag{F2′ revisited}$$

The pure state constraint $g(q(t)) \le 0$ is, in general, very hard to deal with as it does not explicitly involve the control $z(t)$ and can only be regulated indirectly via the ordinary differential equation $dq(t)$. To quantify how 'implicitly' $g(q(t))$ depends on $z(t)$, define g_i^j, $j = 1, 2, ..., \ell$, $i = 1, 2$, recursively as

$$\begin{aligned}
g_i^0(q(t), z(t)) &:= g(q(t)) \\
g_i^1(q(t), z(t)) &:= \nabla_q g_i^0(q(t), z(t))^T f\left(q(t), z(t)\right) \\
&\vdots \\
g_i^{\ell}(q(t), z(t)) &:= \nabla_q g_i^{\ell-1}(q(t), z(t))^T f\left(q(t), z(t)\right).
\end{aligned}$$

If $\nabla_z g_i^j(q(t), z(t)) = 0$ for $0 \leq j \leq \ell - 1$, and $\nabla_z g_i^\ell(q(t), z(t)) \neq 0$, then the state constraint $g_i(q(t))$ is said to be of *order* ℓ. It is easy to see that for (F2′), each pure state constraint is of order 1.

We next introduce a full rank assumption, often referred to as *constraint qualification*, on $g(q(t))$ and $h(z(t))$. In particular, for $g(q(t))$ of order 1, the constraint qualification requires that the matrices

$$\left[\frac{\partial g^1(q(t))}{\partial z}\right] \quad \text{and} \quad \left[\frac{\partial h(z(t))}{\partial z} \;\; \text{diag}\left(h(z(t))\right)\right]$$

have full rank for all $t \geq 0$. In the context of (F2′), we have

$$\text{rank}\left[\frac{\partial g^1(q(t))}{\partial z}\right] = \text{rank}\begin{bmatrix}\mu_1 & 0 \\ 0 & \mu_2\end{bmatrix} = 2,$$

and

$$\text{rank}\left[\frac{\partial h(z(t))}{\partial z} \;\; \text{diag}\left(h(z(t))\right)\right] = \text{rank}\begin{bmatrix}1 & 1 & z_1(t)+z_2(t)-s & 0 & 0 \\ -1 & 0 & 0 & -z_1(t) & 0 \\ 0 & -1 & 0 & 0 & -z_2(t)\end{bmatrix} = 3,$$

as at least one of $z_1(t)$ and $z_2(t)$ is strictly positive at all times. Hence, (F2′) satisfies the constraint qualification.

Under the constraint qualification, Pontryagin's Minimum Principle contains a list of necessary conditions satisfied by any optimal solution to the control problem. The next theorem summarizes some of the necessary conditions we utilize in our development. We refer to the survey paper Hartl et al. (1995) for a comprehensive summary of developments regarding Pontryagin's Minimum Principle for optimal control problems with state constraints.

Theorem 5 (Pontryagin's Minimum Principle (Hartl et al. (1995), Sethi and Thompson (2000))) *Assume that the constraint qualification holds. Let z^* be an optimal solution to* (F2′)*, q^* be the corresponding state trajectory, and τ^* be the optimal hitting time. Then, there exists a non-zero piecewise absolutely continuous adjoint vector $p^* : [0, \tau^*] \to \mathbb{R}^2$ with piecewise continuous derivatives, piecewise absolutely continuous Lagrangian multipliers $\eta^* : [0, \tau^*] \to \mathbb{R}^2$, $\xi^* : [0, \tau^*] \to \mathbb{R}^3$, and a vector $\omega^*(\beta_j) \in \mathbb{R}^2$ for each point β_j of discontinuity of p^* such that for almost every $t \in [0, \tau^*]$,*

1. *Ordinary Differential Equation condition:*

$$q^*(0) = q_0, \quad dq^*(t) = f\left(q^*(t), z^*(t)\right) \tag{ODE}$$

2. *Adjoint Vector condition:*

$$dp^*(t) = -\nabla_q L(q^*(t), z^*(t), p^*(t), \eta^*(t), \xi^*(t)) \tag{ADJ}$$

3. *Minimization condition:*

$$H(q^*(t), z^*(t), p^*(t)) = \min_z \{H(q^*(t), z(t), p^*(t))\} \tag{M}$$

4. *Hamiltonian condition:*

$$H(q^*(t), z^*(t), p^*(t)) = 0 \tag{H}$$

5. *Transversality condition:*

$$\nabla_z L(q^*(t), z^*(t), p^*(t), \eta^*(t), \xi^*(t)) = 0, \tag{T}$$

6. *Complementary condition:*

$$\begin{aligned} &\eta^*(t) \geq 0, \quad \eta^*(t)^T g(q^*(t)) = 0 \\ &\xi^*(t) \geq 0, \quad \xi^*(t)^T h(z^*(t)) = 0, \end{aligned} \tag{C}$$

7. *Jump condition: For any time in a boundary arc or a junction time, β, the adjoint vector p^* and the Hamiltonian H may have a discontinuity, but they must satisfy the following jump conditions:*

$$\begin{aligned} &(J1): p^*(\beta-) = p^*(\beta+) + \omega_1^*(\beta)\nabla_q g_1(q^*(\beta)) + \omega_2^*(\beta)\nabla_q g_2(q^*(\beta)) \\ &(J2): H(q^*(\beta-), z(\beta-), p^*(\beta-)) = H(q^*(\beta+), z(\beta+), p^*(\beta+)) - \omega_1^*(\beta)\nabla_t g_1(q^*(\beta)) - \omega_2^*(\beta)\nabla_t g_2(q^*(\beta)) \\ &(J3): \omega^*(\beta) \geq 0, \quad \omega^*(\beta)^T g\left(q^*(\beta)\right) = 0. \end{aligned} \tag{J}$$

Next, we provide more explanations about the conditions in Pontryagin's Minimum Principle listed in Theorem 5 to complement the discussion in Section 4.3.

1. First, solving the ordinary differential equations in (ADJ) for the dynamic of the adjoint vectors, we get

$$\begin{aligned} p_1^*(t) &= K_1 e^{t\theta_1} + e^{t\theta_1} \int_0^t e^{-s\theta_1} \left(-c_1 + \eta_1^*(s)\right) ds \\ p_2^*(t) &= K_2 e^{t(\theta_2+\gamma)} + \frac{K_1\gamma}{\gamma - \theta_1 + \theta_2} \left(e^{t\theta_1} - e^{t(\theta_2+\gamma)}\right) + \frac{\gamma}{\gamma - \theta_1 + \theta_2} e^{t\theta_1} \int_0^t e^{-s\theta_1} \left(-c_1 + \eta_1^*(s)\right) ds \\ &\quad - \frac{\gamma}{\gamma - \theta_1 + \theta_2} e^{t(\theta_2+\gamma)} \int_0^t e^{s(-\gamma-\theta_2)} \left(-c_1 + \eta_1^*(s)\right) ds + e^{t(\theta_2+\gamma)} \int_0^t e^{s(-\gamma-\theta_2)} \left(-c_2 + \eta_2^*(s)\right) ds, \end{aligned} \tag{23}$$

   where K_1 and K_2 are constants that depends on $p^*(0)$, which we, in general, cannot fully specify based on the necessary conditions.

   The adjoint vector is connected to the value function in optimal control theory. In particular, the value function $\Xi : \mathbb{R}_+^2 \to \mathbb{R}_+$ associated with (F2′) is defined by

$$\Xi(a_1, a_2) = \inf \left\{ \int_0^\tau F(q(t))dt \Big| \, q_1(0) = a_1,\, q_2(0) = a_2,\, q \text{ is a feasible trajectory in (F2}') \right\}.$$

   There exists an adjoint vector $p^*(t)$ such that $p^*(t) = \nabla_q \Xi(q^*(t))$ under the condition that $\nabla_q \Xi(q)$ is well defined (Frankowska 2010). As the cost structure is linear and increasing in $q^*(t)$, it follows that $p^*(t) \geq 0$ for all $t \geq 0$.

2. Minimization condition (M) and the optimal assignment of service capacity in equations (14) - (15) reveal important properties of the optimal control structure. First, observe in (14) - (15) that on the interior arc when both states are strictly positive and the switching curve is non-zero, the optimal control is 'bang-bang'. Namely, it must be the case that one of the two classes is assigned full service capacity s. On the other hand, on the boundary arc when one of the states is at zero, the optimal control is of an 'interior' type. Namely, both $z_1^*(t)$ and $z_2^*(t)$ stay strictly in the interior of the control region, i.e., $z_1^*(t), z_2^*(t) \in (0, s)$.

3. Consider time β, where $\beta < \tau^*$, as a time on a boundary arc or a junction time. If the adjoint vector p^* has a discontinuity at time β, then Jump condition (J) requires that

$$\begin{bmatrix} p_1^*(\beta-) \\ p_2^*(\beta-) \end{bmatrix} = \begin{bmatrix} p_1^*(\beta+) \\ p_2^*(\beta+) \end{bmatrix} + w_1^*(\beta) \begin{bmatrix} -1 \\ 0 \end{bmatrix} + w_2^*(\beta) \begin{bmatrix} 0 \\ -1 \end{bmatrix} = \begin{bmatrix} p_1^*(\beta+) - w_1^*(\beta) \\ p_2^*(\beta+) - w_2^*(\beta) \end{bmatrix},$$

and that

$$w_i^*(\beta) \geq 0, \quad w_i^*(\beta) g_i(q^*(\beta)) = 0, \quad i = 1, 2.$$

Note that if $q_1^*(\beta) = 0$, then $q_2^*(\beta) > 0$ and thus $w_2^*(\beta) = 0$. The same holds true for q_2^*, namely, if $q_2^*(\beta) = 0$, then $q_1^*(\beta) > 0$ and thus $w_1^*(\beta) = 0$. Hence, only the adjoint vector associated with the queue that is at zero can have a jump, while the other adjoint vector remains continuous at time β.

In addition, since the pure state constraint $g(q(t))$ is time invariant, we have $g_t(q^*(\beta)) = 0$. According to Jump condition (J), the Hamiltonian $H(q^*(t), z^*(t), p^*(t))$ is continuous over boundary arcs and at junction times.

4. Pontryagin's Minimum Principle only requires the necessary conditions to be satisfied 'almost everywhere'. In particular, $q^*(t)$ and $p^*(t)$ can have discontinuities at countably many points. For most problems studied in the literature, jumps only happen at junction times (Hartl et al. 1995). That said, in general, we cannot rule out the possibility of jumps on the boundary or interior arcs. In our analysis, we shall first assume that $p^*(t)$ is *continuous on interior arcs*. We then show that the continuity assumption indeed holds by verifying a sufficient version of Pontryagin's Minimum Principle for the optimal control problem (F2′).

We next introduce the sufficient version of Pontryagin's Minimum Principle. Since the terminal state in problem (F2′) is zero and $F(0) = 0$, (F2′) can be equivalently formulated as an optimal control problem without a terminal state constraint but rather over an infinite time horizon. The following sufficient conditions are adapted from Theorem 8.2 and Theorem 8.4 in (Hartl et al. 1995) for the equivalent version of (F2′) over an infinite time horizon.

Theorem 6 (Arrow-type sufficient condition) *Let (q^*, z^*) be a feasible pair for an equivalent version of problem* (F2′) *with infinite time horizon. Assume that there exists a piecewise continuously differentiable function $p^*(t) : [0, \infty) \to \mathbb{R}^2$ and piecewise continuous functions $\eta^* : [0, \infty) \to \mathbb{R}^2$ and $\xi^* : [0, \infty) \to \mathbb{R}^3$, such that conditions (ODE), (ADJ), (M), (T), (C) hold. Assume further that at all points β of discontinuity of p^*, there exists an $\omega^*(\beta) \in \mathbb{R}^2$ such that (J1) and (J3) in (J) hold. In addition, assume that the following limiting condition holds:*

$$\lim_{t \to \infty} p^*(t)^T (q(t) - q^*(t)) \geq 0 \quad \textit{for every other feasible state trajectory } q.$$

If the minimized Hamiltonian $H(q^*(t), z^*(t), p^*(t)) = \min_z \{H(q^*(t), z(t), p^*(t))\}$ is convex in $q^*(t)$ for all $(p^*(t), t)$, the pure state constraint $g(q(t))$ is quasiconvex in $q(t)$, and the control constraint $h(z(t))$ is quasiconvex in $z(t)$, then (q^*, z^*) is an optimal pair.

We first note that the solution we derive in this paper indeed satisfies the sufficient conditions in Theorem 6 and is thus optimal. More specifically, first, we design the control by ensuring that conditions (ODE), (ADJ), (M), (T), and (C) are satisfied almost everywhere. In particular, in our proposed solution, the state trajectory $q^*(t)$ satisfies (ODE) at all the continuity points of the control $z^*(t)$. The adjoint vector $p^*(t)$ follows the ordinary differential equations in (ADJ) everywhere on the interior arcs. (M), (T), and (C)

hold everywhere over the transient time horizon. Second, Jump condition (J) is guaranteed everywhere over boundary arcs and at junction times. Since $p^*(t)$ is continuous over interior arcs, conditions (J1) and (J3) in (J) indeed hold for all discontinuity points of $p^*(t)$. Third, for any feasible state trajectory $q(t)$ other than $q^*(t)$, $\lim_{t\to\infty} p^*(t)(q(t)-q^*(t)) \geq 0$ holds because $p^*(t), q(t) \geq 0$ for all $t \geq 0$, and $\lim_{t\to\infty} q^*(t) = 0$. Lastly, following (14)-(15), the control $z^*(t)$ is linear in $q^*(t)$ for all $t \geq 0$. Hence, the minimized Hamiltonian $H(q^*(t), z^*(t), p^*(t))$ is linear in $q^*(t)$ for all $(p^*(t), t)$. The convexity conditions on $g(q(t))$ and $h(z(t))$ are also satisfied as $g(q(t))$ and $h(z(t))$ are linear in $q(t)$ and $z(t)$ respectively.

We are now ready to prove the results in Section 4.3 using Pontryagin's Minimum Principle.

B.4. Proof of the auxiliary lemmas

B.4.1. Proof of Lemma 2 PROOF: The proof of Lemma 2 uses Transversality condition (T) and Complementarity condition (C) . Consider a boundary arc $[t_1, t_2]$ and a time epoch $t \in (t_1, t_2)$. First, by (14) - (15), the control over the boundary arc is of an 'interior' type, and the amount of service capacity assigned to both classes $(z_1^*(t), z_2^*(t))$ is strictly positive. By Complementarity condition (C), the multipliers satisfy $\xi_2^*(t) = 0$ and $\xi_3^*(t) = 0$. Then, by Transversality condition (T), we have $\mu_1 p_1^*(t) = \mu_2 p_2^*(t) = \xi_1^*(t)$. Hence, the switching curve satisfies $\psi(t) = \mu_1 p_1^*(t) - \mu_2 p_2^*(t) = 0$ for $t \in (t_1, t_2)$. Q.E.D.

B.5. Proof of Lemma 3

PROOF: Recall that the switching curve is characterized by $\psi(t) = \mu_1 p_1^*(t) - \mu_2 p_2^*(t)$. Since $\psi(t) = 0$ on the boundary arcs and by our construction, $p^*(t)$ does not jump on the interior arcs, the switching curve $\psi(t)$ is continuous at all time $t \in [0, \tau^*]$ if $p^*(t)$ is continuous at the junction times. In the rest of the proof, we establish the continuity of $p^*(t)$ at the junction times.

Following Proposition 4.2 in Hartl et al. (1995) and Proposition 3.63 in Grass et al. (2008), for the optimal control problem (F2′) which has pure state constraints of order 1, the adjoint vector $p^*(t)$ is continuous at a junction time β, i.e., $\omega^*(\beta) = 0$, if the entry or exit is nontangential, i.e., $dq_i^*(\beta-) < 0$ or $dq_i^*(\beta+) > 0$, respectively. Namely, the nontangential condition requires that if β is an entry or contact point for q_i^*, then $dq_i^*(\beta-) < 0$. If β is an exit or contact point for q_i^*, then $dq_i^*(\beta+) > 0$. In what follows, we use this nontangential condition and/or Jump condition (J) to establish continuity of $p^*(t)$ at junction times.

Case I: Junction times for Class 2. First, let β be an entry or contact point for q_2^*. In order to drive q_2^* to zero at β, full service capacity must be assigned to q_2^* right before β, i.e., $z_2^*(\beta-) = s$. It follows that

$$dq_2^*(\beta-) = \lambda_2 - \mu_2 z_2^*(\beta-) - \theta_2 q_2^*(\beta-) - \gamma q_2^*(\beta-) = \lambda_2 - \mu_2 s < 0.$$

Second, let β be an exit or contact point for q_2^*. Since there is always sufficient capacity in the system to maintain q_2^* at zero, it must be the case that priority is switched from Class 2 to Class 1 at time β. This implies $z_2^*(\beta+) = 0$ and

$$dq_2^*(\beta+) = \lambda_2 - \mu_2 z_2^*(\beta+) - (\theta_2 + \gamma) q_2^*(\beta+) = \lambda_2 > 0.$$

Since all the entry and exit trajectories are nontangential, the adjoint vectors $p^*(t)$ are continuous at the junction times associated with Class 2.

Case II: Junction times for Class 1. The proof for the continuity of $p^*(t)$ at the junction times associated with Class 1 gets more complicated due to degradation. The discussion is divided into three cases based on the relative level of service capacity s.

1. $\max\{\frac{\lambda_1}{\mu_1}+\frac{\gamma}{\theta_2+\gamma}\frac{\lambda_2}{\mu_1}, \frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_2}\}<s$**.**

(i) Let β be an entry or contact point for q_1^*.

In order to drive q_1^* to zero, full service capacity must be assigned to q_1^* right before β, i.e., $z_1^*(\beta-)=s$. Hence,

$$dq_1^*(\beta-)=\lambda_1-\mu_1 z_1^*(\beta-)-\theta_1 q_1^*(\beta-)+\gamma q_2^*(\beta-)=\lambda_1-\mu_1 s+\gamma q_2^*(\beta-).$$

In addition, there exists some neighborhood $[\beta-\delta,\beta)$, $0<\delta<\beta$, where $dq_1^*(t)<0$ for all $t\in[\beta-\delta,\beta)$. This implies that

$$q_2^*(t)<(s\mu_1-\lambda_1)/\gamma \quad \text{for all } t\in[\beta-\delta,\beta).$$

We next show that $dq_1^*(\beta-)<0$. Suppose by contradiction $dq_1^*(\beta-)=0$, then it must be the case that $q_2^*(\beta)=(s\mu_1-\lambda_1)/\gamma$. On the other hand,

$$dq_2^*(t)=\lambda_2-\mu_2 z_2^*(t)-(\theta_2+\gamma)q_2^*(t)\le\lambda_2-(\theta_2+\gamma)q_2^*(t),$$

which is strictly negative if

$$q_2^*(t)>\lambda_2/(\theta_2+\gamma).$$

Under Assumption 2, i.e. $s>\max\{\lambda_1/\mu_1+\lambda_2/\mu_2,\ \lambda_1/\mu_1+\lambda_2\gamma/((\gamma+\theta_2)\mu_1)\}$, it holds that

$$q_2^*(\beta)=(s\mu_1-\lambda_1)/\gamma>\lambda_2/(\theta_2+\gamma).$$

Therefore, there exists some $\delta'>0$, such that $dq_2^*(t)<0$ and $q_2^*(t)>q_2^*(\beta)$ for $t\in(\beta-\delta',\beta)$. It follows that $dq_1^*(t)>0$ for $t\in(\beta-\delta',\beta)$, which contradicts that $dq_1^*(t)<0$ for all $t\in[\beta-\delta,\beta)$. Therefore, $dq_1^*(\beta-)<0$ at entry or contact point β.

(ii) Let β be an exit or contact point for q_1^*. Similar arguments as in Case II.1.(i) apply and we can show that $dq_1^*(\beta+)>0$.

Since all the entry and exit trajectories are nontangential, the adjoint vectors $p^*(t)$ are continuous at the junction times associated with Class 1.

2. $\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_2}<s=\frac{\lambda_1}{\mu_1}+\frac{\gamma}{\theta_2+\gamma}\frac{\lambda_2}{\mu_1}$**.**

(i) Let β be an entry point for $q_1^*(t)$.

First, if $dq_1^*(\beta-)<0$, then it follows from the nontangential condition that there is no jump in the adjoint vector $p^*(t)$ at time β.

Second, suppose for the sake of contradiction that $dq_1^*(\beta-)=0$. It then follows that

$$q_2^*(\beta)=(s\mu_1-\lambda_1)/\gamma=\lambda_2/(\theta_2+\gamma).$$

Note that the point $(0,\lambda_2/(\theta_2+\gamma))$ is a locally asymptotically stable equilibrium point for the joint queue length process under priority to Class 1, while $(0,0)$ is the equilibrium under priority to Class 2. Hence, priority must be switched from Class 1 to Class 2 at time β; otherwise, keeping priority to

Class 1 generates extra cost but does not change the queue lengths. Hence, β cannot be an entry point for $q_1^*(t)$, a contradiction.

Therefore, $dq_1^*(\beta-) < 0$ at entry point β for q_1^*.

(ii) Let β be an exit point for q_1^*.

First, if $dq_1^*(\beta+) > 0$, then it follows from the nontangential condition that there is no jump in the adjoint vector $p^*(t)$ at time β.

Second, suppose for the sake of contradiction that $dq_1^*(\beta+) = 0$. Then, it is

$$q_2^*(\beta) = (s\mu_1 - \lambda_1)/\gamma = \lambda_2/(\theta_2 + \gamma).$$

Again, since the point $(0, \lambda_2/(\theta_2+\gamma))$ is a locally asymptotically stable equilibrium point for the joint queue length process, priority must be switched from Class 1 to Class 2 at time β. (Following the same reasoning as in Case II.2.(i), keeping priority to Class 1 generates extra cost but does not change the queue lengths.) Hence, $dz_1^*(\beta+) = 0$ and

$$dq_1^*(\beta+) = \lambda_1 - \mu_1 z_1^*(\beta+) + \gamma q_2^*(\beta+) > 0,$$

a contradiction.

Therefore, $dq_1^*(\beta+) > 0$ at exit point β for q_1^*.

(iii) Let β be a contact point for q_1^*.

First, if $dq_1^*(\beta-) < 0$ and $dq_1^*(\beta+) > 0$, then $p^*(t)$ does not have any jump at time β due to the nontangential condition.

Second, if $dq_1^*(\beta-) = 0$, then following the same arguments as in Case II.2.(i) and Case II.2.(ii), it holds that $q_2^*(\beta) = \lambda_2/(\theta_2+\gamma)$ and priority is switched from Class 1 to Class 2 at time β. In this case, Jump condition (J) requires the adjoint vector $p^*(t)$ to have no jump at time β. To see this, suppose for the sake of contradiction that $p^*(t)$ jumps at β. Then, Jump condition (J) characterizes that $p_1^*(\beta+) = p_1^*(\beta-) + w_1^*(\beta)$, for some $w_1^*(\beta) > 0$. Recall that the switching curve is defined as $\psi(t) = \mu_1 p_1^*(t) - \mu_2 p_2^*(t)$. Since Class 1 is prioritized right before β, it holds that $\psi(\beta-) \geq 0$. If $p_1^*(t)$ has a jump with strictly positive size $w_1^*(\beta)$ at time β, then $\psi(\beta+) > 0$. However, this implies that priority cannot be switched to Class 2 at time β, which is a contradiction.

Third, the case where $dq_1^*(\beta+) = 0$ is ruled out by exactly the same arguments in Case II.2.(ii).

In almost all (but one) cases, since all the entry and exit trajectories are nontangential, the adjoint vectors $p^*(t)$ are continuous at the junction times associated with Class 1. In the case where β is a contact point at which priority is switched from Class 1 to Class 2, we have established the continuity of $p^*(t)$ at β using Jump condition (J).

3. $\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} < s < \frac{\lambda_1}{\mu_1} + \frac{\gamma}{\theta_2+\gamma}\frac{\lambda_2}{\mu_1}$**.**

(i) Let β be an entry point for q_1^*.

First, if $dq_1^*(\beta-) < 0$, then $p_1^*(t)$ does not jump at β due to the nontangential condition.

Second, suppose for the sake of contradiction that $dq_1^*(\beta-)=0$. Then, $q_2^*(\beta)=(s\mu_1-\lambda_1)/\gamma<\lambda_2/(\theta_2+\gamma)$. Recall that the dynamic of q_2^* follows $dq_2^*(t)=\lambda_2-\mu_2 z_2^*(t)-(\theta_2+\gamma)q_2^*(t)$. Because priority is kept at Class 1 over the boundary arc following β, there exists some $\delta>0$ such that $dq_2^*(t)>0$ for $t\in[\beta,\beta+\delta)$. This implies that $dq_1^*(t)>0$ for $t\in(\beta,\beta+\delta)$, contradicting the fact that β is an entry point for $q_1^*(t)$.

Therefore, $dq_1^*(\beta-)<0$ at entry point β for q_1^*.

(ii) Let β be an exit point for q_1^*.

First, if $dq_1^*(\beta+)>0$, then $p_1^*(t)$ does not jump at β due to the nontangential condition.

Second, suppose for the sake of contradiction that $dq_1^*(\beta+)=0$. Then, priority must be kept at Class 1 at time β and over some interval $[\beta,\beta+\delta_1)$, $\delta_1>0$; otherwise, $dq_1^*(\beta+)>0$. In addition, we have, $q_2^*(\beta)=(s\mu_1-\lambda_1)/\gamma<\lambda_2/(\theta_2+\gamma)$. It then follows from the dynamic of q_2^* that there further exists some δ_2, $0<\delta_2<\delta_1$, such that $z_1^*(t)=s$, $dq_1^*(t)>0$ and $dq_2^*(t)>0$ for $t\in[\beta,\beta+\delta_2)$. Since $p_1^*(t)\geq 0$, $p_2^*(t)\geq 0$ and $q_2^*(t)>0$ for $t\in[\beta,\beta+\delta_2)$, we have $H(q^*(t),z^*(t),p^*(t))=p_1^*(t)dq_1^*(t)+p_2^*(t)dq_2^*(t)+c_1q_1^*(t)+c_2q_2^*(t)>0$ for $t\in[\beta,\beta+\delta_2)$. However, Hamiltonian condition (H) requires that $H(q^*(t),z^*(t),p^*(t))=0$ almost everywhere, which gives a contradiction.

Therefore, $dq_1^*(\beta+)>0$ at exit point β for q_1^*.

(iii) Let β be a contact point for q_1^*.

First, if $dq_1^*(\beta-)<0$ and $dq_1^*(\beta+)>0$, then $p^*(t)$ does not jump at β due to the nontangential condition.

Second, note that if priority is switched from Class 1 to Class 2 at time β, then Jump condition (J) requires that $p^*(t)$ does not jump at β due to the same reasoning as in Case II.2.(iii). Now, suppose for the sake of contradiction that $dq_1^*(\beta-)=0$ and priority is kept at Class 1 over some interval $[\beta,\beta+\delta_1)$, $\delta_1>0$. Then, following the same arguments as in Case II.3.(ii), there exists some δ_2, $0<\delta_2<\delta_1$, such that $z_1^*(t)=s$, $dq_1^*(t)>0$ and $dq_2^*(t)>0$ for $t\in[\beta,\beta+\delta_2)$, which violates Hamiltonian condition (H), which gives a contradiction.

Third, the case where $dq_1^*(\beta+)=0$ is ruled out by the same arguments as in Case II.3.(ii).

In almost all (but one) cases, since all the entry and exit trajectories are nontangential, the adjoint vectors $p^*(t)$ are continuous at the junction times associated with Class 1. In the case where β is a contact point at which priority is switched from Class 1 to Class 2, we have established the continuity of $p^*(t)$ at β using Jump condition (J).

Taking Cases I and II together, we have shown that the adjoint vectors $p^*(t)$ are continuous at all the junction times. This further implies that the switching curve $\psi(t)$ is continuous at all $t\in[0,\tau^*]$. Q.E.D.

B.5.1. Proof of Lemma 4 PROOF: By Lemma 1, we restrict to trajectories without chattering behavior. For any entry or contact point τ_j, there exists a nontrivial interval $(0,\alpha_j)$ such that for $t\in(0,\alpha_j)$, $q_1^*(\tau_j-t)$ and $q_2^*(\tau_j-t)$ are both strictly positive. Since the multiplier η^* is equal to zero over any interior arc, from (ADJ) we have, for $t\in(0,\alpha_j)$,

$$p_1^*(\tau_j-t)=\frac{c_1}{\theta_1}+e^{\theta_1(-t+\tau_j)}K_1$$

$$p_2^*(\tau_j-t)=\frac{c_2}{\theta_2+\gamma}+\frac{c_1\gamma}{\theta_1(\theta_2+\gamma)}+\frac{e^{\theta_1(-t+\tau_j)}\gamma K_1+e^{(\theta_2+\gamma)(-t+\tau_j)}(-\gamma K_1+(\gamma-\theta_1+\theta_2)K_2)}{\gamma-\theta_j+\theta_2},$$

where K_1 and K_2 are constant of integration that depend on $p^*(\tau_j)$ and τ_j.

The backward switching curve from time τ_j over the interval $(0, \alpha_j)$ is given by

$$\begin{aligned}
\psi(\tau_j - t) &= \mu_1 p_1^*(\tau_j - t) - \mu_2 p_2^*(\tau_j - t) \\
&= \mu_1 \left(\frac{c_1}{\theta_1} + e^{\theta_1(-t+\tau_j)} K_1 \right) \\
&\quad - \mu_2 \left(\frac{c_2}{\theta_2+\gamma} + \frac{c_1\gamma}{\theta_1(\theta_2+\gamma)} + \frac{e^{\theta_1(-t+\tau_j)}\gamma K_1 + e^{(\theta_2+\gamma)(-t+\tau_j)}(-\gamma K_1 + (\gamma - \theta_1 + \theta_2)K_2)}{\gamma - \theta_1 + \theta_2} \right) \\
&= K_1\mu_1 e^{\theta_1(-t+\tau_j)} - \frac{\gamma}{\gamma - \theta_1 + \theta_2}\mu_2 K_1 e^{\theta_1(-t+\tau_j)} - \frac{\mu_2}{\gamma - \theta_1 + \theta_2}(-\gamma K_1 + (\gamma - \theta_1 + \theta_2)K_2)e^{(\theta_2+\gamma)(-t+\tau_j)} \\
&\quad + \frac{c_1}{\theta_1}\mu_1 - \left(\frac{c_2}{\theta_2+\gamma}\mu_2 + \frac{\frac{\gamma}{(\theta_2+\gamma)}c_1}{\theta_1}\mu_2 \right) \\
&= \left(\mu_1 - \frac{\gamma}{\gamma - \theta_1 + \theta_2}\mu_2 \right) K_1 e^{\theta_1\tau_j} e^{-\theta_1 t} - \frac{\mu_2}{\gamma - \theta_1 + \theta_2}(-\gamma K_1 + (\gamma - \theta_1 + \theta_2)K_2)e^{(\theta_2+\gamma)\tau_j} e^{-(\theta_2+\gamma)t} \\
&\quad + \frac{c_1}{\theta_1}\mu_1 - \left(\frac{c_2}{\theta_2+\gamma}\mu_2 + \frac{\frac{\gamma}{(\theta_2+\gamma)}c_1}{\theta_1}\mu_2 \right).
\end{aligned}$$

Define

$$\begin{aligned}
A_1(\tau_j) &:= \left(\mu_1 - \frac{\gamma}{\gamma - \theta_1 + \theta_2}\mu_2 \right) K_1 e^{\theta_1\tau_j} \\
A_2(\tau_j) &:= -\frac{\mu_2}{\gamma - \theta_1 + \theta_2}(-\gamma K_1 + (\gamma - \theta_1 + \theta_2)K_2)e^{(\theta_2+\gamma)\tau_j}.
\end{aligned}$$

We have, for $t \in (0, \alpha_j)$,

$$\psi(\tau_j - t) = A_1(\tau_j)e^{-\theta_1 t} + A_2(\tau_j)e^{-(\theta_2+\gamma)t} + \frac{c_1}{\theta_1}\mu_1 - \left(\frac{c_2}{\theta_2+\gamma}\mu_2 + \frac{\frac{\gamma}{(\theta_2+\gamma)}c_1}{\theta_1}\mu_2 \right).$$

Q.E.D.

B.5.2. Proof of Proposition 3 PROOF: The proof utilizes Proposition 1 and the possible shapes of the switching curve, $\psi(\tau_N - t)$, characterized in Lemma 4. It is divided into two cases, depending on the relationship between the $c\mu$-index and the modified $c\mu/\theta$-index.

Case I. First, we consider the parameter regime where the $c\mu$-rule and the modified $c\mu/\theta$-rule prioritize the same class, namely,

$$(c_1\mu_1 - c_2\mu_2)\left(\frac{c_1}{\theta_1}\mu_1 - \left(\frac{c_2}{\theta_2+\gamma}\mu_2 + \frac{\frac{\gamma}{(\theta_2+\gamma)}c_1}{\theta_1}\mu_2 \right) \right) > 0.$$

For the moment, suppose Class 1 has a higher $c\mu$-index and modified $c\mu/\theta$-index.

By Proposition 1, when the state is in an ϵ-neighborhood of the origin, it is optimal to assign strict priority to Class 1. Recall that τ_N is the last entry or contact point (forward in time) when one of the states hits zero. It follows that τ_N must be the last epoch forward in time when q_1^* hits zero, and q_1^* is then maintained at zero after τ_N, i.e., $q_1^*(t) = 0$ for $t \in [\tau_N, \tau)$. By Lemma 4, the switching curve right before τ_N satisfies for some $\alpha_N < \tau_N$,

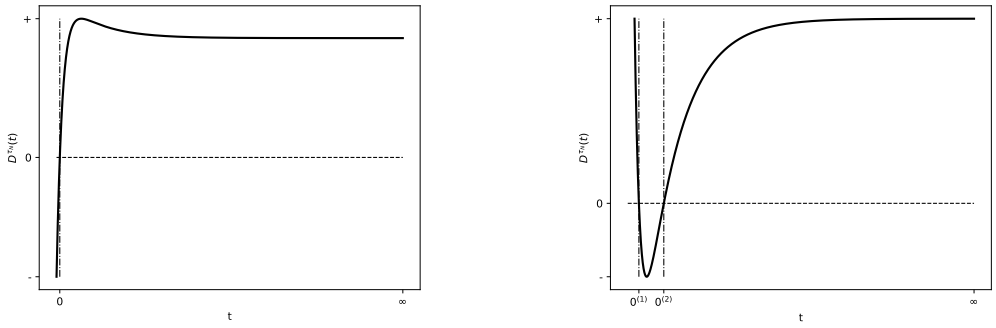
$$\psi(\tau_N - t) = A_1(\tau_N)e^{-\theta_1 t} + A_2(\tau_N)e^{-(\theta_2+\gamma)t} + \frac{c_1}{\theta_1}\mu_1 - \left(\frac{c_2}{\theta_2+\gamma}\mu_2 + \frac{\frac{\gamma}{(\theta_2+\gamma)}c_1}{\theta_1}\mu_2 \right), \quad t \in (0, \tau_N - \alpha_N),$$

where $A_1(\tau_N)$ and $A_2(\tau_N)$ are constants in $\mathbb{R}$. Furthermore, $D^{\tau_N}(t)$, the pseudo switching curve backward from τ_N, satisfies

$$\lim_{t\to\infty} D^{\tau_N}(t) = \frac{c_1}{\theta_1}\mu_1 - \left(\frac{c_2}{\theta_2+\gamma}\mu_2 + \frac{\frac{\gamma}{(\theta_2+\gamma)}c_1}{\theta_1}\mu_2 \right) > 0.$$

The structure of $D^{\tau_N}(t)$ regulates that it can have at most two zeros. With $\lim_{t\to\infty} D^{\tau_N}(t) > 0$, the two possible function shapes $D^{\tau_N}(t)$ can take are demonstrated in Figure 14, with one root in Figure 14(a) and two roots in Figure 14(b). Figure 14 is comprehensive in the sense that any such $D^{\tau_N}(t)$ function shares the same behavior in crossing zero and in the limiting regime as $t \to \infty$. In particular, if $D^{\tau_N}(t)$ has one zero as in Figure 14(a), then it must be that $D^{\tau_N}(t)$ is increasing at the zero point and eventually converges to $\lim_{t\to\infty} D^{\tau_N}(t)$. (For the instance plotted in Figure 14, $\lim_{t\to\infty} D^{\tau_N}(t) = 2.5$.) Once $D^{\tau_N}(t)$ crosses zero, it will never decrease to zero again. Likewise, if $D^{\tau_N}(t)$ has two zeros as in Figure 14(b), then it must be that $D^{\tau_N}(t)$ has negative slope at the first zero, has positive slope at the second zero, and eventually converges to 2.5. Once $D^{\tau_N}(t)$ crosses the second zero point, it will never decrease to zero again. We comment that if the values of $A_1(\tau_N)$ are $A_2(\tau_N)$ are known, then there is no ambiguity in the trajectory of $D^{\tau_N}(t)$, and thus no notion of 'possible' function shapes. Nevertheless, due to the degrees of freedom inherent to Pontryagin's Minimum Principle, it is hard to characterize these coefficients exactly. Therefore, the idea is to infer the structure of the optimal control from the interaction of the coefficients without explicitly characterizing their values.

Figure 14 **Possible trajectory of $D^{\tau_1}(t)$ with $c_1\mu_1 > c_2\mu_2$, modified $c_1\mu_1/\theta_1 >$ modified $c_2\mu_2/\theta_2$**



(a) One zero point

(b) Two zero points

We first note that the interval $[\tau_N, \tau]$ is a boundary arc over which q_1^* is maintained at zero. It follows that $\psi(t) = 0$ for $t \in (\tau_N, \tau)$ (Lemma 2), and $\psi(t)$ is continuous in time so that $\psi(\tau_N) = 0$ (Lemma 3). Furthermore, since the optimal control is 'bang-bang' right before τ_N, in order to drive q_1^* to zero at time τ_N, strict priority must be given to Class 1 in some non-trivial neighborhood before τ_N. Namely, there exists $\epsilon_{\tau_N} > 0$ such that $\psi(t) > 0$ for $t \in (\tau_N - \epsilon_{\tau_N}, \tau_N)$. For $D^{\tau_N}(t)$, this implies that $D^{\tau_N}(0) = 0$ and $D^{\tau_N}(t) > 0$ for $t \in (0, \epsilon_{\tau_N})$. Thus for the possible structures in Figure 14, if $D^{\tau_N}(t)$ has one zero (Figure 14(a)), then $D^{\tau_N}(0)$ is at this unique zero point. If $D^{\tau_N}(t)$ has two zeros (Figure 14(b)), then $D^{\tau_1}(0)$ is at the second zero. This implies that as long as the dynamic of the switching curve $\psi(\tau_N - t)$ follows that of $D^{\tau_N}(t)$, $\psi(\tau_N - t) > 0$. It is important to note that the trajectory of $\psi(\tau_N - t)$ agrees with $D^{\tau_N}(t)$ for t in some non-degenerative interval $(0, \tau_N - \alpha_N)$.

Next, taking the derivative of $D^{\tau_N}(t)$ with respect to t, it is easy to see that $dD^{\tau_N}(t)$ can have at most one root. Since $D^{\tau_N}(0) = 0$ and $D^{\tau_N}(t) > 0$ for $t \in (0, \epsilon_{\tau_N})$, it holds that for any interval $[0, \ell)$, $\ell > 0$, either $D^{\tau_N}(t)$ is strictly increasing over $[0, \ell)$ or $D^{\tau_N}(\ell) > \lim_{t\to\infty} D^{\tau_N}(t) - \delta$ for some $\delta > 0$ arbitrarily small. In either case, $D^{\tau_N}(\ell) > \delta'$ for some $\delta' > 0$. If $\eta_1^*(\tau_N - t) = 0$ and $\eta_2^*(\tau_N - t) = 0$ for $t \in [0, \ell)$, then the same

holds true for the backward switching curve $\psi(\tau_1 - t)$ over the interval $t \in [0, \ell)$. To this end, it is only possible for $\psi(\tau_N - t)$ to deviate from the dynamic of $D^{\tau_N}(t)$ if $\eta_2^*(\tau_N - \beta)$ becomes strictly positive at some time $0 < \beta \leq t$. (Naturally, $\beta \geq \alpha_N$.) Note that $\eta_1^*(\tau_N - t) = 0$ and $\eta_2^*(\tau_N - t) = 0$ for all $t \in [0, \beta)$. Now, as $D^{\tau_N}(\beta) > \delta'$ for some $\delta' > 0$ and $\eta_2^*(\tau_N - \beta) > 0$, it follows that $\psi(\tau_N - \beta) \geq \delta' > 0$. However, $\eta_2^*(\tau_N - \beta)$ becomes positive only if $q_2^*(\tau_N - \beta) = 0$, which implies that strict priority is given to Class 2 right before time $(\tau_N - \beta)$, i.e., $\psi((\tau_N - \beta) -) \leq 0$. However, due to the continuity of the switching curve, this contradicts the fact that $\psi(\tau_N - \beta) \geq \delta' > 0$. Therefore, for all $t \in [0, \tau_N]$, $\psi(\tau_N - t)$ follows the dynamic of $D^{\tau_N}(t)$ and remains strictly positive. We then conclude that strict priority to Class 1 is optimal throughout the transient time horizon.

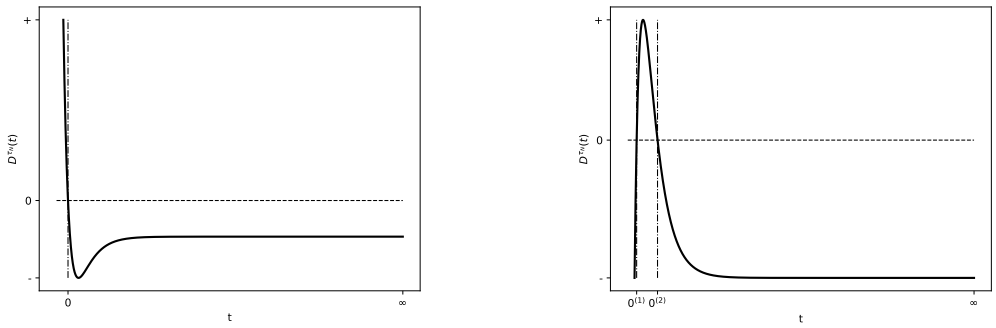
The proof for the case where Class 2 has a higher $c\mu$-index and higher modified $c\mu/\theta$-index follows similarly. In this case, strict priority to Class 2 is optimal throughout the transient time horizon.

Case II. We consider the case where the $c\mu$-rule and the modified $c\mu/\theta$-rule prioritize different classes, namely,

$$(c_1\mu_1 - c_2\mu_2)\left(\frac{c_1}{\theta_1}\mu_1 - \left(\frac{c_2}{\theta_2+\gamma}\mu_2 + \frac{\frac{\gamma}{(\theta_2+\gamma)}c_1}{\theta_1}\mu_2\right)\right) < 0.$$

For the moment, suppose Class 1 has a higher $c\mu$-index and Class 2 has a higher modified $c\mu/\theta$-index. Following similar lines of arguments as in Case I, the backward switching curve $\psi(\tau_N - t)$ follows the dynamic of $D^{\tau_N}(t)$ for some non-trivial time interval $t \in (0, \alpha_N)$. Again, the structure of $D^{\tau_N}(t)$ guarantees that it can have at most two zeros. With Class 2 having a higher modified $c\mu/\theta$-index, the two possible shapes for $D^{\tau_N}(t)$ are demonstrated in Figure 15, with Figure 15(a) crossing zero once and Figure 15(b) crossing zero twice. In particular, if $D^{\tau_N}(t)$ has one zero as in Figure 15(a), then it must be that $D^{\tau_N}(t)$ is decreasing at the zero point and eventually converges to $\lim_{t\to\infty} D^{\tau_N}(t) < 0$. Once $D^{\tau_N}(t)$ crosses zero, it will never increase to zero again. Likewise, if $D^{\tau_N}(t)$ has two zeros as in Figure 15(b), then it must be that $D^{\tau_N}(t)$ has positive slope at the first zero, has negative slope at the second zero, and eventually converges to $\lim_{t\to\infty} D^{\tau_N}(t) < 0$. Once $D^{\tau_N}(t)$ crosses the second zero point, it will never increase to zero again.

Figure 15 **Possible trajectory of $D^{\tau_1}(t)$ with $c_1\mu_1 > c_2\mu_2$, modified $c_1\mu_1/\theta_1 <$ modified $c_2\mu_2/\theta_2$**



(a) Two zero points (b) One zero point

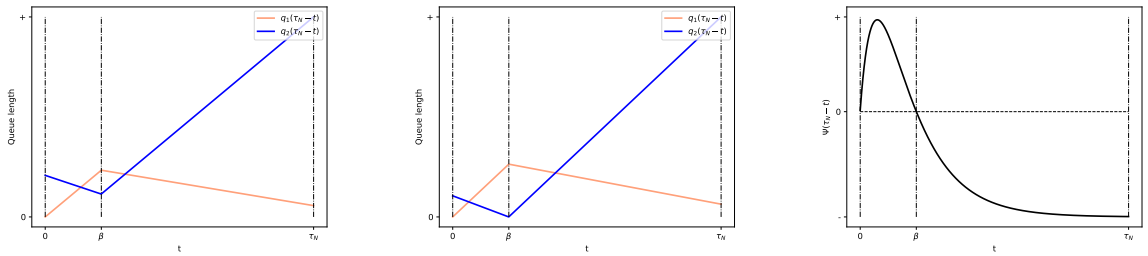
By Proposition 1, for $c_1\mu_1 > c_2\mu_2$, it is optimal to give strict priority to Class 1 when the system state is close enough to the origin. Therefore, τ_N is the last time before τ when q_1^* hits zero. In order to empty q_1^*,

strict priority must be given to Class 1 for some non-trivial time interval right before τ_N. This implies that there exits $\epsilon_{\tau_N} > 0$ such that $D_{\tau_N}(0) = 0$ and $D_{\tau_N}(t) > 0$ for $t \in (0, \epsilon_{\tau_N})$. In this case, we can rule out Figure 15(a). $D_{\tau_N}(0)$ must be at the first zero in Figure 15(b). Now, let time $\beta > 0$ denote the second zero in Figure 15(b), i.e., $D^{\tau_N}(\beta) = 0$. Then, one of the following three scenarios holds.

Scenario 1. $\tau_N \leq \beta$. The backward switching curve $\psi(\tau_N - t)$ agrees with $D^{\tau_N}(t)$ for all $t \in [0, \tau_N]$. Because $\psi(\tau_N - t) > 0$ for all $t \in [0, \tau_N)$, strict priority is given to Class 1 throughout the transient time horizon.

Scenario 2. $\tau_N > \beta$. The backward switching curve $\psi(\tau_N - t)$ follows $D^{\tau_N}(t)$ for $t \in [0, \beta)$. Both $q_1^*(\tau_N - t)$ and $q_2^*(\tau_N - t)$ stay strictly positive over $t \in (0, \beta)$. At time $t = \beta$, priority is switched from Class 1 to Class 2 (backward in time). In this case, either both queues are strictly positive at $t = \beta$ as in Figure 16(a), or β is a contact point as in Figure 16(b). In either case, the multipliers $\eta_1^*(\tau_N - t)$ and $\eta_2^*(\tau_N - t)$ stay at zero (or become positive only at one point). Then the backward switching curve $\psi(\tau_N - t)$ further follows $D^{\tau_N}(t)$ for some non-trivial interval, $(\beta, \beta + \delta)$ for some $\delta > 0$. Following similar arguments as in Case I, once crossing zero at $t = \beta$, the backward switching curve $\psi(\tau_N - t)$ remains strictly negative afterwards as shown in Figure 16(c). In this case, the optimal control (forward in time) switches priority once from Class 2 to Class 1.

Figure 16 Backward state trajectory and switching curve in Scenario 2



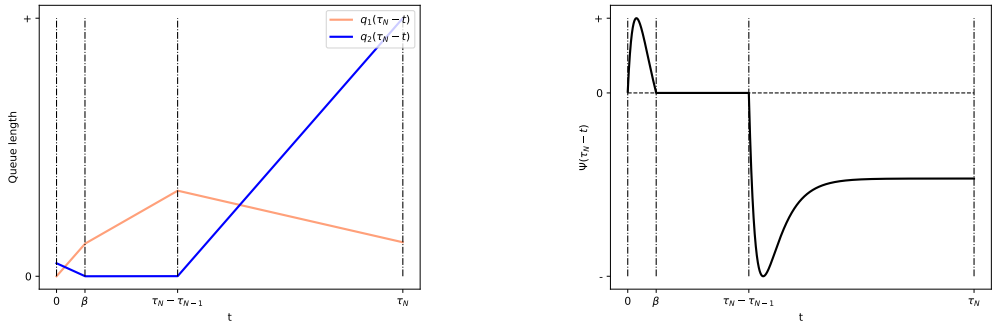
(a) non-junction time β (b) contact point β (c) Backward switching curve

Scenario 3. $\tau_N > \beta$. The backward switching curve $\psi(\tau_N - t)$ follows $D^{\tau_N}(t)$ for all $t \in [0, \beta)$. Both $q_1^*(\tau_N - t)$ and $q_2^*(\tau_N - t)$ stay strictly positive over $t \in (0, \beta)$. Different from the above case, β is an exit point (forward in time) for the trajectory of q_2^*; see Figure 17(a). Correspondingly, the entry point is τ_{N-1}. At time τ_{N-1}, the switching curve $\psi(\tau_{N-1}) = 0$. Now, we repeat the structural derivation for the backward switching curve starting from τ_{N-1}, namely, for the function $\psi(\tau_{N-1} - t)$. In order to drive q_2^* to zero at time τ_{N-1}, strict priority must be assigned to q_2^* for some amount of time right before τ_{N-1}. As such, there exits $\epsilon_{\tau_{N-1}} > 0$ such that $D^{\tau_{N-1}}(0) = 0$ and $D^{\tau_{N-1}}(t) < 0$ for $t \in (0, \epsilon_{\tau_{N-1}})$. Again, following similar arguments as in Case I, we can show that once crossing zero at τ_{N-1}, the switching curve $\psi(\tau_{N-1} - t)$ remains strictly negative afterwards. In this case, the optimal control (forward in time) switches priority once from Class 2 to Class 1. The structure of the backward switching curve in this case is illustrated in Figure 17(b).

In all the three scenarios above, the optimal control either assigns strict priority to Class 1 throughout, or switches from priority once from Class 2 to Class 1.

When Class 2 has a higher $c\mu$-index and Class 1 has a higher modified $c\mu/\theta$-index, the proof holds in a similar fashion. In this case, the optimal control either invariantly assigns strict priority to Class 2, or switches once from prioritizing Class 1 to Class 2. Q.E.D.

Figure 17 Backward state trajectory and switching curve in Scenario 3



(a) exit time β (b) Backward switching curve

B.6. Proof of Proposition 2

PROOF: First, as shown in Proposition 3, the if the $c\mu$-rule and the modified $c\mu/\theta$-rule prioritize the same class, then the modified $c\mu/\theta$-rule (the $c\mu$-rule) is optimal throughout the transient time horizon and the claim follows. Now, we consider the case where the $c\mu$-rule and the modified $c\mu/\theta$-rule prioritize different classes, namely,

$$(c_1\mu_1 - c_2\mu_2)\left(\frac{c_1}{\theta_1}\mu_1 - \left(\frac{c_2}{\theta_2+\gamma}\mu_2 + \frac{\frac{\gamma}{(\theta_2+\gamma)}c_1}{\theta_1}\mu_2\right)\right) < 0.$$

By Propositions 1 and 3, when the $c\mu$-rule and the modified $c\mu/\theta$-rule prioritize different classes, the optimal control follows the $c\mu$-rule near the origin and switches priority at most once along the trajectory. However, it remains to be shown whether or not the optimal control will ever switch priority. Namely, it remains unknown whether there exists a set of initial conditions from which the optimal trajectories switch priority from one class to the other. In this proof, we establish the existence of such initial values and provide a partial characterization of the states at which the system will follow the modified $c\mu/\theta$-rule. The proof is divided into two cases.

Case I. First, consider the case where the $c\mu$-rule prioritizes Class 2 and the modified $c\mu/\theta$-rule prioritizes Class 1. Suppose by contradiction that the optimal control follows the $c\mu$-rule throughout. Then, the optimal trajectory must be such that q_2^* is first driven to zero with full service capacity and maintained at zero afterwards. After q_2^* hits zero, q_1^* starts receiving positive service capacity and will be driven to zero at time τ. We next show that for systems initialized large enough, the cost of such trajectory can be decreased by imposing a change in the priority rule over a small time interval, thus a contradiction.

Consider two systems with the same, large initial workload $q_{10} + q_{20} > M$. The first system follows strict priority rule to Class 2; the second system gives strict priority to Class 2 until Class 2 hits zero, at τ_1, switches priority to Class 1 for ℓ amount of time, and then switches back to prioritizing Class 2 until the end of the transient time horizon τ. Because the two systems have identical trajectories up to τ_1 – i.e., the time epoch when q_2 hits zero, from τ_1 on, it is equivalent to consider the two systems initialized with q_{10} large and $q_{20} = 0$.

For the **first system**, priority is assigned to Class 2 throughout. Solving the trajectories directly, we get the following formula for the state trajectory $q^*(t)$, the hitting time τ, and the cumulative holding cost C:

$$
\begin{aligned}
q_1^*(t) &= \frac{e^{-t\theta_1}\left(\left(-1+e^{t\theta_1}\right)\lambda_2\mu_1 + q_{10}\theta_1\mu_2 + \left(-1+e^{t\theta_1}\right)\left(\lambda_1 - s\mu_1\right)\mu_2\right)}{\theta_1\mu_2}, \quad t\in[0,\tau] \\
q_2^*(t) &= 0, \quad t\in[0,\tau] \\
\tau &= \frac{1}{\theta_1}\log\left(\frac{\lambda_2\mu_1 - q_{10}\theta_1\mu_2 + \lambda_1\mu_2 - s\mu_1\mu_2}{\lambda_2\mu_1 + \lambda_1\mu_2 - s\mu_1\mu_2}\right) \\
C^* &= \int_0^\tau \left(c_1 q_1^*(t) + c_2 q_2^*(t)\right) dt.
\end{aligned}
$$

For the **second system**, priority is assigned to Class 1 for ℓ amount of time and then given to Class 2. We next calculate the cost over three intervals, $[0,\ell]$, $[\ell,\tau_N]$, and $[\tau_N,\tau]$ separately.

For the first interval where we assign s servers to Class 1 and 0 server to Class 2, we have

$$
\begin{aligned}
q_1^{(1)}(t) &= -\frac{1}{\theta_1(\theta_2+\gamma)(\gamma-\theta_1+\theta_2)}e^{-t(\theta_2+\gamma)}\left(-\gamma\theta_1\lambda_2 - e^{t(\theta_2+\gamma)}(\gamma-\theta_1+\theta_2)(\gamma\lambda_1+\theta_2\lambda_1+\gamma\lambda_2 - s(\theta_2+\gamma)\mu_1)\right. \\
&\quad \left. -e^{t(\gamma-\theta_1+\theta_2)}(\theta_2+\gamma)(q_{10}\theta_1(\gamma-\theta_1+\theta_2) + (\theta_1-\theta_2)(\lambda_1-s\mu_1) - \gamma(\lambda_1+\lambda_2-s\mu_1))\right), \quad t\in[0,T^{(1)}] \\
q_2^{(1)}(t) &= -\frac{\left(-1+e^{t(-\gamma-\theta_2)}\lambda_2\right)}{\theta_2+\gamma}, \quad t\in[0,T^{(1)}] \\
T^{(1)} &= \ell \\
C^{(1)} &= \int_0^{T^{(1)}} \left(c_1 q_1^{(1)}(t) + c_2 q_2^{(1)}(t)\right) dt.
\end{aligned}
$$

For the second interval where we assign 0 server to Class 1 and s servers to Class 2 to drive Class 2 to zero, we have

$$
\begin{aligned}
q_1^{(2)}(t) &= -\frac{1}{\theta_1\left(\theta_2+\gamma\right)\left(\gamma-\theta_1+\theta_2\right)}e^{-\ell(\theta_2+\gamma)-t(\gamma+\theta_1+\theta_2)}\Bigg(-e^{t\theta_1}\gamma\theta_1\lambda_2 - e^{t(\theta_2+\gamma)+\ell(\gamma-\theta_1+\theta_2)}(\theta_2+\gamma) \\
&\quad (q_{10}\theta_1(\gamma-\theta_1+\theta_2) + (\theta_1-\theta_2)(\lambda_1-s\mu_1) - \gamma(\lambda_1+\lambda_2-s\mu_1)) + e^{t\theta_1+\ell(\theta_2+\gamma)}s\gamma\theta_1\mu_2 \\
&\quad + e^{(\ell+t)(\theta_2+\gamma)}s(\theta_2+\gamma)\left((\gamma-\theta_1+\theta_2)\mu_1 - \gamma\mu_2\right) - e^{\ell(\theta_2+\gamma)+t(\gamma+\theta_1+\theta_2)}(\gamma-\theta_1+\theta_2) \\
&\quad (\theta_2\lambda_1 + \gamma(\lambda_1+\lambda_2-s\mu_2))\Bigg), \quad t\in[0,T^{(2)}] \\
q_2^{(2)}(t) &= \frac{1}{\theta_2+\gamma}\left(\lambda_2 - e^{-(\ell+t)(\theta_2+\gamma)}\lambda_2 + \left(-1+e^{-t(\theta_2+\gamma)}\right)s\mu_2\right), \quad t\in[0,T^{(2)}] \\
T^{(2)} &= \frac{1}{\theta_2+\gamma}\log\left(\frac{-e^{-\ell(\theta_2+\gamma)}\lambda_2 + s\mu_2}{-\lambda_2+s\mu_2}\right) =: \tau^N - \ell \\
C^{(2)} &= \int_0^{T^{(2)}} \left(c_1 q_1^{(2)}(t) + c_2 q_2^{(2)}(t)\right) dt.
\end{aligned}
$$

For the third interval where we maintain Class 2 at zero while driving Class 1 to zero, we have

$$\begin{aligned}
q_1^{(3)}(t) &= \frac{1}{\theta_1(-\gamma+\theta_1-\theta_2)} e^{-(\ell+t)\theta_1} \Bigg(\Bigg(q_{10}\theta_1(-\gamma+\theta_1-\theta_2) - (\theta_1-\theta_2)\left(\lambda_1 + (-1+e^{\ell\theta_1})s\mu_1\right) \\
&\quad + \gamma\left(\lambda_1+\lambda_2 - s\mu_1 + e^{\ell\theta_1}s(\mu_1-\mu_2)\right)\Bigg)\left(\frac{e^{-\ell(\theta_2+\gamma)}\lambda_2 - s\mu_2}{\lambda_2 - s\mu_2}\right)^{-\frac{\theta_1}{\theta_2+\gamma}} \\
&\quad - \frac{1}{\mu_2} e^{\ell\theta_1}\left((\lambda_2 - s\mu_2)(-(\gamma-\theta_1+\theta_2)\mu_1 + \gamma\mu_2) + e^{t\theta_1}(\gamma-\theta_1+\theta_2)(\lambda_2\mu_1 + (\lambda_1 - s\mu_1)\mu_2)\right)\Bigg), \\
&\quad t \in [0, T^{(3)}] \\
q_2^{(3)}(t) &= 0, \quad t \in [0, T^{(3)}] \\
T^{(3)} &= -\ell + \frac{1}{\theta_1}\log\Bigg(\frac{1}{(\gamma-\theta_1+\theta_2)(\lambda_2\mu_1 + (\lambda_1 - s\mu_1)\mu_2)}\Bigg(\Bigg(q_{10}\theta_1(-\gamma+\theta_1-\theta_2) - (\theta_1-\theta_2) \\
&\quad \left(\lambda_1 + \left(-1+e^{\ell\theta_1}\right)s\mu_1\right) + \gamma\left(\lambda_1+\lambda_2 - s\mu_1 + e^{\ell\theta_1}s(\mu_1-\mu_2)\right)\Bigg)\mu_2\left(\frac{e^{-\ell(\theta_2+\gamma)}\lambda_2 - s\mu_2}{\lambda_2 - s\mu_2}\right)^{-\frac{\theta_1}{\theta_2+\gamma}} \\
&\quad + e^{\ell\theta_1}(-\lambda_2 + s\mu_2)(-(\gamma-\theta_1+\theta_2)\mu_1 + \gamma\mu_2)\Bigg)\Bigg) =: \tau - \tau_N \\
C^{(3)} &= \int_0^{T^{(3)}} \left(c_1 q_1^{(3)}(t) + c_2 q_2^{(3)}(t)\right) dt.
\end{aligned}$$

Comparing the costs associated with System 1 and System 2, we have for ℓ small,

$$\begin{aligned}
C^* - \left(C^{(1)} + C^{(2)} + C^{(3)}\right) &= \frac{\ell^2 s\lambda_2}{2(\lambda_2 - s\mu_2)}\left(c_2\mu_2 + \frac{c_1(-\lambda_2\mu_1^2 + \mu_2\left(\mu_1(q_{10}(\theta_2+\gamma) - \lambda_1 + s\mu_1) - q_{10}\gamma\mu_2\right))}{\lambda_2\mu_1 + (-q_{10}\theta_1 + \lambda_1 - s\mu_1)\mu_2}\right) \\
&\quad + o(\ell^2),
\end{aligned}$$

and as $q_{10} \to \infty$,

$$C^* - \left(C^{(1)} + C^{(2)} + C^{(3)}\right) \to \frac{\ell^2 s\lambda_2\left(-c_1(\theta_2+\gamma)\mu_1 + c_1\gamma\mu_2 + c_2\theta_1\mu_2\right)}{2\theta_1(\lambda_2 - s\mu_2)} + o(\ell^2),$$

which is positive if and only if Class 1 has a higher modified $c\mu/\theta$-index. This implies that the cost difference $C^* - \left(C^{(1)} + C^{(2)} + C^{(3)}\right) > 0$ for sufficiently small ℓ and large q_{10}, and that we can find a policy that achieves a smaller cost than strict priority to Class 2, a contradiction. Therefore, if the states are initialized with $q_1(0) + q_2(0) > M$ for some M sufficiently large, the optimal control must switch from prioritizing Class 1 to prioritizing Class 2 at some time epoch $w \in (0, \tau)$. Moreover, it follows that the modified $c\mu/\theta$-rule is optimal at time t if $q_1(t) + q_2(t) > M$.

Case II. We next consider the case where the $c\mu$-rule prioritizes Class 1 and the modified $c\mu/\theta$-rule prioritizes Class 2. Note that the proof technique used for the previous case no longer applies. This is because under the strict priority rule to Class 1, the content of q_2 cannot be arbitrarily large when q_1 gets emptied due to degradation. In particular, let τ_1 denote the time when q_1 first gets emptied. Then it holds that for any $\epsilon > 0$,

$$q_2(\tau_1) < \left(s\mu_1 - \lambda_1\right)/\gamma + \epsilon.$$

In this case, our proof makes use of the backward switching curve and the fact that $q_2(\tau_1)$ is properly bounded.

We first note that by the definition of τ_1, both queues are strictly positive for $t < \tau_1$. Thus, the multiplies $\eta_1^*(t) = \eta_2^*(t) = 0$ for $t < \tau_1$. By Lemma 4, the backward switching curve before τ_1 is characterized as follows

$$\psi(\tau_1 - t) = A_1(\tau_1)e^{-\theta_1 t} + A_1(\tau_1)e^{-(\theta_2+\gamma)t} + \frac{c_1}{\theta_1}\mu_1 - \left(\frac{c_2}{\theta_2+\gamma}\mu_2 + \frac{\frac{\gamma}{(\theta_2+\gamma)}c_1}{\theta_1}\mu_2\right), \tag{24}$$

where $A_1(\tau_1), A_2(\tau_1)$ are constants in $\mathbb{R}$. Furthermore, since $q_1^*(\tau_1) = 0$ and $q_2^*(\tau_1)$ is bounded for any initialization, using the fact that $p^*(t) = \nabla_q \Xi(q^*(t))$, it holds that $p_1^*(\tau_1)$ and $p_2^*(\tau_1)$ are bounded for any initialization.

Lastly from the form of $A_1(\tau_1)$ and $A_2(\tau_1)$ in the proof of Lemma 4 (Appendix B.5.1), it can be shown that $A_1(\tau_1)$ and $A_2(\tau_1)$ are bounded if $p_1^*(\tau_1)$ and $p_2^*(\tau_1)$ are bounded.

Now, if the system is initialized with a large queue, τ_1, the time to empty queue 1 for the first time forward in time, is large. As t approaches τ_1 in (24), the sign of the backward switching curve will eventually be governed by the difference between the modified $c\mu/\theta$-indices associated with Class 1 and 2 respectively. In other words, if the states are initialized with $q_1(0) + q_2(0) > M$ for some M sufficiently large, then the optimal control switches priority at some time epoch $s \in (0, \tau^*)$ from Class 2 to Class 1. Similarly to the previous case, it follows that the modified $c\mu/\theta$-rule is optimal at time t if $q_1(t) + q_2(t) > M$. Q.E.D.

B.7. Proof of Proposition 4

PROOF: Theorem 4 indicates that a one-time switch in priority from Class 1 to Class 2 takes place if $c_1\mu_1 < c_2\mu_2$ and $\frac{c_1}{\theta_1}\mu_1 > \left(\frac{c_2}{\theta_2+\gamma}\mu_2 + \frac{\frac{\gamma}{(\theta_2+\gamma)}c_1}{\theta_1}\mu_2\right)$. To derive the policy curve at which (state) the switching takes place, we apply Hamiltonian condition (H). In particular, let (a_1, a_2) be a state where priority is just switched from Class 1 to Class 2, i.e., (a_1, a_2) is on the policy curve, where $a_1 \geq 0$ and $a_2 > 0$. We denote the time of the switching by t_1. We also denote $t_2 > t_1$ as the time Class 2 gets emptied and $t_3 = \tau^* > t_2$ as the time Class 1 gets emptied. Since there is no other switch in priority (Proposition 3) after t_1, starting from time t_1, the dynamic of the adjoint vector for $p^*(t)$ is specified by (ADJ) as

$$\begin{aligned} p_1^*(t) &= \frac{c_1}{\theta_1} + e^{\theta_1 t}K_1 \quad \text{for } t \in [t_1, t_3], \\ p_2^*(t) &= \frac{c_2}{\theta_2+\gamma} + \frac{c_1\gamma}{\theta_1(\theta_2+\gamma)} + \frac{e^{\theta_1 t}\gamma K_1 + e^{(\theta_2+\gamma)t}(-\gamma K_1 + (\gamma - \theta_1 + \theta_2)K_2)}{\gamma - \theta_j + \theta_2} \quad \text{for } t \in [t_1, t_2], \end{aligned} \tag{25}$$

where K_1 and K_2 are constants that depend on the value of $p^*(t_1)$ and t_1. The rest of the analysis is divided into three intervals. For each one of the three intervals, we characterize the state trajectory $q^*(t)$ and the adjoint vector $p^*(t)$. Then, plugging the values of $q^*(t)$ and $p^*(t)$ into the Hamiltonian and utilizing Hamiltonian condition (H) that $H(q^*(t), z^*(t), p^*(t)) = 0$, we are able to characterize the integration constants K_1, K_2 in (25) as well as the policy curve. These steps will become self-explanatory as the proof proceeds.

Case 1: q_1^* is strictly positive or has just reached zero at time t_1. In this case, full service capacity s is assigned to Class 1 at time t_1-.

Interval 1: At time t_1-, we assign s servers to Class 1 and 0 servers to Class 2.

$$\begin{aligned} q_1^{*(1)}(t_1-) &= a_1 \\ q_2^{*(1)}(t_1-) &= a_2 \\ H^{(1)}(q^*(t_1-), z^*(t_1-), p^*(t_1-)) &= c_1a_1 + c_2a_2 + (a_2\gamma - a_1\theta_1 + \lambda_1 - s\mu_1)\left(\frac{c_1}{\theta_1} + K_1\right) \\ &\quad + (-a_2(\theta_2+\gamma) + \lambda_2)\left(\frac{c_1\gamma + c_2\theta_1}{\gamma\theta_1 + \theta_1\theta_2} + K_2\right). \end{aligned}$$

Interval 2: Over $[t_1, t_2)$, we assign 0 servers to Class 1 and s server to Class 2, and Class 2 gets emptied at time t_2.

$$\begin{aligned}
q_1^{*(2)}(t) &= -\frac{1}{\theta_1(\theta_2+\gamma)(\gamma-\theta_1+\theta_2)} e^{-t(\gamma+\theta_1+\theta_2)} \Big(e^{t\theta_1}\gamma\theta_1(a_2(\theta_2+\gamma)-\lambda_2+s\mu_2) \\
&\quad - e^{t(\theta_2+\gamma)}(\theta_2+\gamma)(a_2\gamma\theta_1+a_1\theta_1(\gamma-\theta_1+\theta_2)-\gamma\lambda_1+\theta_1\lambda_1-\theta_2\lambda_1-\gamma\lambda_2+s\gamma\mu_2) \\
&\quad - e^{t(\gamma+\theta_1+\theta_2)}(\gamma-\theta_1+\theta_2)(\theta_2\lambda_1+\gamma(\lambda_1+\lambda_2-s\mu_2)) \Big), \quad t\in[0, t_2-t_1) \\
q_2^{*(2)}(t) &= \frac{1}{\theta_2+\gamma} e^{-t(\theta_2+\gamma)} \left(a_2(\theta_2+\gamma) + (-1+e^{t(\theta_2+\gamma)})(\lambda_2-s\mu_2) \right), \quad t\in[0, t_2-t_1) \\
t_2 - t_1 &= \frac{1}{\theta_2+\gamma} \log\left(\frac{-a_2\gamma - a_2\theta_2 + \lambda_2 - s\mu_2}{\lambda_2 - s\mu_2} \right) \\
H^{(2)}(q^*(t), z^*(t), p^*(t)) &= \frac{1}{\theta_1(\theta_2+\gamma)} \Big(c_1\theta_2\lambda_1 + c_2\theta_1(\lambda_2-s\mu_2) + c_1\gamma(\lambda_1+\lambda_2-s\mu_2) \\
&\quad - \theta_1(\theta_2+\gamma)\left(a_1\theta_1K_1 - \lambda_1K_1 + a_2\theta_2K_2 - \lambda_2K_2 + s\mu_2K_2 + a_2\gamma(-K_1+K_2)\right) \Big).
\end{aligned}$$

Solving $H^{(1)}(q^*(t_1), z^*(t_1), p^*(t_1)) = 0$ and $H^{(2)}(q^*(t), z^*(t), p^*(t)) = 0$ for K_1 and K_2, we get

$$\begin{aligned}
K_1 &= \frac{c_1(-a_2(\theta_2+\gamma)+\lambda_2)\mu_1 + c_2a_2\theta_1\mu_2 + c_1(a_2\gamma+\lambda_1-s\mu_1)\mu_2}{\theta_1(a_2(\theta_2+\gamma)-\lambda_2)\mu_1 + \theta_1(-a_2\gamma+a_1\theta_1-\lambda_1+s\mu_1)\mu_2} \\
K_2 &= -\frac{c_1\gamma + c_2\theta_1}{\gamma\theta_1+\theta_1\theta_2} + \frac{(c_1a_1+c_2a_2)\mu_1}{a_2(\theta_2+\gamma)\mu_1 - \lambda_2\mu_1 - a_2\gamma\mu_2 + (a_1\theta_1-\lambda_1+s\mu_1)\mu_2}.
\end{aligned} \tag{26}$$

Interval 3: Over $[t_2, t_3]$, we assign enough servers to maintain Class 2 at zero and the rest service capacity to Class 1. Class 1 gets emptied at time t_3.

$$\begin{aligned}
q_1^{*(3)}(t) &= \frac{e^{-t\theta_1}}{\theta_1(-\gamma+\theta_1-\theta_2)} \Big(-(a_2\gamma\theta_1 + a_1\theta_1(\gamma-\theta_1+\theta_2) - \gamma\lambda_1 + \theta_1\lambda_1 - \theta_2\lambda_1 - \gamma\lambda_2 + s\gamma\mu_2) \\
&\quad \left(1 + \frac{a_2(\theta_2+\gamma)}{-\lambda_2+s\mu_2}\right)^{-\frac{\theta_1}{\theta_2+\gamma}} - \frac{1}{\mu_2}\big((\lambda_2-s\mu_2)(-(\gamma-\theta_1+\theta_2)\mu_1+\gamma\mu_2) \\
&\quad + e^{t\theta_1}(\gamma-\theta_1+\theta_2)(\lambda_2\mu_1 + (\lambda_1-s\mu_1)\mu_2)\big) \Big), \quad t\in[0, t_3-t_2] \\
q_2^{*(3)}(t) &= 0, \quad t\in[0, t_3-t_2] \\
t_3 - t_2 &= \frac{1}{\theta_1} \log\Bigg(\frac{1}{(\gamma-\theta_1+\theta_2)(\lambda_2\mu_1 + (\lambda_1-s\mu_1)\mu_2)} \Big((\lambda_2-s\mu_2)((\gamma-\theta_1+\theta_2)\mu_1 - \gamma\mu_2) \\
&\quad - \mu_2\left(a_2\gamma\theta_1 + a_1\theta_1(\gamma-\theta_1+\theta_2) - \gamma\lambda_1 + \theta_1\lambda_1 - \theta_2\lambda_1 - \gamma\lambda_2 + s\gamma\mu_2\right) \\
&\quad \left(1 + \frac{a_2(\theta_2+\gamma)}{-\lambda_2+s\mu_2}\right)^{-\frac{\theta_1}{\theta_2+\gamma}} \Big) \Bigg).
\end{aligned}$$

Note that $[t_2, t_3)$ is a boundary arc for q_2^* and an interior arc for q_1^*. As $dq_2^{*(3)}(t) = 0$, we have

$$H^{(3)}(q^*(t), z^*(t), p^*(t)) = p_1^*(t)dq_1^{*(3)}(t) + p_2^*(t)dq_2^{*(3)}(t) + c_1q_1^{*(3)}(t) + c_2q_2^{*(3)}(t) = p_1^*(t)dq_1^{*(3)}(t) + c_1q_1^{*(3)}(t).$$

Since the dynamic of $p_1^*(t)$ does not change over interior arcs, $p_1^*(t)$ still follows (25) over the interval $[t_2, t_3)$. Then, plugging the expression of $q_1^{*(3)}(t)$ into $H^{(3)}(q^*(t), z^*(t), p^*(t))$, we get

$$\begin{aligned}
H^{(3)}(q^*(t), z^*(t), p^*(t)) &= \frac{K_1}{\mu_2(\gamma-\theta_1+\theta_2)} \Big(-\mu_2\big(a_2\gamma\theta_1 + a_1\theta_1(\gamma-\theta_1+\theta_2) - \gamma\lambda_1 + \theta_1\lambda_1 - \theta_2\lambda_1 - \gamma\lambda_2 + s\gamma\mu_2\big) \\
&\quad + (\lambda_2-s\mu_2)((\gamma-\theta_1+\theta_2)\mu_1 - \gamma\mu_2)\left(1 + \frac{a_2(\theta_2+\gamma)}{-\lambda_2+s\mu_2}\right)^{\frac{\theta_1}{\theta_2+\gamma}} \Big) \\
&\quad + \frac{c_1(\lambda_2\mu_1 + (\lambda_1-s\mu_1)\mu_2)}{\theta_1\mu_2}.
\end{aligned}$$

Plugging the value of K_1 in (26) into the equality $H^{(3)}(q^*(t), z^*(t), p^*(t)) = 0$ establishes the relationship (a_1, a_2) must satisfy and retrieves the policy curve in Proposition 4.

Case 2: q_1^* is equal to zero at time t_1 and has been maintained at zero over interval $[t_1 - \epsilon, t_1]$ for some $\epsilon > 0$. In this case, the right amount of service capacity is assigned to Class 1 at time t_1- to maintain q_1^* at zero.

Interval 1: At time t_1-, we assign $(\lambda_1 + \gamma q_2^*(t_1-))/\mu_1$ servers to Class 1 and the rest servers to Class 2.

$$
\begin{aligned}
q_1^{*(1)}(t_1-) &= 0 \\
q_2^{*(1)}(t_1-) &= a_2 \\
H^{(1)}(q^*(t_1-), z^*(t_1-), p^*(t_1-)) &= c_2 a_2 + \left(-a_2(\gamma+\theta_2) + \lambda_2 - s\mu_2 + \frac{(a_2\gamma+\lambda_1)\mu_2}{\mu_1}\right)\left(\frac{c_1\gamma + c_2\theta_1}{\gamma\theta_1 + \theta_1\theta_2} + K_2\right)
\end{aligned}
$$

Interval 2: Over $[t_1, t_2)$, we assign 0 servers to Class 1 and s server to Class 2, and Class 2 gets emptied at time t_2.

$$
\begin{aligned}
q_1^{*(2)}(t) &= -\frac{1}{\theta_1(\theta_2+\gamma)(\gamma-\theta_1+\theta_2)} e^{-t(\gamma+\theta_1+\theta_2)} \Big(e^{t\theta_1}\gamma\theta_1(a_2(\theta_2+\gamma) - \lambda_2 + s\mu_2) \\
&\quad - e^{t(\theta_2+\gamma)}(\theta_2+\gamma)(a_2\gamma\theta_1 - \gamma\lambda_1 + \theta_1\lambda_1 - \theta_2\lambda_1 - \gamma\lambda_2 + s\gamma\mu_2) \\
&\quad - e^{t(\gamma+\theta_1+\theta_2)}(\gamma-\theta_1+\theta_2)(\theta_2\lambda_1 + \gamma(\lambda_1+\lambda_2-s\mu_2))\Big), \quad t \in [0, t_2 - t_1) \\
q_2^{*(2)}(t) &= \frac{1}{\theta_2+\gamma} e^{-t(\theta_2+\gamma)} \left(a_2(\theta_2+\gamma) + (-1 + e^{t(\theta_2+\gamma)})(\lambda_2 - s\mu_2)\right), \quad t \in [0, t_2 - t_1) \\
t_2 - t_1 &= \frac{1}{\theta_2+\gamma} \log\left(\frac{-a_2\gamma - a_2\theta_2 + \lambda_2 - s\mu_2}{\lambda_2 - s\mu_2}\right) \\
H^{(2)}(q^*(t), z^*(t), p^*(t)) &= \frac{1}{\theta_1(\theta_2+\gamma)} \Big(c_1\theta_2\lambda_1 + c_2\theta_1(\lambda_2 - s\mu_2) + c_1\gamma(\lambda_1 + \lambda_2 - s\mu_2) \\
&\quad - \theta_1(\theta_2+\gamma)\left(-\lambda_1 K_1 + a_2\theta_2 K_2 - \lambda_2 K_2 + s\mu_2 K_2 + a_2\gamma(-K_1 + K_2)\right)\Big).
\end{aligned}
$$

Solving $H^{(1)}(q^*(t_1), z^*(t_1), p^*(t_1)) = 0$ and $H^{(2)}(q^*(t), z^*(t), p^*(t)) = 0$ for K_1 and K_2, we get

$$
\begin{aligned}
K_1 &= \frac{c_1(-a_2(\theta_2+\gamma) + \lambda_2)\mu_1 + c_2 a_2\theta_1\mu_2 + c_1(a_2\gamma + \lambda_1 - s\mu_1)\mu_2}{\theta_1(a_2(\theta_2+\gamma) - \lambda_2)\mu_1 + \theta_1(-a_2\gamma - \lambda_1 + s\mu_1)\mu_2} \\
K_2 &= -\frac{c_1\gamma + c_2\theta_1}{\gamma\theta_1 + \theta_1\theta_2} + \frac{c_2 a_2\mu_1}{a_2(\theta_2+\gamma)\mu_1 - \lambda_2\mu_1 - a_2\gamma\mu_2 + (-\lambda_1 + s\mu_1)\mu_2}.
\end{aligned} \tag{27}
$$

Interval 3: Over $[t_2, t_3]$, we assign enough servers to maintain Class 2 at zero and the rest service capacity to Class 1. Class 1 gets emptied at time t_3.

$$
\begin{aligned}
q_1^{*(3)}(t) &= \frac{e^{-t\theta_1}}{\theta_1(-\gamma+\theta_1-\theta_2)} \Big(-(a_2\gamma\theta_1 - \gamma\lambda_1 + \theta_1\lambda_1 - \theta_2\lambda_1 - \gamma\lambda_2 + s\gamma\mu_2) \\
&\quad \left(1 + \frac{a_2(\theta_2+\gamma)}{-\lambda_2 + s\mu_2}\right)^{-\frac{\theta_1}{\theta_2+\gamma}} - \frac{1}{\mu_2}\big((\lambda_2 - s\mu_2)(-(\gamma-\theta_1+\theta_2)\mu_1 + \gamma\mu_2) \\
&\quad + e^{t\theta_1}(\gamma-\theta_1+\theta_2)(\lambda_2\mu_1 + (\lambda_1 - s\mu_1)\mu_2)\big)\Big), \quad t \in [0, t_3 - t_2] \\
q_2^{*(3)}(t) &= 0, \quad t \in [0, t_3 - t_2] \\
t_3 - t_2 &= \frac{1}{\theta_1} \log\Bigg(\frac{1}{(\gamma-\theta_1+\theta_2)(\lambda_2\mu_1 + (\lambda_1 - s\mu_1)\mu_2)} \Big((\lambda_2 - s\mu_2)((\gamma-\theta_1+\theta_2)\mu_1 - \gamma\mu_2) \\
&\quad - \mu_2\left(a_2\gamma\theta_1 - \gamma\lambda_1 + \theta_1\lambda_1 - \theta_2\lambda_1 - \gamma\lambda_2 + s\gamma\mu_2\right)\left(1 + \frac{a_2(\theta_2+\gamma)}{-\lambda_2 + s\mu_2}\right)^{-\frac{\theta_1}{\theta_2+\gamma}}\Big)\Bigg).
\end{aligned}
$$

Note that $[t_2, t_3)$ is a boundary arc for q_2^* and an interior arc for q_1^*. As $dq_2^{*(3)}(t) = 0$, we have

$$H^{(3)}(q^*(t), z^*(t), p^*(t)) = p_1^*(t) dq_1^{*(3)}(t) + p_2^*(t) dq_2^{*(3)}(t) + c_1 q_1^{*(3)}(t) + c_2 q_2^{*(3)}(t) = p_1^*(t) dq_1^{*(3)}(t) + c_1 q_1^{*(3)}(t).$$

Since the dynamic of $p_1^*(t)$ does not change over interior arcs, $p_1^*(t)$ still follows (25) over the interval $[t_2, t_3)$. Then, plugging the expression of $q_1^{*(3)}(t)$ into $H^{(3)}(q^*(t), z^*(t), p^*(t))$, we get

$$\begin{aligned} H^{(3)}(q^*(t), z^*(t), p^*(t)) = \frac{K_1}{\mu_2(\gamma - \theta_1 + \theta_2)} \Big(& -\mu_2\big(a_2\gamma\theta_1 - \gamma\lambda_1 + \theta_1\lambda_1 - \theta_2\lambda_1 - \gamma\lambda_2 + s\gamma\mu_2\big) \\ & + (\lambda_2 - s\mu_2)((\gamma - \theta_1 + \theta_2)\mu_1 - \gamma\mu_2) \left(1 + \frac{a_2(\theta_2 + \gamma)}{-\lambda_2 + s\mu_2}\right)^{\frac{\theta_1}{\theta_2 + \gamma}} \Big) \\ & + \frac{c_1(\lambda_2\mu_1 + (\lambda_1 - s\mu_1)\mu_2)}{\theta_1\mu_2}. \end{aligned}$$

Plugging the value of K_1 in (27) into the equality $H^{(3)}(q^*(t), z^*(t), p^*(t)) = 0$ establishes the relationship a_2 must satisfy in order for priority to be switched from P_1 to P_2 given that q_2^* is at level a_2 and q_1^* has been maintained at zero for some amount of time. It is easy to see that setting $H^{(3)}(q^*(t), z^*(t), p^*(t)) = 0$ in Case 2 retrieves the point $(0, a_2)$ on the switching curve established in Case 1.

It is important to note that the switching point $(0, a_2)$ analyzed in Case 2 assumes that q_1^* has been maintained at zero before priority is switched. On the other hand, the switching point $(0, a_2)$ on the policy curve derived in Case 1 assumes that q_1^* just hits zero when priority is switched from P_1 to P_2. It is well expected that the switching points in the two cases coincide with each other. Our proof rigorously verified this. Q.E.D.

Appendix C: Special Cases without Abandonment or Degradation

As mentioned in Remark 2, the same line of analysis allows us to establish optimality results in analogue to Theorem 4 when some or all of the abandonment and degradation rates are zero. In particular, if $\theta_1 = \theta_2 = \gamma = 0$, then the $c\mu$-rule is optimal; if $\gamma = 0$ but $\theta_1, \theta_2 > 0$, the $c\mu$-rule is optimal close to the origin and the standard $c\mu/\theta$-rule is optimal when the states are far from the origin.

Corollary 1 *If $\gamma = \theta_1 = \theta_2 = 0$ and $s > \lambda_1/\mu_1 + \lambda_2/\lambda_2$, the $c\mu$-rule is optimal for the transient fluid optimal control problem (F2′).*

Corollary 2 *Suppose $\gamma = 0$, $\theta_1, \theta_2 > 0$, and $s > \lambda_1/\mu_1 + \lambda_2/\lambda_2$. For the transient fluid optimal control problem (F2′):*

I. If the $c\mu$-rule and the standard $c\mu/\theta$-rule both prioritize class i, $i = 1, 2$, then the strict priority rule to class i is optimal for any $t \in [0, \tau]$.

II. If the $c\mu$-rule prioritizes class i but the standard $c\mu/\theta$-rule prioritizes class j, for $i \neq j, i, j = 1, 2$, then there exist positive real numbers ϵ and M such that it is optimal to prioritize class i when $q_1 + q_2 < \epsilon$ and prioritize class j when $q_1 + q_2 > M$. Furthermore, the optimal scheduling policy switches priority at most once over the transient time horizon $[0, \tau]$.

Since the proof using Pontryagin's Minimum Principle for the case with $\theta_1 = \theta_2 = \gamma = 0$ is very concise and nicely illustrates the main idea of our proof strategy, we present the proof of Corollary 1 here.

PROOF: [Proof of Corollary 1] Suppose without loss of generality that $c_1\mu_1 > c_2\mu_2$. The queue length process evolves as

$$dq_1(t) = \lambda_1 - \mu_1 z_1(t) \quad \text{and} \quad dq_2(t) = \lambda_2 - \mu_2 z_2(t).$$

The Hamiltonian is

$$\begin{aligned} H(q(t), z(t), p(t)) &= p_1(t)dq_1(t) + p_2(t)dq_2(t) + c_1 q_1(t) + c_2 q_2(t) \\ &= p_1(t)\left(\lambda_1 - \mu_1 z_1(t)\right) + p_2(t)\left(\lambda_2 - \mu_2 z_2(t)\right) + c_1 q_1(t) + c_2 q_2(t). \end{aligned}$$

The augmented Halmiltonian takes the form

$$\begin{aligned} L(q(t), z(t), p(t), \eta(t), \xi(t)) &= H(x, s, p) + \eta(t)^T g(q(t)) + \xi(t)^T h(z(t)) \\ &= p_1(t)\left(\lambda_1 - \mu_1 z_1(t)\right) + p_2(t)\left(\lambda_2 - \mu_2 z_2(t)\right) + c_1 q_1(t) + c_2 q_2(t) \\ &\quad + \eta_1(t)(-q_1(t)) + \eta_2(t)(-q_2(t)) + \xi_1(t)(z_1(t) + z_2(t)) \\ &\quad + \xi_2(t)(-z_1(t)) + \xi_3(t)(-z_2(t)). \end{aligned}$$

Since $dp^*(t) = -\nabla_q L(q^*(t), z^*(t), p^*(t), \eta^*(t), \xi^*(t))$, we have

$$dp_1^*(t) = -c_1 + \eta_1^*(t) \quad \text{and} \quad dp_2^*(t) = -c_2 + \eta_2^*(t). \tag{28}$$

Hence,

$$p_1^*(t) = -c_1 t + \int_0^t \eta_1^*(s)ds + K_1 \quad \text{and} \quad p_2^*(t) = -c_2 t + \int_0^t \eta_2^*(s)ds + K_2,$$

where K_1 and K_2 are constants that depend on $p^*(0)$.

The switching curve is

$$\psi(t) = \mu_1 p_1^*(t) - \mu_2 p_2^*(t).$$

Proposition 1 still holds in this case. Hence, when the queue length process is arbitrarily close to the origin, the $c\mu$-rule is optimal and Class 1 should be given strict priority. Let τ_N be the last time epoch (forward in time) $q_1^*(t)$ hits zero, i.e.,

$$\tau_N = \sup\left\{t : q_1^*(t) = 0,\, q_1^*(t-\epsilon) > 0 \text{ for any } \epsilon > 0\right\}.$$

Following the same lines arguments as in Lemmas 2 and 3, we have the switching curve $\psi(t) = 0$ for $t \geq \tau_N$.

We next characterize the optimal control before τ_N. To this end, observe that by construction, both queues are strictly positive before τ_N. Therefore, there exists a non-trivial period $[0, \beta], \beta < \tau_N$, such that for $t \in [0, \beta]$, the backward switching curve is characterized by

$$\psi(\tau_N - t) = \psi(\tau_N) + (c_1\mu_1 - c_2\mu_2)t + \left(\mu_2 \int_{\tau_N - t}^{\tau_N} \eta_2^*(s)ds - \mu_1 \int_{\tau_N - t}^{\tau_N} \eta_1^*(s)ds\right) = (c_1\mu_1 - c_2\mu_2)t. \tag{29}$$

Since $c_1\mu_1 > c_2\mu_2$, the significance of (29) is that strict priority must be assigned to Class 1 during this period. Moreover, as no queue has the possibility to hit zero over this period, the characterization of the switching curve (29) indeed holds for all $t \in [0, \tau_N]$. Namely, strict priority to Class 1 is optimal throughout $[0, \tau^*]$. Q.E.D.

C.1. Full Characterization of the Dual Vectors When $\gamma = \theta_1 = \theta_2 = 0$

When establishing the optimal scheduling policy, we use Pontryagin's Minimum Principle to derive structural properties of the dual vectors $(p^*(t),\ \eta^*(t),\ \xi^*(t))$ without characterizing their expressions explicitly. The latter step can be prohibitively hard for systems with convoluted dynamics, as is the case for our model with both abandonment and degradation. On the other hand, for simplified systems without abandonment or degradation, we can provide a full characterization of the dual vectors. We next illustrate the derivation.

By Corollary 1, the $c\mu$-rule is optimal at all time for systems without abandonment and without degradation. Suppose without loss of generality that the $c\mu$-rule prioritizes Class 1, i.e., $c_1\mu_1 > c_2\mu_2$. In this case, the value function associated with state (a_1, a_2) is equal to the cost of emptying the system under P_1 when the system is initialized at (a_1, a_2). We can then calculate the value function by solving the state trajectory and the cost directly. Specifically, the value function takes the form

$$\Xi(a_1,a_2) = \frac{1}{2(\lambda_1 - s\mu_1)}\left(-c_1 a_1^2 + \frac{c_2\left(a_2^2\mu_1(-\lambda_1+s\mu_1) + a_1^2\lambda_2\mu_2 - 2a_1a_2(\lambda_1 - s\mu_1)\mu_2\right)}{\lambda_2\mu_1 + (\lambda_1 - s\mu_1)\mu_2}\right).$$

For a fixed initial condition, q_0, let $q^*(t)$ denote the (optimal) state trajectory under P_1, which can be solved directly. Along the optimal state trajectory, τ_1 is the time epoch when q_1^* first gets emptied. q_1^* is then maintained at zero after time τ_1, until q_2^* reaches zero at time τ^*.

Using the fact that there exists an adjoint vector $p^*(t) = \nabla_q \Xi(q_1^*(t), q_2^*(t))$, we have

$$\begin{aligned} p_1^*(t) &= \frac{1}{\lambda_1 - s\mu_1}\left(-c_1 q_1^*(t) + \frac{c_2(-q_2^*(t)\lambda_1 + q_1^*(t)\lambda_2 + sq_2^*(t)\mu_1)\mu_2}{\lambda_2\mu_1 + (\lambda_1 - s\mu_1)\mu_2}\right), \quad t\in[0,\tau^*] \\ p_2^*(t) &= \frac{c_2(q_2^*(t)\mu_1 + q_1^*(t)\mu_2)}{-\lambda_2\mu_1 - \lambda_1\mu_2 + s\mu_1\mu_2}, \quad t\in[0,\tau^*]. \end{aligned} \tag{30}$$

The switching curve is then given by

$$\psi(t) = \mu_1 p_1^*(t) - \mu_2 p_2^*(t), \quad t\in[0,\tau^*],$$

where $p^*(t)$ is calculated explicitly in (30).

In addition, it follows from (28) that at all regular points of $p_i^*(t)$ where $p_i^*(t)$ is differentiable with respect to t, $\eta_i^*(t) = dp_i^*(t) + c_i$, $i = 1, 2$. In this case,

$$\begin{aligned} \eta_1^* &= \begin{cases} 0, & t\in[0,\tau_1] \\ c_1 - c_2\mu_2/\mu_1, & t\in[\tau_1,\tau^*] \end{cases} \\ \eta_2^* &= 0, \quad t\in[0,\tau^*]. \end{aligned}$$

Lastly, we can infer from Transversality condition (T) and Complementarity condition (C) that

$$\begin{aligned} \xi_1^*(t) &= \mu_1 p_1^*(t), \quad t\in[0,\tau^*] \\ \xi_2^*(t) &= 0, \quad t\in[0,\tau^*] \\ \xi_3^*(t) &= \begin{cases} \mu_1 p_1^*(t) - \mu_2 p_2^*(t), & t\in[0,\tau_1] \\ 0, & t\in[\tau_1,\tau^*]. \end{cases} \end{aligned}$$

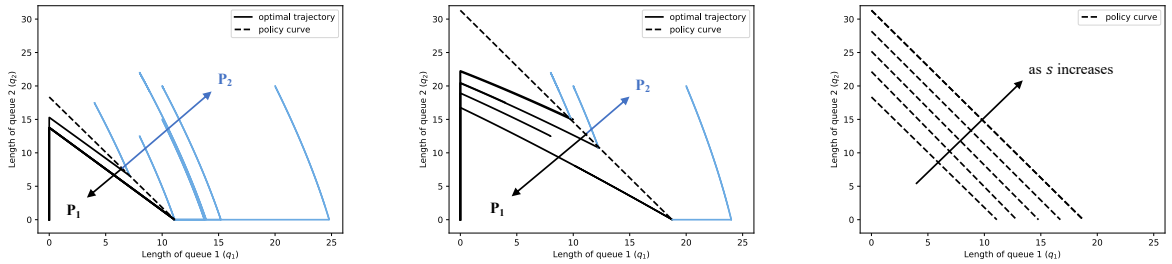
We comment that similar analysis to delineate the dual vectors is not replicable for the general system with both abandonment and degradation. Consider the scenario where the $c\mu$-rule prioritizes Class 2 and the modified $c\mu/\theta$-rule prioritizes Class 1. With the policy curve explicitly characterized in Proposition 4, one

can potentially calculate the value function (by calculating the optimal state trajectory starting from any state) and derive the dual vectors as above. However, due to the intertwined system dynamics introduced by degradation, we have not found a way to fully characterize the optimal state trajectory analytically, particularly in the segment where strict priority is given to Class 1. In the other scenario where the $c\mu$-rule prioritizes Class 1 and the modified $c\mu/\theta$-rule prioritizes Class 2, the analysis is hindered by not being able to characterize the policy curve as well as the optimal state trajectory.

Appendix D: Supplementary Sensitivity Analysis

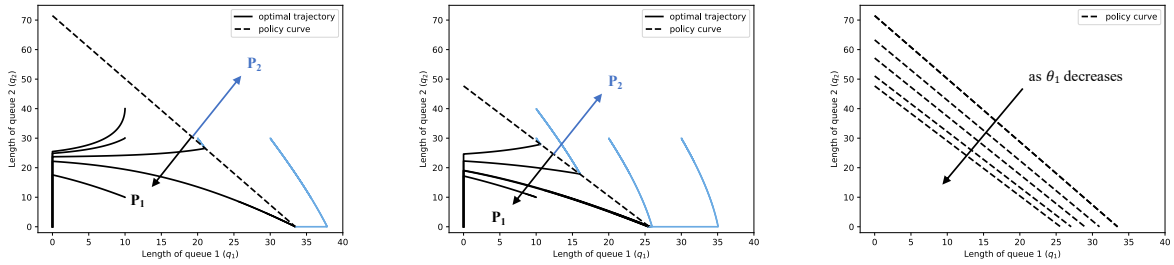
In this section, we provide some additional numerical experiments on the policy curve from P_2 to P_1 when we vary the number of servers s (Figure 18) or the abandonment rate of Class 1 θ_1 (Figure 19). The policy curves (plotted in dashed line) is interpolated from the switching epochs of the optimal trajectories. We observe that the region where we prioritize Class 1 increases as s increases (Figure 18 (c)), and decreases as θ_1 decreases (Figure 19 (c)).

Figure 18 **Sensitivity analysis of the policy curve with respect to s**
($\lambda_1 = 10, \lambda_2 = 20, \mu_1 = 1, \mu_2 = 2.5, \theta_1 = 0.1, \theta_2 = 0.2, \gamma = 0.4, c_1 = 5, c_2 = 1$)



(a) $s = 25$ (b) $s = 29$ (c) Increasing $s = 25, 26, 27, 28, 29$

Figure 19 **Sensitivity analysis of the policy curve with respect to θ_1**
($\lambda_1 = 10, \lambda_2 = 20, \mu_1 = 1, \mu_2 = 2.5, \theta_2 = 0.45, \gamma = 0.4, s = 26, c_1 = 5, c_2 = 1$)



(a) $\theta_1 = 0.18$ (b) $\theta_1 = 0.02$ (c) Decreasing $\theta_1 = 0.18, 0.14, 0.1, 0.06, 0.02$