Optimal Scheduling of Proactive Service with Customer Deterioration and Improvement

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

Service systems are typically limited resource environments where scarce capacity is reserved for the most urgent customers. However, there has been a growing interest in the use of proactive service when a less urgent customer may become urgent while waiting. On one hand, providing service for customers when they are less urgent could mean that fewer resources are needed to fulfill their service requirement. On the other hand, utilizing limited capacity for customers who may never need the service in the future takes the capacity away from other more urgent customers who need it now. To understand this tension, we propose a multi-server queueing model with two customer classes: moderate and urgent. We allow customers to transition classes while waiting. In this setting, we characterize how moderate and urgent customers should be prioritized for service when proactive service for moderate customers is an option. We identify an index, 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 times, abandonments, and transitions between customer classes.

Key words: Proactive Service, Multi-Class Queue, Optimal Control, Equilibrium, Transient Analysis

1. Introduction

With recent advancements of predictive analytics and data availability, considerable efforts have been made to develop predictive tools for service systems. For example, in healthcare settings, 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. In call centers, predictive models have been developed to identify customers who are likely to contact their insurance company based on past claims data (Jerath et al. 2015).

From the operations perspective, predictive information on customers' future service needs brings the opportunity of developing approaches to provide effective proactive service and, potentially, improve system performance. In healthcare, there is 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 that forecast patient deterioration, can help reduce treatment delays and improve patient outcomes (Hu et al. 2018). In the insurance company call center example, Jerath et al.

(2015) advocate reaching out proactively to customers who have a high probability of calling to increase customer satisfaction and reduce peak demand.

Isolating the potential impact of proactive service is not straightforward. On one hand, advancing service for customers when they are less urgent could mean that fewer resources are needed to fulfill their service requirement. This has the potential benefit of reducing the overall workload of the system. On the other hand, utilizing limited capacity for customers who are less critical may take capacity away from other more critical customers whose service needs are more urgent. Moreover, some of these less critical customers may be satisfied without ever needing the critical service. Thus, providing proactive service 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 service, we propose a multi-class queueing system that explicitly models customers' deterioration and improvement behavior, and study the optimal scheduling policy for proactive service based on the model.

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 service systems, there are very few works analyzing proactive service with predictive information about customers' future needs (see Section 1.1 for a detailed review of some related works). Our modeling approach aims to provide a systematic way to capture the key tradeoffs in the limited resource environment: the potential benefit of serving customers early on with fewer resources versus the potential cost of delaying service for the more urgent customers and generating more overall workload to the system. Moreover, our analysis provides insights on how the accuracy of the predictive information affects the prioritization of services.

We conduct analysis on both the long-run average performance and the transient performance, with the focus on developing structural insights into the optimal scheduling policy. The long-run average performance analysis provides guidance on scheduling proactive service when the system is in its "normal" state of operation. That said, service systems often operate in a highly non-stationary environment. A surge in demand due to random shocks, e.g., disease outbreaks or mass casualty events for hospitals and insurance companies, weather patterns resulting in mass flight cancellations for airline call centers, etc., can bring the system far from its normal state of operation. It is thus important to study the transient optimal control and to develop an understanding of the most cost-effective way to bring the system back to normal.

Our analysis quantifies the merits of proactive service. We are able to characterize settings where proactive service can be beneficial and others where it is better to focus all resources on the most urgent customers. Our main contributions can be summarized as follows.

Queueing model with dynamic class types. We propose a Markovian multi-server queue with two customer classes: urgent and moderate. The key feature we incorporate is that a moderate customer who does not receive timely service may resolve their problem and leave without requiring service, or may deteriorate and become an urgent customer. Similarly, an urgent customer who does not receive service may leave the system, e.g., through adversarial events such as abandonment, or may improve to the moderate class. If we assume there is a classifier (e.g. an early warning system) that classifies potentially risky customers into the moderate class, then the proportion of moderate customers 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 service.

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 corresponding modified $c\mu/\theta$-index accounts for the class-transition dynamics in addition to the holding costs, service rates, and abandonment rates. The exact expression of this index lends itself to a very intuitive interpretation of which parameters – pre or post transition of class types – impact the performance.

Transient optimal control. To minimize the cumulative transient cost (until reaching the equilibrium point with zero queue) for the fluid model, we show that the optimal policy may switch priority depending on the interplay between two indices: 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 relatively easily. We conduct sensitivity analysis on the policy curve and quantify the effect of 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 Minimum 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. Section 5 considers some model extensions. Lastly, we conclude in Section 6. All the proofs are provided in the appendix.

1.1. Related Literature

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

Proactive Service. There are a number of works on proactive service in service systems, most of which focus on optimal screening strategies in healthcare. 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. In various service settings, there are also works modeling proactive service when providers have advance information about customers' future service needs, but they do not model the dynamic change of customer class types 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 service into account. These aspects include a limited resource environment, customer deterioration and amelioration, different service needs, and different waiting costs. We also derive structural insights on the optimal scheduling policy for proactive service.

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 a lot of cases, a simple index-based policy can be shown to be optimal. For example, the $c\mu$-rule is shown to be optimal for a 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 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) model the wait list for donated organs as a multi-class overloaded queue. Disease evolution is captured by allowing customers to transition 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. 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). Most of these works rely on exact or numerical analysis of the corresponding Markov decision process (MDP), where the analysis can become prohibitively challenging when the scale of the system becomes large or more features 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, even without the added complexity of optimizing over different control policies. Only a limited set of numerical and approximation techniques have been developed for transient performance analysis. These include inverting Laplace transforms (Abate and Whitt 1988, 2006), heavy-traffic asymptotics (Honnappa et al. 2015), etc. Our study uses a fluid approximation and employs tools from the optimal control theory for dynamical systems to derive the optimal transient scheduling policy; see (Sethi and Thompson 2000, Grass et al. 2008) for an overview of continuous-time control theory and its wide applications. In particular, we utilize Pontryagin's Minimum Principle (Hartl et al. 1995), which is a common methodology used for both linear and nonlinear continuous control problems. 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 static (fixed) class types. 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 component of dynamic class types is a highly nontrivial 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 navigating optimal control problems with state constraints; this approach may shed insights for other queueing control problems.

2. The Model

To explore the potential benefits of proactive service, we propose a Markovian two-class multi-server queueing system as depicted in Figure 1. Customers (jobs) are defined by their need for service. Without loss of generality, we refer to Class 1 as the *urgent class*: those with immediate need for service. Focusing resource allocation to just these customers is a common approach in the service operations literature. In this work, we also consider a *moderate class* (Class 2): those who currently do not need as high level of service as Class 1, but are at risk of becoming urgent. The novel feature we incorporate is dynamic class types. We allow Class 2 customers to transition to Class 1 while waiting and refer to this as a degradation. *Proactive service* (preventive service), i.e. providing service to Class 2 customers, can prevent Class 2 customers from becoming Class 1 customers. We also allow Class 1 customers to transition to Class 2 while waiting, and refer to this as improvement. Note that, mathematically, our model is symmetric. We differentiate customers as urgent and moderate to better facilitate discussions of real-world applications and derive managerial insights.

These type of dynamics may arise in a lot of service operations applications. For example, in hospitals, Class 1 customers may correspond to patients who are physiologically unstable and in need of care in an Intensive Care Unit (ICU), while Class 2 customers may correspond to patients in the general medical ward who are *at risk* of deteriorating. Those who are in the general medical ward, but are known to have no risk of needing ICU care, would be outside of our modeling framework. Many patients in 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 severely critical is becoming a reality (Hu et al. 2018). What remains is to understand when and how such care should be utilized.

Another example is airline call centers following massive flight cancellations, e.g. due to severe weather issues. In this case, urgent customers are those with complicated and urgent travel needs, and thus require immediate assistance from the agents. Moderate customers are those who can either rebook through an agent or rebook online themselves. However, some moderate customers may develop negative emotions while trying to find another flight themselves and may require more service time to satisfactorily address their needs once they have joined the urgent queue for agent assistance (Altman et al. 2019).

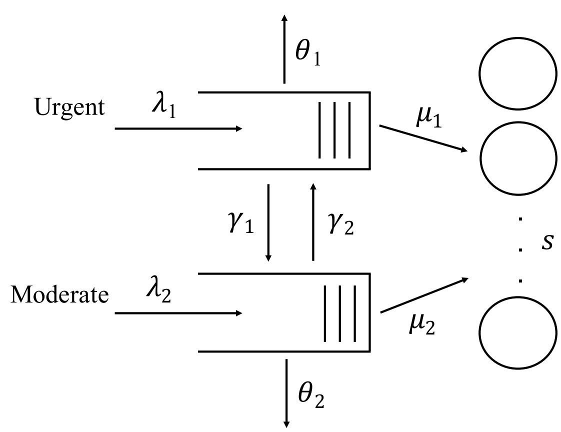
We consider a system with s identical servers, i.e. they offer the same quality of service. Class i customers, $i=1,2$, arrive to the system according to a time-homogeneous Poisson process with rate λ_i. Class 1 customers have independent and identically distributed service requirements following an exponential distribution with mean $1/\mu_1$. While waiting to receive service, a Class 1 customer may improve and transition to the Class 2 queue according to an exponentially distributed clock with rate γ_1. A Class 1 customer can also abandon the queue if its waiting time exceeds its patience time. The patience time is exponentially distributed with mean $1/\theta_1$ and is independent of everything else. For Class 1 customers, one can interpret this abandonment as an undesirable event. For example, in the healthcare setting, urgent patients could be placed in an off-service unit, transferred to another hospital, or even die. In a call center setting, customers may abandon and their patronage may be lost.

Class 2 customers 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 service to a Class 2 customer, its service time is exponentially distributed with mean $1/\mu_2$. Deterioration and abandonment happen according to two independent exponential clocks with rate γ_2 and θ_2, respectively. For Class 2 customers, one can interpret the abandonment as a desirable outcome. For example, in the healthcare example, the abandonment for moderate patients can be the event that the patient is no longer at risk for deterioration, i.e., the patient self-cures.

Remark 1 *We make two remarks about our modeling assumptions. First, the Markovian assumption on system primitives, including exponential deterioration and upgrade times, is quite common in the literature; see, for example, Down and Lewis (2010), Cao and Xie (2016), Xie et al. (2017). This is in part because the assumption greatly facilitates the theoretical analysis of system dynamics. Second, in practice, it is natural to assume that the service times while in Class 1 and Class*

2 for the same customer should be correlated. This can be achieved by assuming that the "base" service requirement for a customer is characterized by a rate 1 *exponential random variable,* V_0*. When the customer is served as Class* i*,* $i = 1, 2$*, its service time is* V_0/μ_i*. In this case, we keep the marginal service time distribution in Class* i *as exponential with rate* μ_i *while maintaining the order of the service times, e.g., if* $\mu_1 < \mu_2$*, then* $V_0/\mu_1 > V_0/\mu_2$ *with probability* 1*. Due to the memoryless property of exponential random variables, introducing such correlation will not affect the dynamics of the system. For simplicity in the subsequent development, we treat these service times as independent random variables.*

Figure 1 **Two-class queue**



We next provide a useful interpretation for the ratio $\phi := \gamma_2/(\theta_2 + \gamma_2)$. Note that if no proactive service is provided to Class 2 customers, $\gamma_2/(\theta_2 + \gamma_2)$ of them will deteriorate into the urgent class. Suppose Class 2 customers are identified via a classifier that determines customers who are "at risk" of deteriorating (e.g., Escobar et al. (2012)), then $\gamma_2/(\theta_2 + \gamma_2)$ can be interpreted as the true positive rate of this classifier. That is, it measures the accuracy of the classifier. For example, if we know with certainty that Class 2 customers will eventually deteriorate into Class 1 customers, then $\theta_2 = 0$ and $\gamma_2/(\theta_2 + \gamma_2) = 1$.

To understand the key tradeoffs we are trying to capture with this model, we start by discussing the extreme case where $\gamma_1 = \theta_1 = \theta_2 = 0$. In this case, if no service is provided to Class 2 customers, each Class 2 customer generates an average workload of $\gamma_2/(\mu_1(\theta_2 + \gamma_2))$ to the system. This is because $\gamma_2/(\theta_2 + \gamma_2)$ of the Class 2 customers will deteriorate into Class 1 and all Class 1 customers must be served. On the other hand, if we can provide proactive service to all Class 2 customers, then each Class 2 customer will generate an average workload of $1/\mu_2$. The magnitude of $\gamma_2/(\theta_2 + \gamma_2)$ impacts whether it may be more or less beneficial, from a workload perspective, to provide proactive service to Class 2 customers. Of course, the actual problem we are facing is more complicated than minimizing the system workload. In particular, the different waiting, abandonment, and/or class-transition 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 customers in the system at time t, $t \geq 0$. We denote by $Z_i(t)$ the number of servers assigned to Class i customers, 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 (X, Q) only. In particular, the policy is non-anticipating. Under this class of scheduling policies, which we denote by set $\mathscr{S}$, $\{(X(t), Q(t)) : t \geq 0\}$ forms a Markov process.

As the process $\{(X(t), Q(t)) : t \geq 0\}$ 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 customers that have abandoned the Class i queue by time t, and $\Gamma_i^{\Pi}(t)$ as the cumulative number of customers that have changed type from Class i to the other by time t. In what follows, we shall drop the superscript Π when it can be understood from the context.

We incur costs for all customers who wait, abandon, or transition classes. In particular, for each Class i customer, we denote h_i as the holding cost per unit time waiting in queue, α_i as the fixed cost of abandonment, and ν_i as the fixed cost of changing class types. Our goal is to minimize the aggregated cost incurred, namely,

$$\mathbb{E}\left[\int_0^T \sum_{i=1,2} h_i Q_i(t) dt + \sum_{i=1,2} \left(\alpha_i R_i(T) + \nu_i \Gamma_i(T)\right)\right]. \tag{1}$$

Note that under the Markovian modeling assumption, we have

$$\mathbb{E}\left[R_i(T)\right] = \theta_i \, \mathbb{E}\left[\int_0^T Q_i(t) dt\right] \quad \text{and} \quad \mathbb{E}\left[\Gamma_i(T)\right] = \gamma_i \, \mathbb{E}\left[\int_0^T Q_i(t) dt\right], \quad i=1,2.$$

Thus, (1) can be equivalently written as

$$\mathbb{E}\left[\int_0^T \left(c_1 Q_1(t) + c_2 Q_2(t)\right) dt\right], \quad \text{where } c_i = h_i + \alpha_i \theta_i + \nu_i \gamma_i \text{ for } i=1,2.$$

This implies that we can incorporate the abandonment costs and the class-transition costs into the holding costs. In what follows, we shall use c_1 and c_2 to denote the "generalized" holding costs.

Note that we defined Class 1 as the *urgent* class in order to facilitate interpretation and draw managerial insights. For example, this can correspond to defining Class 1 customers as those having a higher generalized holding cost, i.e., $c_1 > c_2$.

Remark 2 *Due to our Markovian assumptions and our holding cost criteria, the system performance is agnostic to the order customers are served within a class. The policy development focuses on which class to prioritize; customers within the same class can be served in any order, e.g., first-come-first-served. That said, when looking at individual customers, depending on the transition dynamics and scheduling policies, it is possible that a customer's waiting time may* increase *or* decrease *after changing class. Indeed, our policy development leverages the fact that customers may be able to afford to wait longer after transition (due to the smaller holding cost) and so we can focus resources to the higher priority customers. If we wanted to take waiting-time related fairness into account, we would need to modify our objective function to add some cost of fairness or adapt the optimization problem to incorporate a fairness constraint. Quantifying the fairness of a scheduling policy is an interesting and challenging problem which is outside the scope of this work. We refer to Wierman (2011) for more discussions on the topic.*

In this paper, we focus on two cost measures. One is the long-run average cost; the other is the cumulated transient cost. The two cost formulations have different focuses and are both relevant in practice. The long-run average cost formulation involves minimizing the cost when the system is in its "normal" state of operation. When shocks bring the system far from its normal state of operation, the transient cost formulation aims to minimize the cost incurred to bring the system back to normal. More precisely, the **long-run average cost minimization problem** is

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

It is significant that the long-run average problem is not capacity specific, namely, the system can be staffed to operate in an underloaded or overloaded regime. For the **cumulated transient cost minimization problem**, we define

$$\mathcal{T} := \inf\{t \geq 0 : Q_1(t) + Q_2(t) = 0\}.$$

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

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

These cost minimization problems are MDP's. 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 into the optimal scheduling policy, we employ a fluid approximation approach; a similar method has been used in, for example, Whitt (2006), Perry and Whitt (2009), Atar et al. (2010).

2.1. The Fluid Model

To construct the fluid model, we replace the stochastic arrival, service, abandonment and class-transition 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 take the form

$$\begin{aligned} dq_1(t) &= \lambda_1 - z_1(t)\mu_1 - \theta_1 q_1(t) - \gamma_1 q_1(t) + \gamma_2 q_2(t) \\ dq_2(t) &= \lambda_2 - z_2(t)\mu_2 - \theta_2 q_2(t) - \gamma_2 q_2(t) + \gamma_1 q_1(t). \end{aligned} \tag{2}$$

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 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 $q_i(t)$ 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 the context.

We comment that the fluid dynamics capture the mean dynamics of the stochastic system well, as we will demonstrate later with numerical experiments. 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 scaling (Whitt 2002, Reed and Ward 2008). In this limiting regime, we scale up the arrival rates and the service rates while scale down the space

(Alternatively, we can scale up time while scale down the abandonment rates, the class-transition rates, and the space). The number of servers is held fixed[1].

2.2. Problem Formulation

In this section, we introduce the fluid counterparts of the stochastic cost minimization problems. Note that for the long-run average optimization problem, we only require that the amount of service capacity is non-negative, $s \geq 0$.

Fluid long-run average cost optimization problem:

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

For the transient optimization problem, let $\tau := \inf\{t \geq 0 : q_1(t) + q_2(t) = 0\}$, which is the first time when the total fluid queue reduces to 0. We assume that there is ample capacity s such that for any $q(0) = q_0$, $\tau < \infty$. As will be explained in Section 4, the precise condition is $s > \lambda_1/\mu_1 + \lambda_2/\mu_2$.

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 definition.

Definition 1 *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 (F1) in Section 3. 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 rates and service capacity, we impose the following assumption on the abandonment and class-transition rates.

Assumption 1 *(i) $\theta_1 + \gamma_1\theta_2 > 0$ and $\theta_2 + \gamma_2\theta_1 > 0$. (ii) $\frac{1}{\mu_2} \neq \frac{1}{\mu_1}\frac{\gamma_2}{\theta_2+\gamma_2}$ and $\frac{1}{\mu_1} \neq \frac{1}{\mu_2}\frac{\gamma_1}{\theta_1+\gamma_1}$.*

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

Part (i) of Assumption 1 requires that the system has the "necessary" abandonment for stability even with no service. For example, if the abandonment rate from Class 1 is zero ($\theta_1 = 0$), then a Class 1 customer can leave the system by converting to Class 2 and eventually abandoning the Class 2 queue ($\gamma_1\theta_2 > 0$). Part (ii) of the assumption requires that there is a workload difference based on when (i.e. before versus after class-transitions occur) service is provided. This imposes a tradeoff when deciding which class to prioritize (see, Appendix A.1 for more details).

The long-run average cost minimization problem can be explicitly written as

$$\begin{aligned}
\min_{z_1,z_2,q_1,q_2} \quad & \limsup_{T\to\infty} \frac{1}{T}\int_0^T (c_1 q_1(t) + c_2 q_2(t))\, dt \\
s.t. \quad & dq_1(t) = \lambda_1 - \mu_1 z_1(t) - \theta_1 q_1(t) - \gamma_1 q_1(t) + \gamma_2 q_2(t) \\
& dq_2(t) = \lambda_2 - \mu_2 z_2(t) - \theta_2 q_2(t) - \gamma_2 q_2(t) + \gamma_1 q_1(t) \\
& z_1(t) + z_2(t) \le s, \quad t \ge 0 \\
& z_1(t), z_2(t), q_1(t), q_2(t) \ge 0, \quad t \ge 0.
\end{aligned}$$

This is an infinite dimensional linear program (LP). We first make an important observation that allows us to reformulate the problem as a finite dimensional LP. This observation 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.

$$\begin{aligned}
\min_{z_1^e,z_2^e,q_1^e,q_2^e} \quad & c_1 q_1^e + c_2 q_2^e \\
s.t. \quad & \lambda_1 - \mu_1 z_1^e - \theta_1 q_1^e - \gamma_1 q_1^e + \gamma_2 q_2^e = 0 \\
& \lambda_2 - \mu_2 z_2^e - \theta_2 q_2^e - \gamma_2 q_2^e + \gamma_1 q_1^e = 0 \\
& z_1^e + z_2^2 \le s \\
& z_1^e, z_2^e, q_1^e, q_2^e \ge 0.
\end{aligned} \tag{4}$$

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

$$\begin{aligned}
\max_{z_1^e,z_2^e} \quad & \left(\frac{c_1}{\theta_1 + \gamma_1 \frac{\theta_2}{\gamma_2+\theta_2}} + \frac{c_2 \frac{\gamma_1}{\gamma_1+\theta_1}}{\theta_2 + \gamma_2 \frac{\theta_1}{\gamma_1+\theta_1}} \right) \mu_1 z_1^e + \left(\frac{c_2}{\theta_2 + \gamma_2 \frac{\theta_1}{\gamma_1+\theta_1}} + \frac{c_1 \frac{\gamma_2}{\theta_2+\gamma_2}}{\theta_1 + \gamma_1 \frac{\theta_2}{\theta_2+\gamma_2}} \right) \mu_2 z_2^e \\
s.t. \quad & z_1^e + z_2^e \le s \\
& \frac{(\theta_2+\gamma_2)\lambda_1 + \gamma_2\lambda_2}{(\theta_2+\gamma_2)\theta_1 + \gamma_1\theta_2} - \frac{(\theta_2+\gamma_2)\mu_1}{(\theta_2+\gamma_2)\theta_1 + \gamma_1\theta_2} z_1^e - \frac{\gamma_2\mu_2}{(\theta_2+\gamma_2)\theta_1 + \gamma_1\theta_2} z_2^e \ge 0 \\
& \frac{(\theta_1+\gamma_1)\lambda_2 + \gamma_1\lambda_1}{(\theta_1+\gamma_1)\theta_2 + \gamma_2\theta_1} - \frac{(\theta_1+\gamma_1)\mu_2}{(\theta_1+\gamma_1)\theta_2 + \gamma_2\theta_1} z_2^e - \frac{\gamma_1\mu_1}{(\theta_1+\gamma_1)\theta_2 + \gamma_2\theta_1} z_1^e \ge 0 \\
& z_1^e, z_2^e \ge 0.
\end{aligned} \tag{5}$$

It is easy to see that the optimal solution to (5) tends to assign a larger value to the z_i^e with a larger coefficient 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

$$r_1 := \left(\frac{c_1}{\theta_1 + \gamma_1 \frac{\theta_2}{\gamma_2+\theta_2}} + \frac{c_2 \frac{\gamma_1}{\gamma_1+\theta_1}}{\theta_2 + \gamma_2 \frac{\theta_1}{\gamma_1+\theta_1}} \right) \mu_1, \tag{6}$$

and **the modified $c\mu/\theta$ index for Class 2** is

$$r_2 := \left(\frac{c_2}{\theta_2 + \gamma_2 \frac{\theta_1}{\gamma_1+\theta_1}} + \frac{c_1 \frac{\gamma_2}{\theta_2+\gamma_2}}{\theta_1 + \gamma_1 \frac{\theta_2}{\theta_2+\gamma_2}} \right) \mu_2. \tag{7}$$

From (6) and (7), we observe that when $\gamma_i = 0$ we recover the standard $c\mu/\theta$-index (Atar et al. 2010). When $\gamma_i \neq 0$, the extra terms are to account for the class-transition dynamics. To interpret the index r_1 in (6) (r_2 in (7) follows by symmetry), we note that the first term corresponds to the standard $c\mu/\theta$-structure for Class 1 customers. In particular, these customers incur a cost at rate c_1. The effective abandonment rate is $\theta_1 + \gamma_1\theta_2/(\theta_2 + \gamma_2)$. This is because "abandonment" in this case consists of the nominal abandonment, which happens at rate θ_1, as well as the improvement. The improvement happens at rate γ_1, but we also have to adjust for the fact that $\gamma_2/(\gamma_2 + \theta_2)$ of those customers may deteriorate and transition back to Class 1. Thus, the net improvement rate is $\gamma_1\theta_2/(\gamma_2 + \theta_2)$. The second term takes into account the Class 1 customers who improve to Class 2. These customers will incur a cost at rate c_2 when in Class 2. Because the proportion of Class 1 customers who improve to Class 2 is $\gamma_1/(\theta_1 + \gamma_1)$, the expected cost rate is $c_2\gamma_1/(\theta_1 + \gamma_1)$. When in Class 2, these customers abandon at rate θ_2, and deteriorate at rate γ_2 with a feedback probability $\gamma_1/(\theta_1 + \gamma_1)$.

Formally, we have the following theorem characterizing the optimal scheduling policy based on the modified $c\mu/\theta$-index.

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 $r_1 \geq r_2$, for r_1, r_2 defined in* (6) *and* (7), *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 ensure that the fluid dynamical system converges to the desired equilibrium point under the strict priority rule implied by the modified $c\mu/\theta$-index. We provide detailed analysis on the long-run regularity of the fluid model under the strict priority rules in Appendix A. These convergence analyses are interesting in their own right, as they reveal important characteristics of the system dynamics. Moreover, we show that an interesting *bi-stability* phenomenon, i.e. the presence of two equilibria, can arise when the sub-optimal strict priority rule is employed. We provide more discussions about this phenomenon in Section 3.1.

We next numerically compare the long-run average costs of the fluid models to those of the corresponding stochastic systems under different strict priority rules. We denote P_1 as strict priority to Class 1 and P_2 as strict priority to Class 2[2]. Figure 2 plots the long-run average costs for systems with different numbers of servers s. The fluid costs are plotted in dashed lines while the costs for the stochastic systems 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. Figure 2(a) illustrates the scenario where the modified $c\mu/\theta$-index suggests prioritizing Class 1, while Figure 2(b) has the modified $c\mu/\theta$-index suggesting prioritizing Class 2. We first note that the long-run average fluid cost approximates the long-run average cost of the stochastic system reasonably well, especially when s is small (the system is in the so-called overloaded regime) and when s is large (the system is in the so-called underloaded regime). Second, we observe that when comparing the strict priority rules, prioritizing the class with a larger modified $c\mu/\theta$-index always leads to a lower cost in the stochastic system. Thus, even when the cost of fluid system may deviate from that of the corresponding stochastic system, the resulting policy recommendations are consistent. Lastly, we note that in Figure 2(b), when $18 \leq s \leq 22$, the fluid model under strict priority to Class 1 has two different equilibria (bi-stability). Which equilibrium the fluid system converges to depends on its initial condition. For the corresponding stochastic system, it will fluctuate around one equilibrium point for a while before transitioning to the region around the other equilibrium point. Thus, the corresponding long-run average cost is a weighted average of the costs around the two equilibria.

3.1. Bi-Stability

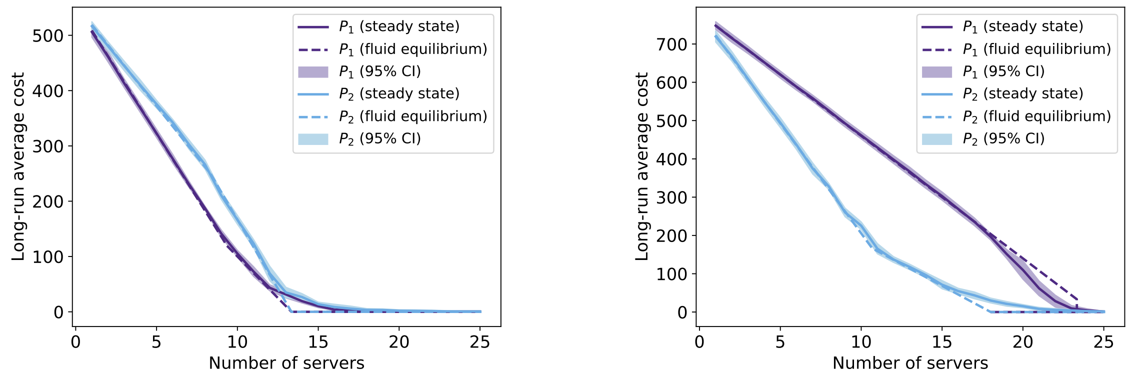
Due to the dynamic class types, applying the strict priority rule that does not agree with the modified $c\mu/\theta$-index can lead to a bi-stability phenomenon. Motivated by proactive service applications, in this section, we study in more depth a special case where bi-stability arises. Specifically, we consider the system parameters for which Theorem 4 in Appendix A.1 suggests that if we prioritize the urgent class, the system exhibits *bi-stability*. While, in this parameter regime, following the modified $c\mu/\theta$-rule is the optimal policy, from a practical standpoint, the service provider may prefer to give priority to the urgent class, as long as it does not degrade system performance. We

[2] Throughout this manuscript, all numerical experiments for the stochastic system are conducted with preemption, though we emphasize this has no impact on the fluid analysis.

Figure 2 **Optimal long-run scheduling policy**

((a): $\lambda_1 = 10, \lambda_2 = 20, \mu_1 = 1.5, \mu_2 = 3, \gamma_1 = 0.1, \gamma_2 = 0.1, \theta_1 = 0.1, \theta_2 = 0.4, c_1 = 5, c_2 = 3$

(b): $\lambda_1 = 10, \lambda_2 = 20, \mu_1 = 1, \mu_2 = 2.5, \gamma_1 = 0.2, \gamma_2 = 0.4, \theta_1 = 0.1, \theta_2 = 0.2, c_1 = 5, c_2 = 1$**)**



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

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

explore whether it is reasonable to (sometimes) give priority to the urgent class even though one of the optimal long-run average policies indicates priority should be given to the moderate class.

The parameter regime we are interested in is when the urgent class (Class 1) has a higher $c\mu$-index, i.e., $c_1\mu_1 > c_2\mu_2$, but $\mu_1 < \frac{\gamma_2}{\theta_2+\gamma_2}\mu_2$, which implies that the moderate class (Class 2) has a higher modified $c\mu/\theta$-index, i.e., $r_2 \geq r_1$. In this case, from the workload perspective, it is more efficient to serve moderate customers before they deteriorate, i.e.,

$$\frac{1}{\mu_2} < \frac{\gamma_2}{\gamma_2 + \theta_2}\frac{1}{\mu_1}.$$

Additionally, the capacity is in the critical region

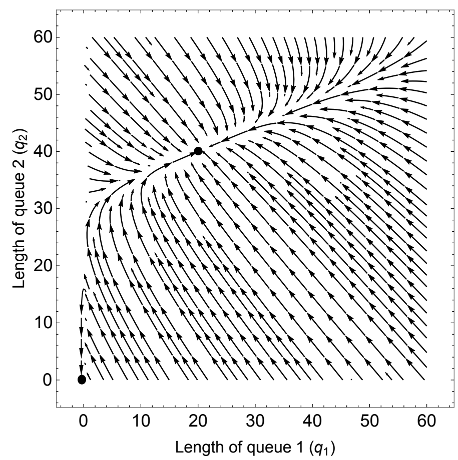
$$\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} < s \leq \frac{\lambda_1}{\mu_1} + \frac{\gamma_2}{\theta_2 + \gamma_2}\frac{\lambda_2}{\mu_1}. \tag{8}$$

Figure 3 provides an illustration of the vector field under bi-stability. We note that there are two locally asymptotically stable equilibrium points. Which equilibrium point the queue process converges to depends on its initialization.

Intuitively, the bi-stability arises because if we delay service for moderate customers, they will end up generating more workload on average when they deteriorate into the urgent class. When the system is critically loaded as in (8), even though we have enough capacity to serve both classes when service 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 customers when service for Class 2 is delayed, i.e.,

$$s \leq \frac{\lambda_1}{\mu_1} + \frac{\gamma_2}{\theta_2 + \gamma_2}\frac{\lambda_2}{\mu_1}.$$

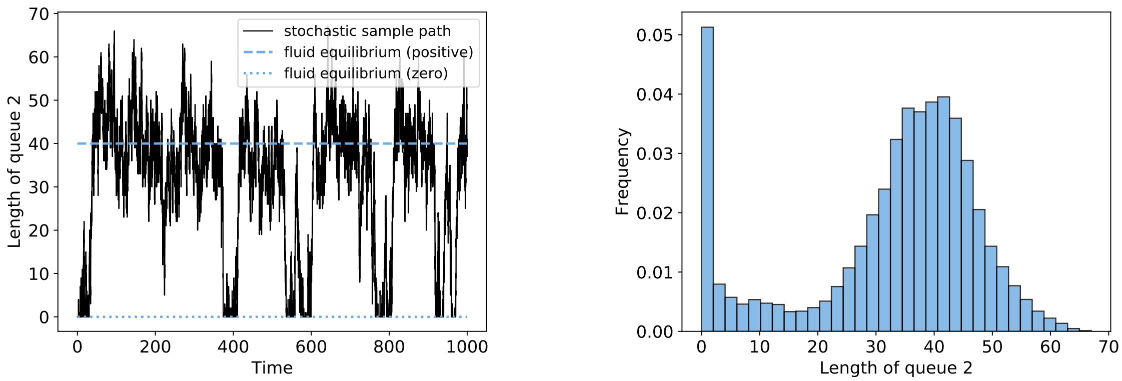
Figure 3 **Vector field under bi-stability**
($\lambda_1 = 10, \lambda_2 = 20, s = 20, \mu_1 = 1, \mu_2 = 2.5, \gamma_1 = 0.2, \gamma_2 = 0.4, \theta_1 = 0.1, \theta_2 = 0.2$**)**



Under bi-stability, we note that one of the equilibrium points leads to very good performance – zero holding cost, while the other equilibrium point has positive queues for at least one class (Figure 3 and Theorem 4). Ideally, we want to avoid the "bad" equilibrium regardless of where we start. One way to ensure global convergence to the "good" equilibrium is to switch priority to the moderate class as suggested by the modified $c\mu/\theta$-rule. However, there are many systems where it may be preferable to give priority to the urgent class for obvious administrative reasons. Thus, we propose an alternative intervention, which we refer to as the bi-stability control. For a fixed threshold $\alpha_0 > 0$, when $\bar{q}_1(t) + \bar{q}_2(t) \leq \alpha_0$, we prioritize the urgent class; otherwise, we prioritize the moderate class. Following a similar Lyapunov argument as in Appendix A.1, when α_0 is sufficiently small, $q(t)$ will converge to $(0,0)$ regardless of its initialization q_0, i.e., $(0,0)$ is a globally asymptotically stable equilibrium under this control. As such, both the modified $c\mu/\theta$ rule and the bi-stability control with properly chosen threshold lead to the same optimal long-run average cost in this case. However, when studying the transient cost, i.e., the cost incurred to restore system to zero when it is initialized far from zero, the bi-stability control can lead to a lower cost than the modified $c\mu/\theta$ rule as we will explain in Section 4.

We next elaborate on the implications of the fluid bi-stability phenomenon for the stochastic system. When bi-stability arises in the fluid system, the queue length process of the corresponding stochastic system will fluctuate around one equilibrium for a while before transitioning to the region around the other equilibrium. Figure 4(a) shows a typical sample path of the stochastic queue length process, i.e., we plot $Q_2(t)$ for $t \in [0, 1000]$. Figure 4(b) provides the histogram of $Q_2(t)$. We observe that it follows a bi-modal distribution where the two peaks are around the two fluid equilibria.

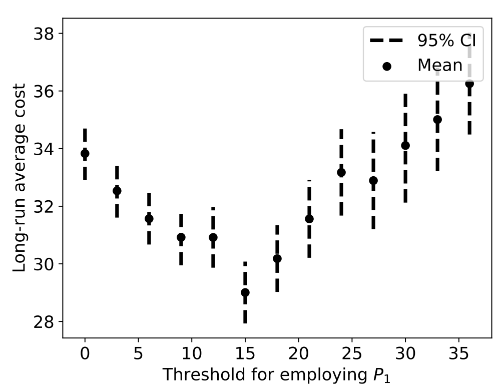
Figure 4 **Bi-stability in the stochastic system**
($\lambda_1 = 10, \lambda_2 = 20, \mu_1 = 1, \mu_2 = 2.5, \gamma_1 = 0.2, \gamma_2 = 0.4, \theta_1 = 0.1, \theta_2 = 0.2, s = 20$**)**



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

Figure 5 plots the long-run average cost for the stochastic system under the bi-stability control for different values of α_0 (point estimates together with the corresponding 95% confidence intervals). In the stochastic system, if $Q_1(t)+Q_2(t) \leq \alpha_0$, we prioritize the urgent class. Otherwise, we prioritize the moderate class. Note that $\alpha_0 = 0$ is equivalent to assigning strict priority to the moderate class, i.e., the modified $c\mu/\theta$-rule. Interestingly, we observe that for certain values of α_0, the bi-stability control achieves a smaller long-run average cost than the modified $c\mu/\theta$-rule. As surprising as the observation may seem at first glance, this phenomenon is due to stochastic fluctuations that bring the system away from the equilibrium, i.e., zero queue. To restore the system to zero in the most cost-effective way, the experiments suggest that we should prioritize the moderate class when the queues are large, and prioritize the urgent class when the queues are small. We explore this more formally in our transient analysis in Section 4.

Figure 5 **Long-run average cost under the bi-stability control**
($\lambda_1 = 10, \lambda_2 = 20, \mu_1 = 1, \mu_2 = 2.5, \gamma_1 = 0.2, \gamma_2 = 0.2, \theta_1 = 0.1, \theta_2 = 0.2, s = 19, c_1 = 10, c_2 = 1$**)**



4. Optimal Transient Scheduling Policy

Service systems often operate in highly non-stationary environments. In healthcare settings, for example, random shocks like disease outbreaks or mass casualty events can push the system far from its normal state, i.e., equilibrium. When such a demand shock 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 find the optimal clearing of backlogs. In particular, we derive the optimal scheduling policy to help the system recover from demand shocks.

We start by focusing on the after-shock control. In particular, we assume $q(0) = q_0 > 0$ [3] but we now 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 $s > \lambda_1/\mu_1 + \lambda_2/\mu_2$.

Later in Section 5.1, we generalize the arrival-rate pattern to include the period of demand shock in our planning horizon. In particular, the shock raises the arrival rates for a fixed amount of time, during which the service capacity is insufficient and so the backlogs increase. Importantly, the arrival rates during the demand shock period can be time-varying and violate Assumption 2. After the initial shock, the arrival rates restore to normal and satisfy Assumption 2. It is significant that the structure of the optimal control does not change under this more general arrival-rate model (see Theorem 3). The after-shock control studied in this section builds the basis for the cases with more general arrival rates.

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 Theorem 4, Assumption 2 implies that there exists a scheduling policy π, under which, for any $q(0) = q_0 > 0$, $\tau < \infty$. We also note from our long-run regularity analysis in Appendix A.1 that under Assumption 2, both strict priority to Class 1 and strict priority to Class 2 lead to the same long-run average holding cost – zero. However, our following analysis will reveal important differences in their transient performance.

The optimal transient scheduling policy depends on the interplay between two index rules. We define the *$c\mu$-rule* as a policy that prioritizes the class with a higher $c_i\mu_i$ value, $i = 1, 2$, i.e., the $c\mu$-index. Similarly, the *modified $c\mu/\theta$-rule* is a policy that prioritizes the class with a higher r_i value, $i = 1, 2$, i.e., the modified $c\mu/\theta$-index as defined in (6) and (7). To capture the differential effect of each of these rules, we impose the following assumption on the indices.

[3] We define a vector $a > 0$ if all its components are nonnegative and there is at least one component that is strictly positive.

Assumption 3 *$c_1\mu_1 \neq c_2\mu_2$ and $r_1 \neq r_2$.*

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 $f_1(q,z) = \lambda_1 - z_1\mu_1 - \theta_1 q_1 - \gamma_1 q_1 + \gamma_2 q_2$ and $f_2(q,z) = \lambda_2 - z_2\mu_2 - \theta_2 q_2 - \gamma_2 q_2 + \gamma_1 q_1$. 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 $h_1(z) = z_1 + z_2 - s$, $h_2(z) = -z_1$, and $h_3(z) = -z_2$. 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 optimal control theory, optimization problems of the form (F2′) are referred to as *optimal control with state constraints*. Despite a rich body of literature in optimal control, problems with state constraints are, in general, very difficult to solve explicitly as they impose extra boundary conditions (Trélat 2012). While some results can be derived in special cases, there is no systematic way to deal with these problems; we refer to the survey paper Hartl et al. (1995) for an overview.

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. In particular, as we shall explain in Theorem 2, the optimal scheduling policy switches priority at most once and priorities can be characterized by two simple index rules. Then solving for the optimal scheduling policy reduces to finding the policy curve that governs where in the state space the switch in priority happens. We provide a closed form characterization of the policy curve in Proposition 4 for a special case, and provide an efficient numerical scheme to construct the policy curve for the other cases.

The next theorem characterizes the structure of the transient optimal scheduling policy. Let τ^* denote the time to empty the queue under the optimal policy.

Theorem 2 *Under Assumptions 1, 2, and 3, for the transient optimal control problem* (F2′)*:*

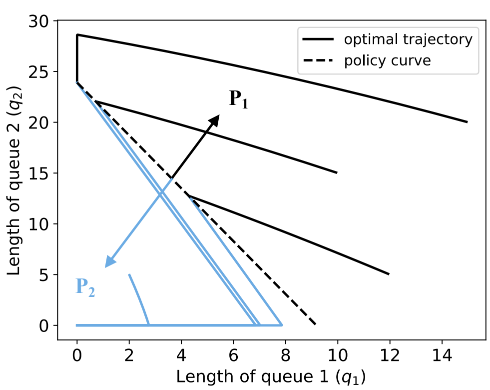
I. *If the $c\mu$-rule and the modified $c\mu/\theta$-rule both prioritize Class i, $i = 1,2$, then the strict priority rule to Class i is optimal for any $t \in [0, \tau^*]$.*

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 with $0 < \epsilon < M$, 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 $[0,\tau^*]$.*

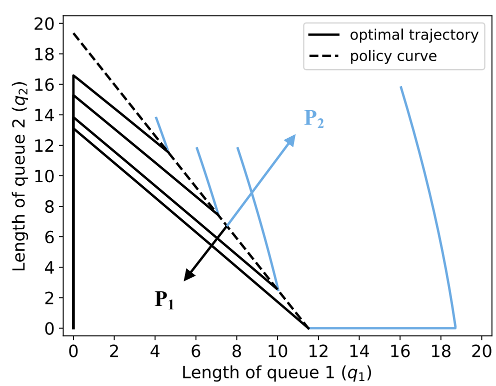
Based on Theorem 2, if the $c\mu$-rule and the modified $c\mu/\theta$-rule agree with each other, it is optimal to give strict priority to the class with a higher $c\mu$-index (and correspondingly a higher modified $c\mu/\theta$-index) for any $q\in\mathbb{R}_+^2$. If the two index rules do not agree, 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, and the time at which 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 a numerical illustration 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 values (derived by solving the a discretized version of (F2′)). We also plot the corresponding policy curve (dashed line). Figure 6(b) shows the case where the modified $c\mu/\theta$-rule prioritizes Class 2 while the $c\mu$-rule prioritizes Class 1. We will provide more discussions about the policy curve in Section 4.3.3.

Figure 6 **Optimal transient queue length trajectory**
((a): $\lambda_1=10,\lambda_2=20,\mu_1=1.5,\mu_2=3,\gamma_1=0.1,\gamma_2=0.1,\theta_1=0.1,\theta_2=0.4,s=17,c_1=5,c_2=3$
(b): $\lambda_1=10,\lambda_2=20,\mu_1=1,\mu_2=2.5,\gamma_1=0.2,\gamma_2=0.4,\theta_1=0.1,\theta_2=0.2,s=26,c_1=5,c_2=1$**)**



(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 3 *Even though Theorem 2 is stated under Assumption 1, following the same lines of analysis, we can show that if $\theta_1=\theta_2=\gamma_1=\gamma_2=0$, we can recover the well-known optimality of* *the cμ-rule throughout the transient time horizon (see Corollary 2 in Appendix D). Furthermore, if $\gamma_1 = \gamma_2 = 0$ but $\theta_1, \theta_2 > 0$, we should follow the cμ-rule when we are close to the origin and the ordinary $c\mu/\theta$-rule when we are far from the origin (This is a special case of Theorem 2). In this case, we recover the results in Larrañaga (2015). Nevertheless, the approaches utilized in literature to study the special cases are not directly generalizable to our setting with dynamic class types.*

We next provide the general strategy of proving Theorem 2. It includes three main parts. We first provide some formal definitions to describe the boundary behavior and rule out some "irregular" behaviors 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, utilizing Pontryagin's Minimum Principle. We believe this framework can be applied to derive the structure of the optimal policy for other transient control problems for queues.

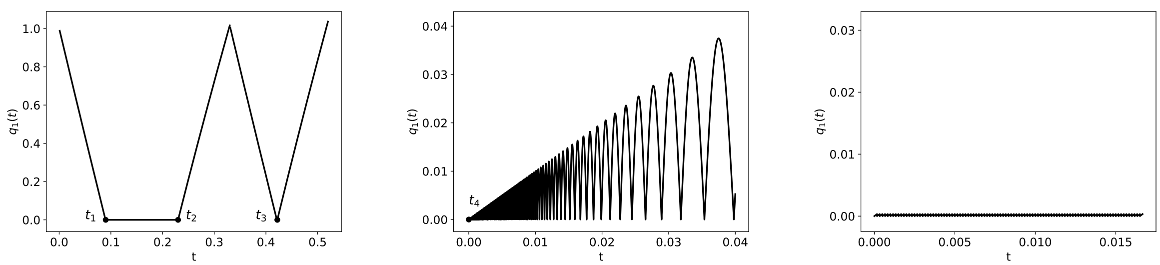
4.1. Boundary Behavior

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 $q_i(t)$, $i=1,2$, 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, 2, and 3, for the transient optimal control problem* (F2′)*, if $q_1(t),q_2(t)\in[0,\epsilon)$, with $\epsilon>0$ sufficiently small, then 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 dynamic for the system comes from service completion, which has an order ϵ effect. The effect of abandonment and class-transition is only second-order, namely, order ϵ^2. 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, $i = 1, 2$. 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 class-transition 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. $g(q(t)) \leq 0$, and the pure-control constraints i.e. $h(z(t)) \leq 0$. 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 is defined as:

$$\begin{aligned}
H(q(t), z(t), p(t)) &:= p(t)^T f\left(q(t), z(t)\right) + F\left(q(t)\right) \\
&= p_1(t)dq_1(t) + p_2(t)dq_2(t) + c_1q_1(t) + c_2q_2(t) \\
&= p_1(t)\left(\lambda_1 - \mu_1z_1(t) - \theta_1q_1(t) - \gamma_1q_1(t) + \gamma_2q_2(t)\right) \\
&\quad + p_2(t)\left(\lambda_2 - \mu_2z_2(t) - \theta_2q_2(t) - \gamma_2q_2(t) + \gamma_1q_1(t)\right) + c_1q_1(t) + c_2q_2(t).
\end{aligned}$$

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

$$\begin{aligned}
L(q(t), z(t), p(t), \eta(t), \xi(t)) &:= H(q(t), z(t), p(t)) + \eta(t)^T g(q(t)) + \xi(t)^T h(z(t)) \\
&= p_1(t)\left(\lambda_1 - \mu_1z_1(t) - \theta_1q_1(t) - \gamma_1q_1(t) + \gamma_2q_2(t)\right) \\
&\quad + p_2(t)\left(\lambda_2 - \mu_2z_2(t) - \theta_2q_2(t) - \gamma_2q_2(t) + \gamma_1q_1(t)\right) + c_1q_1(t) + c_2q_2(t) \\
&\quad - \eta_1(t)q_1(t) - \eta_2(t)q_2(t) + \xi_1(t)(z_1(t) + z_2(t) - s) - \xi_2(t)z_1(t) - \xi_3(t)z_2(t).
\end{aligned}$$

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" $q^*(t)$:

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

2) *Adjoint Vector condition* (ADJ) specifies the dynamics of the "optimal dual trajectory" $p^*(t)$:

$$dp_1^*(t) = (\theta_1 + \gamma_1)p_1^*(t) - \gamma_1 p_2^*(t) - c_1 + \eta_1^*(t), \quad dp_2^*(t) = (\theta_2 + \gamma_2)p_2^*(t) - \gamma_2 p_1^*(t) - c_2 + \eta_2^*(t). \tag{ADJ}$$

In general, we cannot fully characterize $p^*(t)$ due to 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:

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

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. 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

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

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

$$z_1^*(t) = \begin{cases} s & \text{if } q_1^*(t) > 0 \\ \min\left\{s, \frac{\lambda_1 + \gamma_2 q_2^*(t)}{\mu_1}\right\} & \text{if } q_1^*(t) = 0, \end{cases} \quad \text{and} \quad z_2^*(t) = s - z_1^*(t). \tag{9}$$

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

$$z_1^*(t) = s - z_2^*(t), \quad \text{and} \quad z_2^*(t) = \begin{cases} s & \text{if } q_2^*(t) > 0 \\ \min\left\{s, \frac{\lambda_2 + \gamma_1 q_1^*(t)}{\mu_2}\right\} & \text{if } q_2^*(t) = 0. \end{cases} \tag{10}$$

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, $i = 1, 2$. 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: There exits a vector $\omega^*(\beta)=(\omega_1^*(\beta),\omega_2^*(\beta))\in\mathbb{R}^2$, such that

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

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 characterizes 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 establishes 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 $q_2^*(t)$ before τ^*. 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 τ_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*

$$\psi(\tau_j-t)=r_1-r_2+(\mu_1A_1(\tau_j)-\mu_2A_2(\tau_j))\,e^{\upsilon_1(\tau_j-t)}-(\mu_1A_1(\tau_j)-\mu_2A_2(\tau_j))\,e^{\upsilon_2(\tau_j-t)},$$

where r_1,r_2 are defined in (6) and (7), υ_1,υ_2 are positive constants that depend on γ_i's and θ_i's, and $A_1(\tau_j),A_2(\tau_j)$ 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

$$\begin{aligned}D^{\tau_j}(t):=\;&r_1-r_2+(\mu_1A_1(\tau_j)-\mu_2A_2(\tau_j))\,e^{\upsilon_1(\tau_j-t)}\\&-(\mu_1A_1(\tau_j)-\mu_2A_2(\tau_j))\,e^{\upsilon_2(\tau_j-t)},\quad\text{for } t\geq 0,\ j=1,...,N.\end{aligned}$$

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 β, i.e., $\eta_i^*(\tau_j-\beta)>0$ for some $i=1,2$, the switching curve may deviate from the pseudo switching curve for $t\geq\beta$.

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

$$\lim_{t\to\infty}D^{\tau_1}(t)=r_1-r_2,\quad\text{for } r_1,r_2 \text{ in (6) and (7).}\tag{11}$$

The sign of the right-hand-side of (11) 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 (11) for the backward switching curve $\psi(\tau_1 - t)$. Because $\psi(\tau_1 - t)$ only equals to $D^{\tau_1}(t)$ 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 such that the sign of $\psi(\tau_1 - t)$ will be governed by the modified $c\mu/\theta$-index. However, we need to note that the constants $A_1(\tau_1)$ and $A_2(\tau_1)$ also 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 analyses together, we are able to establish the following result.

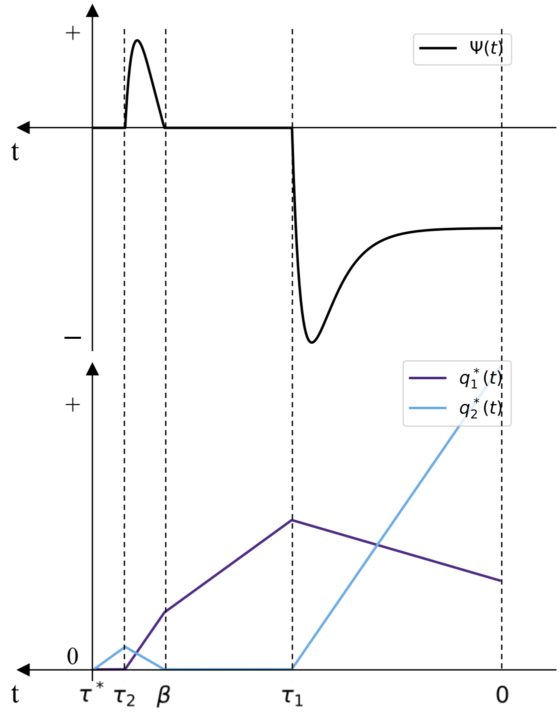
Proposition 2 *Under Assumptions 1, 2, and 3, for the transient optimal control problem (F2′), there exists a positive real number M such that when $q_1(t) + q_2(t) > M$, the modified $c\mu/\theta$-rule is optimal at time t, $t \geq 0$.*

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, 2, and 3, 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 rules 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 optimal 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 (following 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 (following the $c\mu$-rule). Note that the Class 1 queue decreases and the Class 2 queue increases over $[\beta, \tau_2)$. Lastly, priority is kept at Class 1 on $[\tau_2, \tau^*]$ to maintain its queue at zero.

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



4.3.3. The Policy Curve In this section, we first focus on the special case where there is only one-way transition from Class 2 to Class 1, namely, $\gamma_1 = 0$. From Theorem 2, 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 that 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 the Hamiltonian condition (H), we are able to characterize (and approximate) the policy curve $\mathcal{P}$ 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, 2, and 3, for the transient optimal control problem* (F2′) *with $\gamma_1 = 0$, if $c_1\mu_1 < c_2\mu_2$ and $r_1 > r_2$, the policy curve $\mathcal{P}$ for switching from P_1 to P_2 is given by*

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

where

$$B_1(\mathbf{a_2}) := (c_1(-\mathbf{a_2}(\theta_2+\gamma_2)+\lambda_2)\mu_1 + c_2\mathbf{a_2}\theta_1\mu_2 + c_1(\mathbf{a_2}\gamma_2+\lambda_1-s\mu_1)\mu_2)$$

$$\begin{aligned}B_2(\mathbf{a_1},\mathbf{a_2}) := \Bigg(& -\mu_2(\mathbf{a_2}\gamma_2\theta_1 + \mathbf{a_1}\theta_1(\gamma_2-\theta_1+\theta_2) - \gamma_2\lambda_1 + \theta_1\lambda_1 - \theta_2\lambda_1 - \gamma_2\lambda_2 + s\gamma_2\mu_2) \\ & + (\lambda_2 - s\mu_2)((\gamma_2-\theta_1+\theta_2)\mu_1 - \gamma_2\mu_2)\left(1+\frac{\mathbf{a_2}(\theta_2+\gamma_2)}{-\lambda_2+s\mu_2}\right)^{\frac{\theta_1}{\theta_2+\gamma_2}}\Bigg)\end{aligned}$$

$$B_3(\mathbf{a_1},\mathbf{a_2}) := (\gamma_2-\theta_1+\theta_2)\,(\theta_1(\mathbf{a_2}(\theta_2+\gamma_2)-\lambda_2)\mu_1 + \theta_1(-\mathbf{a_2}\gamma_2+\mathbf{a_1}\theta_1-\lambda_1+s\mu_1)\mu_2)\,.$$

If $\gamma_2 = 0$, $c_1\mu_1 > c_2\mu_2$ and $r_1 < r_2$, we can derive the policy curve $\mathcal{P}$ for switching from P_2 to P_1 by symmetry from Proposition 4.

If $c_1\mu_1 > c_2\mu_2$ and $r_1 < r_2$ (still with $\gamma_1 = 0$), the policy curve for switching from P_2 to P_1 cannot be characterized in closed form. This is due to the class-transition 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 deterioration, $\eta_1^*(t)$ 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 $q_2^*(t) = 0$, $\eta_2^*(t)$ affects $p_2^*(t)$ only. See Appendix B.10 for a more detailed discussion.

We note that the policy curve characterized in Proposition 4 is close to being, but not exactly, linear. More generally, to characterize the policy curve for switching from the modified $c\mu/\theta$-rule to the $c\mu$-rule in the presence of transitions from both class types, i.e., $\gamma_1, \gamma_2 > 0$, 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′). Record the n corresponding switching points.

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

We conduct extensive numerical experiments on $\mathcal{P}$ for different system parameters. In all cases, 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.

We next provide some sensitivity analysis on the policy curve through numerical experiments. In particular, we use the numerical scheme outlined above to construct the policy curve. For simplicity of illustration, we focus on 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, i.e., $c_1\mu_2 > c_2\mu_2$ and $r_1 < r_2$.

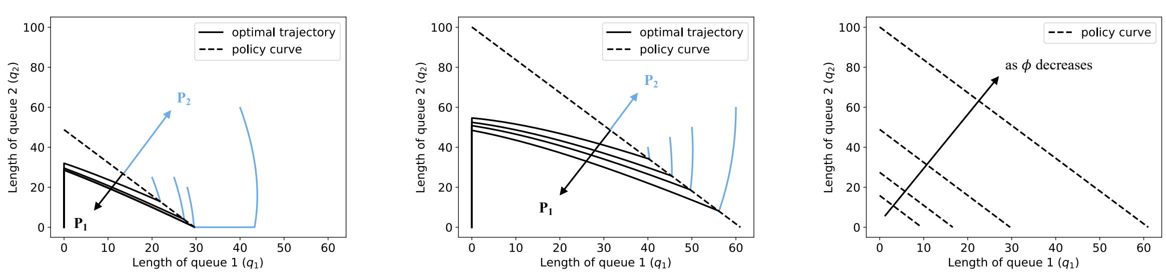
Our first group of numerical experiments is on the value of $\phi := \gamma_2/(\theta_2+\gamma_2)$. As mentioned earlier, if Class 2 patients are identified by a classifier, e.g., an early warning system, ϕ measures the true positive rate of the classifier. Figures 9 and 10 illustrate how the policy curve changes as ϕ decreases. Since γ_2 and θ_2 both affect the value of ϕ, we first keep θ_2 fixed and vary the value of γ_2 (Figure 9). Then, we keep γ_2 fixed and vary the values of θ_2 (Figure 10). In both figures, we vary the values of ϕ from 0.7 to 0.4 in increments of size -0.1. We first observe that the policy curve contracts inwards as ϕ increases. In the example of a classifier, this observation suggests that as the quality of the classifier improves, the region in which the optimal scheduling policy prioritizes Class 2 increases. When $\phi = 1$, the size of the region where it is optimal to prioritize Class 1 is minimized, but the region is still non-trivial. On the other hand, as ϕ decreases, a phase transition in the prioritization rule occurs because r_1 will become larger than r_2. In particular, there exists a threshold ϕ_0 such that once $\phi < \phi_0$, $r_1 > r_2$ and the policy curve "vanishes", namely, we should give strict priority to Class 1 for all states. We also note that given the complex nature of system dynamics, the effect of increasing θ_2 and decreasing γ_2 would be different. In particular, when comparing Figure 9 to Figure 10, we observe that even for the same value of ϕ, the policy curves in the two figures are different. To look further into this, in Figure 11, we fix $\phi = 0.6$ and vary θ_2 and γ_2 simultaneously. We observe that as θ_2 and γ_2 decrease, the policy curve contracts outwards. This is because as abandonment and class-transition occur at slower rates, the effect of service completion is more dominant. Thus, the region in which we adopt the $c\mu$-rule increases.

Similar to the above sensitivity analysis on the policy curve with respect to ϕ via θ_2 or γ_2, we also conduct sensitivity analysis for different values of capacity s; see Figure 12. We observe that the policy curve contracts inwards as s decreases. Define the nominal traffic intensity as

$$\rho := (\lambda_1/\mu_1 + \lambda_2/\mu_2)/s.$$

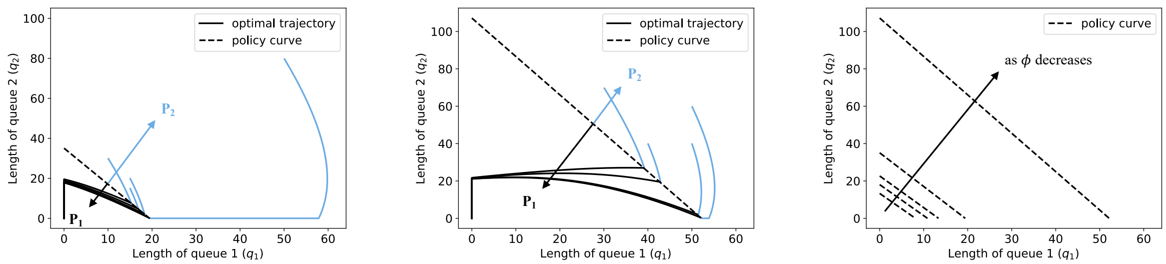
Figure 12 indicates that as the system becomes more heavily loaded (i.e., as ρ increases), the region where we prioritize according to the $c\mu$-rule shrinks.

Figure 9 **Sensitivity analysis of the policy curve with respect to $\gamma_2/(\theta_2+\gamma_2)$ by varying γ_2 ($\lambda_1 = 10, \lambda_2 = 20, \mu_1 = 1, \mu_2 = 2.5, \gamma_1 = 0.2, \theta_1 = 0.1, \theta_2 = 0.2, s = 26, c_1 = 5, c_2 = 1$)**



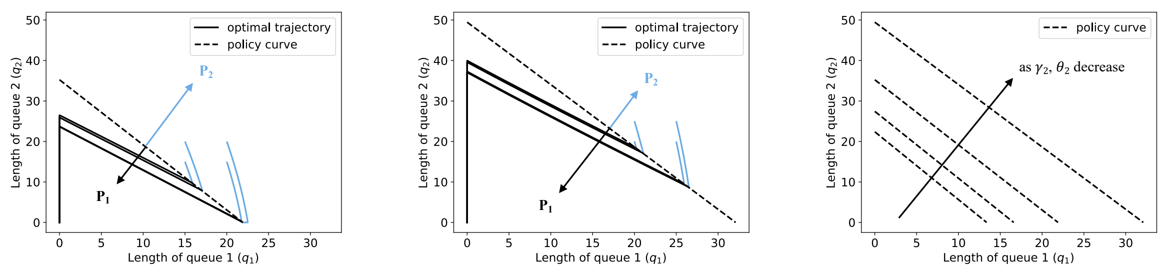
(a) $\phi = 0.5, \gamma_2 = 0.2$ (b) $\phi = 0.4, \gamma_2 = 0.13$ (c) Decreasing $\phi = 0.7, 0.6, 0.5, 0.4$

Figure 10 **Sensitivity analysis of the policy curve with respect to $\gamma_2/(\theta_2+\gamma_2)$ by varying θ_2**
($\lambda_1 = 10, \lambda_2 = 20, \mu_1 = 1, \mu_2 = 2.5, \gamma_1 = 0.2, \gamma_2 = 0.4, \theta_1 = 0.1, s = 26, c_1 = 5, c_2 = 1$)



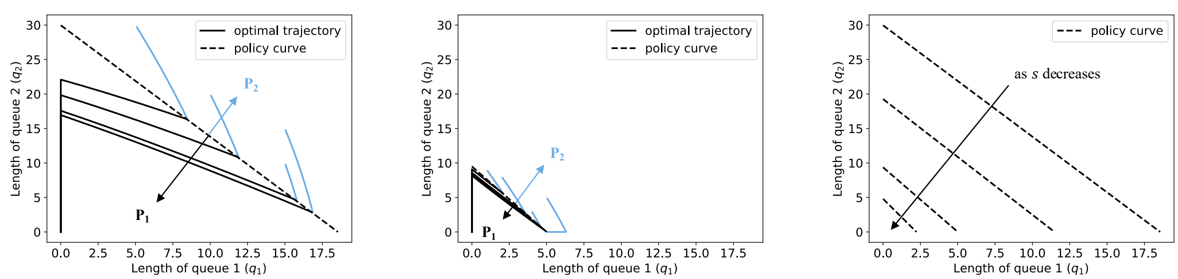
(a) $\phi = 0.5, \theta_2 = 0.4$ (b) $\phi = 0.4, \theta_2 = 0.6$ (c) Decreasing $\phi = 1,\ 0.7,\ 0.6,\ 0.5,\ 0.4$

Figure 11 **Sensitivity analysis of the policy curve with respect to γ_2 and θ_2 for fixed $\gamma_2/(\theta_2+\gamma_2)$**
($\lambda_1 = 10, \lambda_2 = 20, \mu_1 = 1, \mu_2 = 2.5, \gamma_1 = 0.2, \theta_1 = 0.1, \phi = 0.6, s = 26, c_1 = 5, c_2 = 1$)



(a) $\gamma_2 = 0.2, \theta_2 = 0.13$ (b) $\gamma_2 = 0.1, \theta_2 = 0.07$ (c) Decreasing $\gamma_2 = 0.4,\ 0.3,\ 0.2,\ 0.1$

Figure 12 **Sensitivity analysis of the policy curve with respect to s**
($\lambda_1 = 10, \lambda_2 = 20, \mu_1 = 1, \mu_2 = 2.5, \gamma_1 = 0.2, \gamma_2 = 0.4, \theta_1 = 0.1, \theta_2 = 0.2, c_1 = 5, c_2 = 1$)



(a) $s = 30,\ \rho = 0.6$ (b) $s = 22,\ \rho = 0.8$ (c) Decreasing $s = 30,\ 26,\ 22,\ 20$; increasing $\rho = 0.6,\ 0.7,\ 0.8,\ 0.9$

4.4. Numerical Experiments for the Stochastic System

We conclude this section by generalizing the insights from the fluid model analysis to the original stochastic system. The quality of the generalization depends how closely the fluid model approximates the corresponding stochastic system.

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 first elaborate on the scaling under heavy traffic and then conduct numerical comparisons between the fluid trajectory and scaled stochastic sample paths.

Consider a sequence of stochastic systems indexed by n, $n \in \mathbb{N}$. For Class i in the nth system, the arrival and service rates satisfy $\lambda_i^n := \lambda_i n$, $\mu_i^n := \mu_i n$, $i = 1, 2$. 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 nth stochastic system. Given the initial fluid queue length q_0, the nth system has initial queue length $Q^n(0) := \lceil q_0 n \rceil$. Given the fluid policy curve $\mathcal{P}$, for the nth stochastic system, a switch in priority will happen at time t if $\bar{Q}^n(t) \in \mathcal{P}^n$, where

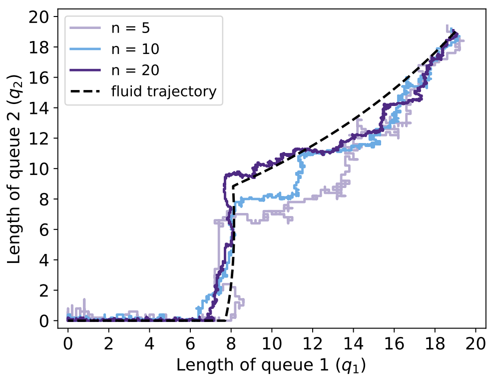
$$\mathcal{P}^n := \left\{ (\bar{Q}_1^n, \bar{Q}_2^n) : \bar{Q}_1^n \in [q_1 - 1/n, q_1 + 1/n], \bar{Q}_2^n \in [q_2 - 1/n, q_2 + 1/n], (q_1, q_2) \in \mathcal{P} \right\}.$$

Figure 13 compares the fluid trajectory with a simulated sample path for the corresponding 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.

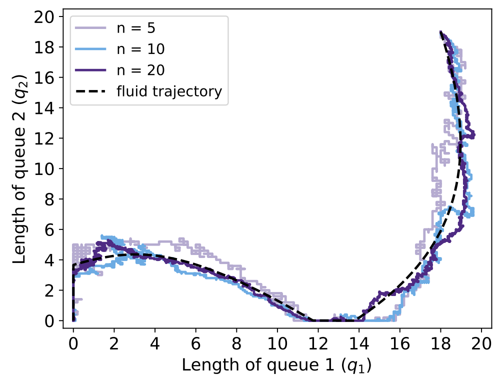
For systems with a very small number of servers, we can solve the MDP (S2) numerically; see Appendix E for details about our solution method. In Figure 14, we plot the MDP solutions together with the fluid policy curves for four 3-server systems with the nominal traffic intensity ρ varying from 0.6 to 0.9. We observe that for lightly and moderately loaded systems (with $\rho = 0.6, 0.7, 0.8$), the optimal scheduling policy for the stochastic system shares the same structure as the optimal fluid control, i.e., it switches priority once from the modified $c\mu/\theta$-rule to the $c\mu$-rule. We also plot the corresponding fluid policy curve (solid line in Figure 14). We observe that when the system is lightly loaded, i.e., $\rho = 0.6, 0.7$, the policy curve of the MDP solution matches the fluid policy curve well. For critically loaded system (with $\rho = 0.9$), the optimal policy follows the modified $c\mu/\theta$-rule throughout; namely, the neighborhood near the origin where the $c\mu$-rule is optimal does not exist. In addition, note that the region where the fluid policy employs the $c\mu$-rule is also very small in this case. Despite some slight deviations between the MDP solution and the fluid-translated policy, in all cases, the optimality gap of the fluid-translated policy is very small, as shall be seen next.

For each of the four stochastic systems in Figure 14, we randomly select a set $\mathcal{J}$ of initial points; see Appendix E for details on this initialization. For each initialization $q_0 \in \mathcal{J}$, we compare the average transient cost under (i) the MDP policy (the optimal policy), (ii) the fluid-translated policy with an approximating linear policy curve $\mathcal{P}$, (iii) strict priority to Class 1, P_1, and (iv) strict priority to Class 2, P_2. Each cost is estimated based on 1000 independent sample paths. In Table 1, we present the average, minimum and maximum optimality gap for polices (ii), (iii), and (iv). We observe that the fluid-translated policy has a very small optimality gap in all cases. In particular, the maximum optimality gaps are less than 3.8% and the mean optimality gaps are less than 1.6%.

Figure 13 **Comparison of the transient fluid trajectory and the stochastic sample path**
((a) 2 servers: $\lambda_1 = 1, \lambda_2 = 2, \mu_1 = 1.5, \mu_2 = 3, \gamma_1 = 0.1, \gamma_2 = 0.1, \theta_1 = 0.1, \theta_2 = 0.4$
(b) 3 servers: $\lambda_1 = 1, \lambda_2 = 2, \mu_1 = 1.4, \mu_2 = 2.5, \gamma_1 = 0.2, \gamma_2 = 0.4, \theta_1 = 0.1, \theta_2 = 0.2$**)**



(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

Table 1 **Stochastic optimality gap of different policies (percentage gap to the MDP)**
((a) 3 servers: $\lambda_1 = 1, \lambda_2 = 2, \mu_1 = 1.28, \mu_2 = 2.0, \gamma_1 = 0.2, \gamma_2 = 0.4, \theta_1 = 0.1, \theta_2 = 0.2, c_1 = 5, c_2 = 1$
(b) 3 servers: $\lambda_1 = 1, \lambda_2 = 2, \mu_1 = 1.09, \mu_2 = 1.7, \gamma_1 = 0.2, \gamma_2 = 0.4, \theta_1 = 0.1, \theta_2 = 0.2, c_1 = 5, c_2 = 1$
(c) 3 servers: $\lambda_1 = 1, \lambda_2 = 2, \mu_1 = 0.95, \mu_2 = 1.48, \gamma_1 = 0.2, \gamma_2 = 0.4, \theta_1 = 0.1, \theta_2 = 0.2, c_1 = 5, c_2 = 1$
(d) 3 servers: $\lambda_1 = 1, \lambda_2 = 2, \mu_1 = 0.84, \mu_2 = 1.32, \gamma_1 = 0.2, \gamma_2 = 0.4, \theta_1 = 0.1, \theta_2 = 0.2, c_1 = 5, c_2 = 1$**)**

Case (a) $\rho = 0.6$				Case (b) $\rho = 0.7$			
	Fluid policy curve	P_1	P_2		Fluid policy curve	P_1	P_2
Mean gap	**0.22%**	0.69%	8.34%	Mean gap	**1.11%**	2.43%	4.27%
Min gap	**0.01%**	0.17%	3.92%	Min gap	**0.01%**	0.88%	0.33%
Max gap	**0.41%**	1.74%	14.74%	Max gap	**2.85%**	4.60%	16.51%
Case (c) $\rho = 0.8$				Case (d) $\rho = 0.9$			
	Fluid policy curve	P_1	P_2		Fluid policy curve	P_1	P_2
Mean gap	**1.53%**	6.92%	0.77%	Mean gap	**0.95%**	17.31%	0.00%
Min gap	**0.25%**	4.19%	0.10%	Min gap	**0.09%**	12.41%	0.00%
Max gap	**3.74%**	10.28%	1.70%	Max gap	**1.76%**	24.71%	0.00%

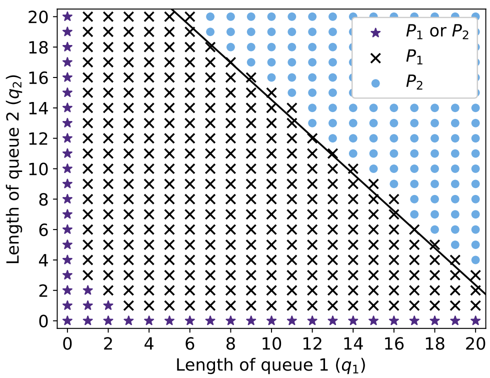
Figure 14 **Exact MDP solutions (Solid line is the corresponding fluid policy curve)**

((a) 3 servers: $\lambda_1 = 1, \lambda_2 = 2, \mu_1 = 1.28, \mu_2 = 2.0, \gamma_1 = 0.2, \gamma_2 = 0.4, \theta_1 = 0.1, \theta_2 = 0.2, c_1 = 5, c_2 = 1$

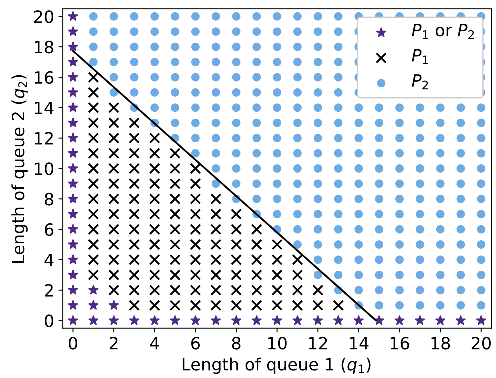
(b) 3 servers: $\lambda_1 = 1, \lambda_2 = 2, \mu_1 = 1.09, \mu_2 = 1.7, \gamma_1 = 0.2, \gamma_2 = 0.4, \theta_1 = 0.1, \theta_2 = 0.2, c_1 = 5, c_2 = 1$

(c) 3 servers: $\lambda_1 = 1, \lambda_2 = 2, \mu_1 = 0.95, \mu_2 = 1.48, \gamma_1 = 0.2, \gamma_2 = 0.4, \theta_1 = 0.1, \theta_2 = 0.2, c_1 = 5, c_2 = 1$

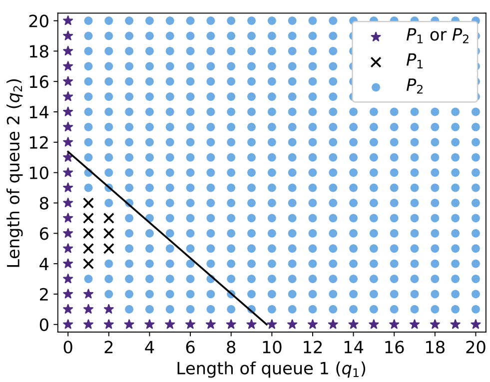
(d) 3 servers: $\lambda_1 = 1, \lambda_2 = 2, \mu_1 = 0.84, \mu_2 = 1.32, \gamma_1 = 0.2, \gamma_2 = 0.4, \theta_1 = 0.1, \theta_2 = 0.2, c_1 = 5, c_2 = 1$**)**



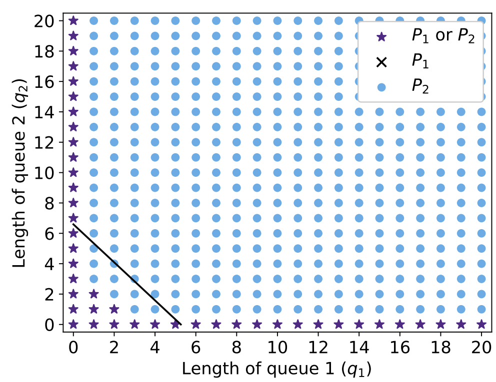
(a) $\rho = 0.6$



(b) $\rho = 0.7$



(c) $\rho = 0.8$



(d) $\rho = 0.9$

5. Model Generalizations

In this section, we consider two generalizations of the model. In the first one, we consider the case with time-varying arrival rates to capture the full demand shock period. In particular, we assume the arrival rates are high for a certain period of time before returning to the "normal" level, and study the transient optimal control problem in this setting. In the second one, we study a system with more than two classes where transition can happen between adjacent classes. We generalize the modified $c\mu/\theta$-index to this setting. As we will demonstrate subsequently, many of the insights we derived in the previous sections still hold in these generalizations.

5.1. Time-Varying Transient Arrival Rate

The transient optimal control problem is motivated by demand shock scenarios. The analysis in Section 4 focuses on an after-shock optimal control problem where the arrival rates satisfy $\lambda_1/\mu_1 + \lambda_2/\mu_2 < s$, but the system can have an arbitrarily large initial backlog. In this section, we consider a generalization where we include in our analysis the period of time during the shock. The shock raises the arrival rates for a fixed amount of time, during which the service capacity falls short and the queue increases. After the initial shock, the arrival rates restore to normal, i.e., the service capacity is able to meet demand and eventually empty the system. Formally, we impose the following assumption on the arrival rates.

Assumption 4 *The arrival rates to the system, $\{\lambda_1(t) : t \geq 0\}$ and $\{\lambda_2(t) : t \geq 0\}$, are non-negative, and there exists some $T \geq 0$ such that*
(1) $\lambda_1(t)$ and $\lambda_2(t)$ are continuously differentiable with respect to t over the time interval $[0,T)$;
(2) $q_1(t) > 0$ and $q_2(t) > 0$ for all $t \in [0,T]$ under any admissible scheduling policy $\pi \in \mathscr{F}$;
(3) $\lambda_1(t) = \lambda_1$ and $\lambda_2(t) = \lambda_2$ for some λ_1, λ_2 that satisfy $\lambda_1/\mu_1 + \lambda_2/\mu_2 < s$, for all $t \geq T$.

Under Assumption 4, $\lambda_1(t)$ and $\lambda_2(t)$ can be any continuously differential functions with argument t over the initial interval $[0,T)$ (condition (1)). The demand shock needs to be high enough such that neither queues empties during the shock (condition (2)). Lastly, after the shock, we have enough resources to bring the queue all the way back to zero (condition (3)). With a slight abuse of notation, we define

$$\tau := \inf\{t \geq T : q_1(t) + q_2(t) = 0\} - T$$

Under Assumption 4 condition (3), $\tau < \infty$.

Recall that the fluid dynamics are defined via $f(q,z) = (f_1(q,z), f_2(q,z))$, where $f_1(q,z) = \lambda_1(t) - z_1\mu_1 - \theta_1 q_1 - \gamma_1 q_1 + \gamma_2 q_2$ and $f_2(q,z) = \lambda_2(t) - z_2\mu_2 - \theta_2 q_2 - \gamma_2 q_2 + \gamma_1 q_1$. For the initial period $[0,T)$ with potentially time-varying arrival rate, we will add a time component to f, i.e., $f(q,z,t)$, to reflect the time dependence. The transient fluid optimization problem can be formulated as a two-stage optimal control problem. In particular, the first-stage problem (over the initial time period $[0,T)$) is expressed as

$$\begin{aligned} \min_{\{z(t):0\leq t<T\}} \quad & \int_0^T F\left(q(t)\right) dt + \Xi(q(T)) \\ s.t. \quad & dq(t) = f\left(q(t), z(t), t\right) \\ & h(z(t)) \leq 0, \end{aligned} \tag{12}$$

where $\Xi(q(T))$ is the terminal cost and is the optimal objective value for the second-stage problem

$$
\begin{aligned}
\min_{\{z(t):T\le t\le T+\tau\}} \quad & \int_{T}^{T+\tau} F\left(q(t)\right) dt \\
s.t. \quad & dq(t) = f\left(q(t), z(t)\right) \\
& g(q(t)) \le 0 \\
& h(z(t)) \le 0.
\end{aligned}
\tag{13}
$$

Note that the first-stage problem (12) is "explicit" without the state constraint $g(q(t)) \le 0$, because under Assumption 4, there does not exist an admissible control under which either of the queues gets emptied during $[0,T]$. Let $q^*(T)$ denote the optimal queue length at the end of the initial time horizon in problem (12). The second-stage problem (13) is the same as (F2′) if we shift the time from $[T, T+\tau]$ to $[0,\tau]$ and set the initial condition $q(0) := q(T)$. Due to this connection, the structural insights from the case of constant arrival rates in Section 4 is maintained in this time-varying case.

Theorem 3 *Under Assumptions 1, 3, and 4, for the transient optimal control problem* (12) - (13)*:*

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

II. If the $c\mu$-rule prioritizes Class i but the modified $c\mu/\theta$-rule prioritizes Class j, for $i \ne j$, $i,j =$ $1,2$, there exist positive real numbers ϵ and M with $0 < \epsilon < M$, such that for $t \in [T, T+\tau^*]$, 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 entire transient time horizon $[0, T+\tau^*]$.

Theorem 3 indicates that for time-varying arrival rates satisfying Assumption 4, the optimal control switches priority at most once from the modified $c\mu/\theta$-rule to the $c\mu$-rule. However, different from the case of fixed arrival rates, the modified $c\mu/\theta$-rule can be optimal during the demand shock (i.e. $[0,T)$), even for very small queues if the demand rate and/or the duration of the shock are sufficiently large. This indicates that when the priority switches is not only state-dependent but also time-dependent. As a simple consequence of Theorem 3, the following corollary characterizes the optimal transient control for sufficiently large demand shocks.

Corollary 1 *For the two-stage transient control problem, let $\mathcal{P}$ be the policy curve from the second-stage problem, and $M \in \mathbb{R}_+$ be defined in Theorem 3. If the arrival rates $\{\lambda_1(t) : t \ge 0\}$ and $\{\lambda_2(t) :$ $t \ge 0\}$ are such that $q_1(T) + q_2(T) > M$ under any admissible control, the optimal control employs* *the modified $c\mu/\theta$-rule over the interval $[0,T]$, and switches to the $c\mu$-rule when $t>T$ and the state crosses the policy curve $\mathcal{P}$, namely, when $q_1(t)+q_2(t)\in\mathcal{P}$.*

With general time-varying arrival rates, characterizing when and where the priority switches can be very complicated and highly case-dependent. For example, the switching point depends on where the system is initialized, how long the demand shock lasts, etc. As such, we leverage the insights from Theorem 3 and Corollary 1, and propose two heuristic policies. In both heuristics, we first derive the policy curve based on the optimal control problem (F2′) with the normal arrival rates, i.e., λ_1 and λ_2 (ignoring the demand shock). In Heuristic 1, we apply a time-homogeneous policy where we follow the $c\mu$-rule when the queues are "below" the policy curve, and follow the modified $c\mu/\theta$-rule when the queues are "above" the policy curve. In Heuristic 2, we modify the policy to be time-dependent. In particular, we employ the modified $c\mu/\theta$-rule for the initial demand-shock period $[0,T)$. Then, for $t\geq T$, we follow the $c\mu/\theta$-rule when the queue is "above" the policy curve, and follow the $c\mu$-rule when the queues are "below" the policy curve. In Table 2, we compare the performance of (i) Heuristic 1, (ii) Heuristic 2, (iii) the $c\mu$-rule, and (iv) the modified $c\mu/\theta$-rule. The problem instances we consider have piecewise constant arrival rates where the arrival rates switch from a fixed high level to a fixed low level, and we vary the duration of the high demand period, T. We observe that when the demand shock lasts for a sufficiently long time, i.e., $T\geq 0.4$, Heuristic 2 performs near optimal. However, when the demand shock lasts for only a short period of time, i.e., $T=0.1, 0.2$, Heuristic 1 performs very well. This is because, in the later case, the queues barely build up during the demand shock, and it is optimal to apply the $c\mu$-rule throughout.

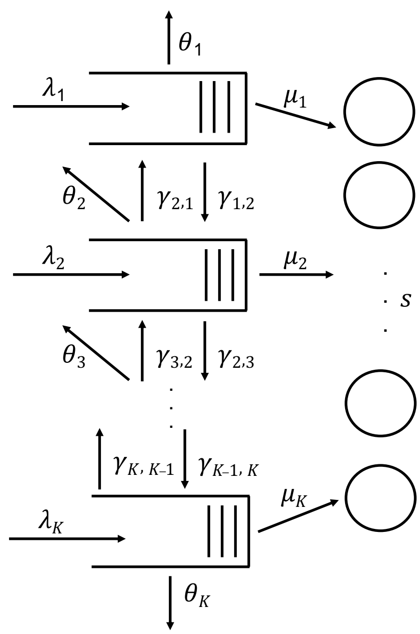
Table 2 Fluid optimality gap of different policies (percentage gap to (12))
($\lambda_1=10, \lambda_2=20, \mu_1=1, \mu_2=2.5, \gamma_1=0.2, \gamma_2=0.4, \theta_1=0.1, \theta_2=0.2, s=26, c_1=5, c_2=1, q(0)=(1,1), \lambda_i(t)=s\mu_i$ for $t\in[0,T), i=1,2$)

Demand shock duration T	Heuristic 1	Heuristic 2	$c\mu$	Modified $c\mu/\theta$
0.1	**0.00%**	18.90%	**0.00%**	46.15%
0.2	**0.00%**	10.15%	**0.00%**	27.41%
0.4	6.14%	**0.00%**	18.48%	6.84%
0.6	9.29%	**0.00%**	72.19%	3.43%
0.8	10.20%	**0.00%**	127.39%	2.05%
1.0	10.33%	**0.00%**	160.89%	1.41%
1.2	10.27%	**0.00%**	179.99%	1.22%
1.4	10.08%	**0.00%**	190.41%	1.03%
1.6	9.54%	**0.00%**	195.29%	0.77%

5.2. Multi-Class System

Thus far, our analysis has focused on a two-class system. We next discuss an extension to a multi-class system with K customer classes as depicted in Figure 15. The customer classes can be interpreted as having different urgency levels, with Class 1 being the most urgent and Class K being the least. Class i is associated with its arrival rate λ_i, service rate μ_i, abandonment rate θ_i, and cost rate c_i, $i=1,...,K$. To capture class-transitions, delayed Class i customers degrade into Class $i-1$ at rate $\gamma_{i,i-1}$, and improve to Class $i+1$ at rate $\gamma_{i,i+1}$, where $\gamma_{1,0}, \gamma_{K,K+1} := 0$. Note that this multi-class model can capture the case where some customers never transition type by setting the corresponding transition rates to 0.

Figure 15 Multi-class queueing network



5.2.1. 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:

$$
\begin{aligned}
\min_{\{z_i^e,\ i=1,\dots,K\}} \quad & \sum_{i=1}^{K} c_i q_i^e \\
s.t. \quad & \lambda_i - \mu_i z_i^e - (\gamma_{i,i-1} + \gamma_{i,i+1} + \theta_i) q_i^e + \gamma_{i+1,i} q_{i+1}^e + \gamma_{i-1,i} q_{i-1}^e = 0, \quad i = 1, ..., K \\
& \sum_{i=1}^{K} z_i^e \le s \\
& z_i^e, q_i^e \ge 0, \quad i = 1, ..., K.
\end{aligned}
\tag{14}
$$

Set $\tilde{\gamma}_{K,K+1} := 0$, and for i decreasing from $K-1$ to 1, we define the modified class improvement rates sequentially as

$$\tilde{\gamma}_{i,i+1} := \gamma_{i,i+1} \frac{\theta_{i+1} + \tilde{\gamma}_{i+1,i+2}}{\gamma_{i+1,i} + \theta_{i+1} + \tilde{\gamma}_{i+1,i+2}}. \tag{15}$$

Similarly, set $\tilde{\gamma}_{1,0} := 0$, and for i increasing from 2 to K, the modified class deterioration rates are sequentially defined by

$$\tilde{\gamma}_{i,i-1} := \gamma_{i,i-1} \frac{\theta_{i-1} + \tilde{\gamma}_{i-1,i-2}}{\gamma_{i-1,i} + \theta_{i-1} + \tilde{\gamma}_{i-1,i-2}}, \tag{16}$$

The modified improvement and deterioration rates in (15) and (16) can be understood as the effective class-transition rates adjusted for potential feedback. For example, the nominal deterioration rate from Class 2 to Class 1 is adjusted from $\gamma_{2,1}$ to $\tilde{\gamma}_{2,1} = \gamma_{2,1}\theta_1/(\theta_1 + \gamma_{1,2})$. Intuitively, if no service is provided, out of the customers that have degraded to Class 1, a proportion $\theta_1/(\theta_1 + \gamma_{1,2})$ will be fed back to Class 2. Thus, the effective degradation rate is $\tilde{\gamma}_{2,1}$.

Rearranging the terms in (14), we can derive that the optimal solution to (14) assigns the maximum value to the z_i^e with a larger modified $c\mu/\theta$-index, r_i, where

$$\begin{aligned} r_i := \mu_i \Bigg(& \frac{c_i}{\theta_i + \tilde{\gamma}_{i,i-1} + \tilde{\gamma}_{i,i+1}} + \sum_{j=1}^{i-1} \frac{c_j}{\theta_j + \tilde{\gamma}_{j,j-1} + \tilde{\gamma}_{j,j+1}} \prod_{k=j+1}^{i} \frac{\gamma_{k,k-1}}{\gamma_{k,k-1} + \tilde{\gamma}_{k,k+1} + \theta_k} \\ & + \sum_{j=i+1}^{K} \frac{c_j}{\theta_j + \tilde{\gamma}_{j,j-1} + \tilde{\gamma}_{j,j+1}} \prod_{k=i}^{j-1} \frac{\gamma_{k,k+1}}{\gamma_{k,k+1} + \tilde{\gamma}_{k,k-1} + \theta_k} \Bigg), \quad \text{for } i = 1, ..., K. \end{aligned}$$

Note that while the expression is more complex, this index has similar interpretation to that when there are only two 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 in (14) is an asymptotically stable equilibrium under the modified $c\mu/\theta$-rule. This requires extending the Lyapunov argument in Appendix A to the multi-class setting. We note that this task will become prohibitively tedious, especially for a large number of classes, K.

5.2.2. Transient Analysis The transient analysis for the two-class 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 class-transition and abandonment on the system dynamics is only second-order. Focusing on the service completions, we can show that 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 $i = 1, ..., K$, for ϵ sufficiently small.

Second, applying Pontryagin's Minimum Principle for the multi-class case, we see that at any time t, the optimal policy prioritizes the class with a larger $p_i^*(t)\mu_i$ value, where p_i^* is the optimal adjoint vector associated with Class i. 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 $\lim_{t\to\infty} p_i^*(\tau_1 - t)\mu_i = r_i$ (assuming we can extend the function to $t > \tau_1$), where r_i is the modified $c\mu/\theta$-index for Class i. This suggests that the modified $c\mu/\theta$-rule is likely to be optimal when the queues are far enough from the origin. However, we emphasize that this is only a heuristic argument. Rigorously establishing such a result requires highly non-trivial derivations.

Lastly, the optimal scheduling policy for areas between the ϵ-neighborhood of the origin and the far from the origin region remains unclear. Noticably, it is not necessarily true that the optimal policy 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. The solutions to (F2$'$) confirm 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 switches priority rule at most once. However, there are also instances where the optimal scheduling policy switches priority more than once, and it follows neither the modified $c\mu/\theta$-rule nor the $c\mu$-rule during part of the transient horizon.

To facilitate implementations, we propose a *one-switch policy*, where we switch priority at most once, and follow the the modified $c\mu/\theta$-rule when the system state is far from the equilibrium and the $c\mu$-rule when the state is close to the equilibrium. Table 3 compares the performance of the *one-switch policy*, the modified $c\mu/\theta$-rule, and the $c\mu$-rule. For the *one-switch policy*, we find the optimal policy curve when imposing that at most one switch is allowed. According to the system parameters, 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 these systems, the optimal LP solution may, under certain initial conditions, prioritize Class 2 over part of the transient horizon. Nevertheless, the sub-optimality gap of the one-switch policy is fairly small, i.e., less than 2.6%, while applying the modified $c\mu/\theta$-rule or the $c\mu$-rule throughout can sometimes lead to very large sub-optimality gaps. In general, we expect the one-switch policy to be a reasonable heuristic policy when the modified $c\mu/\theta$-index and the $c\mu$-index are relatively aligned.

6. Conclusion

In this work, we propose a novel multi-class queueing model to capture the class-transition behavior (e.g., degradation or improvement) in service systems. Our analysis provides insights into how proactive service should be utilized. We identify an important metric, the modified $c\mu/\theta$-index, which plays a critical role in specifying the optimal scheduling policy and lends itself to a very intuitive interpretation. In particular, as in the case of the conventional $c\mu/\theta$-index, the modified

Table 3 Fluid optimality gap of different policies (percentage gap to $(\mathrm{F2}')$)

($\lambda_1 = 10, \lambda_2 = 20, \lambda_3 = 30, \mu_1 = 4, \mu_2 = 5, \mu_3 = 6, \theta_1 = 0.2, \theta_2 = 0.1, \theta_3 = 0.1, \gamma_{2,1} = 0.1, \gamma_{1,2} = 0.2, \gamma_{3,2} = 0.1, \gamma_{2,3} = 0.3, s = 30, c_1 = 20, c_2 = 15, c_3 = 10, c\mu\text{-index} = \{80, 75, 60\}, \text{modified } c\mu/\theta\text{-index} = \{433, 583, 650\}$)

Initialization	One-switch	Modified $c\mu/\theta$	$c\mu$
(5, 5, 5)	**0.00%**	15.83%	**0.00%**
(10, 10, 10)	**0.00%**	14.80%	**0.00%**
(50, 50, 50)	**0.60%**	9.39%	**0.60%**
(100, 100, 100)	**2.21%**	6.56%	**2.21%**
(250, 250, 250)	**2.57%**	2.84%	5.09%
(500, 500, 500)	**0.95%**	1.04%	7.13%
(1000, 1000, 1000)	**0.36%**	0.42%	8.11%

$c\mu/\theta$-index balances the relative importance of holding costs, service times, and abandonment rates. Moreover, it augments the standard $c\mu/\theta$-index by several important additional terms that account for class-transitions.

We study both the long-run average cost and the transient cost minimization problem. When planning the system in the long run, we show that following the modified $c\mu/\theta$-rule is optimal. When considering the most cost-effective way to clear backlogs created by demand shocks, one should employ the modified $c\mu/\theta$-rule when the system has a very large backlog (i.e., when it is essential to account for the abandonment and class-transition dynamics), and follow the $c\mu$-rule when the system has a sufficiently small backlog (i.e., when cost minimization is driven by service completions).

We assume, throughout the paper, that class-transitions and abandonment happen according to independent exponential clocks, i.e., class-transition and abandonment have constant failure rates. It has been shown in a number of service settings, the patience times may have increasing or decreasing failure rates (Puha and Ward 2019). One can potentially extend the long-run average cost minimization problem to incorporate non-exponential class-transitions and abandonments. In particular, characterizing the fluid equilibrium points in these cases follows similar lines of analysis as in Whitt (2006). However, since the optimality of the conventional $c\mu/\theta$-rule may no longer hold for non-exponential patience-time distributions (Puha and Ward 2019), we expect the optimality of the modified $c\mu/\theta$-rule may also not hold with general class-transitions. We leave this as an interesting direction for future research.

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.

Altman, D., S. Ashtar, M. Olivares, G. Yom-Tov. 2019. Do customer emotions affect worker speed? an empirical study of emotional load in online customer contact centers. *Working Paper* .

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.

Jerath, K., A. Kumar, S. Netessine. 2015. An information stock model of customer behavior in multichannel customer support services. *Manufacturing & Service Operations Management* **17**(3) 368–383.

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. L., A. R. Ward. 2019. Scheduling an overloaded multiclass many-server queue with impatient customers. *Operations Research & Management Science in the Age of Analytics*. INFORMS, 189–217.

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.

Whitt, W. 2006. Fluid models for multiserver queues with abandonments. *Operations research* **54**(1) 37–54.

Wierman, A. 2011. Fairness and scheduling in single server queues. *Surveys in Operations Research and Management Science* **16**(1) 39–48.

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.

Appendix A: Long-Run Regularity of the Fluid Model and Proof of Theorem 1

The key to proving Theorem 1 is to establish that the optimal solution to the long-run average fluid optimization problem, (F1), is a globally asymptotically stable equilibrium under the strict priority rule suggested by the modified $c\mu/\theta$-index. In this section, we first establish the long-run regularity of the fluid model under strict priority rules. The stability analysis of strict priority rules can be of independent interest, especially as such policies are often used in practice. Additionally, we identify an interesting bi-stability phenomenon for certain parameter regions under strict priority rules. We then use the long-run regularity results to prove Theorem 1.

A.1. System Stability under Strict Priority Rules

Due to the symmetry of the system, we provide the analysis for strict priority to Class 1 only, i.e., the analysis for strict priority to Class 2 follows identically by symmetry. Under P_1, when $q_1(t)>0$, we will assign all capacity to Class 1. When $q_1(t)=0$, we will assign to Class 1 the minimum amount of capacity necessary 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 $q_1(t)>0$,

$$dq_1(t)=\lambda_1-\mu_1 s-\theta_1 q_1(t)-\gamma_1 q_1(t)+\gamma_2 q_2(t), \quad dq_2(t)=\lambda_2-\theta_2 q_2(t)-\gamma_2 q_2(t)+\gamma_1 q_1(t); \tag{17}$$

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

$$\begin{aligned} dq_1(t) &= \lambda_1-\mu_1\left(\frac{\lambda_1+\gamma_2 q_2(t)}{\mu_1}\wedge s\right)+\gamma_2 q_2(t), \\ dq_2(t) &= \lambda_2-\mu_2\left(s-\frac{\lambda_1+\gamma_2 q_2(t)}{\mu_1}\right)^+-\theta_2 q_2(t)-\gamma_2 q_2(t); \end{aligned} \tag{18}$$

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

$$dq_1(t)=\lambda_1-\mu_1\left(\frac{\lambda_1}{\mu_1}\wedge s\right), \quad dq_2(t)=\lambda_2-\mu_2\left(\left(s-\frac{\lambda_1}{\mu_1}\right)^+\wedge\frac{\lambda_2}{\mu_2}\right). \tag{19}$$

Using a Lyapunov argument, Theorem 4 characterizes the long-run regularity of the fluid dynamical system under strict priority to Class 1.

Theorem 4 *Under Assumption 1, for the dynamical system* (17) - (19)*,*

Case I. *When* $\mu_1>\frac{\gamma_2}{\theta_2+\gamma_2}\mu_2$*,*

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 $\frac{\lambda_1}{\mu_1}+\frac{\gamma_2}{\theta_2+\gamma_2}\frac{\lambda_2}{\mu_1}\leq s<\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_2}$**,** *the system has a globally asymptotically stable equilibrium at*

$$q_1^e=0, \quad q_2^e=\frac{\mu_1\mu_2\left(\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_2}-s\right)}{(\theta_2+\gamma_2)\mu_1-\gamma_2\mu_2}>0.$$

Ic If $s<\frac{\lambda_1}{\mu_1}+\frac{\gamma_2}{\theta_2+\gamma_2}\frac{\lambda_2}{\mu_1}$**.** *the system has a globally asymptotically stable equilibrium at*

$$q_1^e=\frac{\lambda_1+\frac{\gamma_2}{\theta_2+\gamma_2}\lambda_2-s\mu_1}{\theta_1+\gamma_1\frac{\theta_2}{\theta_2+\gamma_2}}>0, \quad q_2^e=\frac{\lambda_2\theta_1+\gamma_1(\lambda_1+\lambda_2-s\mu_1)}{(\theta_2+\gamma_2)\theta_1+\gamma_1\theta_2}>\frac{\lambda_2}{\theta_2+\gamma_2}.$$

Case II. *When* $\mu_1<\frac{\gamma_2}{\theta_2+\gamma_2}\mu_2$*,*

IIa If $\frac{\lambda_1}{\mu_1}+\frac{\gamma_2}{\theta_2+\gamma_2}\frac{\lambda_2}{\mu_1}<s$**,** *the system has a globally asymptotically stable equilibrium at*

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

IIb If $\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_2} < s \le \frac{\lambda_1}{\mu_1}+\frac{\gamma_2}{\theta_2+\gamma_2}\frac{\lambda_2}{\mu_1}$, *the system has two locally asymptotically stable equilibria*

$$q_{11}^e = 0, \quad q_{21}^e = 0; \qquad \textit{and} \qquad q_{12}^e = \frac{\lambda_1 + \frac{\gamma_2}{\theta_2+\gamma_2}\lambda_2 - s\mu_1}{\theta_1 + \gamma_1 \frac{\theta_2}{\theta_2+\gamma_2}} \ge 0, \quad q_{22}^e = \frac{\lambda_2\theta_1 + \gamma_1(\lambda_1+\lambda_2-s\mu_1)}{(\theta_2+\gamma_2)\theta_1+\gamma_1\theta_2} \ge \frac{\lambda_2}{\theta_2+\gamma_2}.$$

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*

$$q_{12}^e = \frac{\lambda_1 + \frac{\gamma_2}{\theta_2+\gamma_2}\lambda_2 - s\mu_1}{\theta_1 + \gamma_1 \frac{\theta_2}{\theta_2+\gamma_2}} > 0, \quad q_{22}^e = \frac{\lambda_2\theta_1 + \gamma_1(\lambda_1+\lambda_2-s\mu_1)}{(\theta_2+\gamma_2)\theta_1+\gamma_1\theta_2} > \frac{\lambda_2}{\theta_2+\gamma_2}.$$

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

$$q_1^e = \frac{\lambda_1 + \frac{\gamma_2}{\theta_2+\gamma_2}\lambda_2 - s\mu_1}{\theta_1 + \gamma_1 \frac{\theta_2}{\theta_2+\gamma_2}} > 0, \quad q_2^e = \frac{\lambda_2\theta_1 + \gamma_1(\lambda_1+\lambda_2-s\mu_1)}{(\theta_2+\gamma_2)\theta_1+\gamma_1\theta_2} > \frac{\lambda_2}{\theta_2+\gamma_2}.$$

Remark 4 *We note that when $\mu_1 = \gamma_2\mu_2/(\theta_2+\gamma_2)$, the system can have uncountably many equilibrium points. In particular, for $s = \lambda_1/\mu_1 + \lambda_2/\mu_2$, any (q_1^e, q_2^e) satisfying $q_1^e = 0$ and $(\lambda_1 + \gamma_2 q_2^e)/\mu_1 < s$ is an equilibrium point. We do not consider this parameter regime, i.e. $\mu_1 = \gamma_2\mu_2/(\theta_2+\gamma_2)$, in this paper.*

Proof: [Proof of Theorem 4] The stability analysis for P_1 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 similar lines 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 proofs for the rest of the cases given the appropriate Lyapunov functions follow similarly and are thus omitted.

The Lyapunov function we utilize to prove each case differs; they are summarized in the table below.

	Lyapunov function
Case Ia	Subcase 1: $\frac{1}{\mu_1}\lvert q_1-q_1^e\rvert + \frac{1}{\mu_2}\lvert q_2-q_2^e\rvert$; Subcase 2: $\lvert q_1-q_1^e\rvert + \lvert q_2-q_2^e\rvert$
Case Ib	Subcase 1: $\frac{1}{\mu_1}\lvert q_1-q_1^e\rvert + \frac{1}{\mu_2}\lvert q_2-q_2^e\rvert$; Subcase 2: $\lvert q_1-q_1^e\rvert + \lvert q_2-q_2^e\rvert$
Case Ic	$\lvert q_1-q_1^e\rvert + \lvert q_2-q_2^e\rvert$
Case IIa	$\lvert q_1-q_1^e\rvert + \frac{\gamma}{\theta_2+\gamma}\lvert q_2-q_2^e\rvert$
Case IIb	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 $\left(\frac{\lambda_1+\frac{\gamma_2}{\theta_2+\gamma_2}\lambda_2-s\mu_1}{\theta_1+\gamma_1\frac{\theta_2}{\theta_2+\gamma_2}}, \frac{\lambda_2\theta_1+\gamma_1(\lambda_1+\lambda_2-s\mu_1)}{(\theta_2+\gamma_2)\theta_1+\gamma_1\theta_2}\right)$: $\lvert q_1-q_1^e\rvert + \lvert q_2-q_2^e\rvert$
Case IIc	$\lvert q_1-q_1^e\rvert + \lvert q_2-q_2^e\rvert$
Case IId	$\lvert q_1-q_1^e\rvert + \lvert q_2-q_2^e\rvert$

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 $q \ne q^e$. 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. $\frac{\gamma_2}{\theta_2+\gamma_2} < \frac{\mu_1}{\mu_2}$**, i.e.,** $\frac{\gamma_2}{\mu_1} - \frac{\theta_2+\gamma_2}{\mu_2} < 0$**.**

Ia. If $\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} \leq s$**.**

Ia.Subcase 1. $\frac{\gamma_2}{\theta_2+\gamma_2} < \frac{\mu_1}{\mu_2} < \frac{\theta_1+\gamma_1}{\gamma_1}$**.**

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 $q_1(t) > 0$,

$$dq_1(t) = \lambda_1 - \mu_1 s - \theta_1 q_1(t) - \gamma_1 q_1(t) + \gamma_2 q_2(t)$$

$$dq_2(t) = \lambda_2 - \theta_2 q_2(t) - \gamma_2 q_2(t) + \gamma_1 q_1(t).$$

$$\begin{aligned}\nabla_q V(q)^T f(q) &= \frac{1}{\mu_1}(\lambda_1 - \mu_1 s - \theta_1 q_1(t) - \gamma_1 q_1(t) + \gamma_2 q_2(t)) + \frac{1}{\mu_2}(\lambda_2 - \theta_2 q_2(t) - \gamma q_2(t) + \gamma_1 q_1(t)) \\ &= \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} - s + \left(\frac{\gamma_1}{\mu_2} - \frac{\theta_1+\gamma_1}{\mu_1}\right) q_1(t) + \left(\frac{\gamma_2}{\mu_1} - \frac{\theta_2+\gamma_2}{\mu_2}\right) q_2(t) \\ &< 0,\end{aligned}$$

where the last inequality follows from the facts that $\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} \leq s$, and $\frac{\gamma_2}{\theta_2+\gamma_2} < \frac{\mu_1}{\mu_2} < \frac{\theta_1+\gamma_1}{\gamma_1}$.

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

(a) if $\frac{\lambda_1+\gamma_2 q_2(t)}{\mu_1} \geq s$,

$$dq_1(t) = \lambda_1 - \mu_1 s + \gamma_2 q_2(t)$$

$$dq_2(t) = \lambda_2 - \theta_2 q_2(t) - \gamma_2 q_2(t).$$

$$\begin{aligned}\nabla_q V(q)^T f(q) &= \frac{1}{\mu_1}(\lambda_1 - \mu_1 s + \gamma_2 q_2(t)) + \frac{1}{\mu_2}(\lambda_2 - \theta_2 q_2(t) - \gamma_2 q_2(t)) \\ &= \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} - s + \left(\frac{\gamma_2}{\mu_1} - \frac{\theta_2+\gamma_2}{\mu_2}\right) q_2(t) \\ &< 0,\end{aligned}$$

where the last inequality follows from the facts that $\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} \leq s$, and $\frac{\gamma_2}{\theta_2+\gamma_2} < \frac{\mu_1}{\mu_2} < \frac{\theta_1+\gamma_1}{\gamma_1}$.

(b) If $\frac{\lambda_1+\gamma_2 q_2(t)}{\mu_1} < s$,

$$dq_1(t) = \lambda_1 - \mu_1\left(\frac{\lambda_1 + \gamma_2 q_2(t)}{\mu_1}\right) + \gamma_2 q_2(t) = 0$$

$$dq_2(t) = \lambda_2 - \mu_2\left(s - \frac{\lambda_1 + \gamma_2 q_2(t)}{\mu_1}\right) - \theta_2 q_2(t) - \gamma_2 q_2(t).$$

$$\begin{aligned}\nabla_q V(q)^T f(q) &= \frac{1}{\mu_2}\left(\lambda_2 - \mu_2\left(s - \frac{\lambda_1 + \gamma_2 q_2(t)}{\mu_1}\right) - \theta_2 q_2(t) - \gamma_2 q_2(t)\right) \\ &= \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} - s + \left(\frac{\gamma_2}{\mu_1} - \frac{\theta_2+\gamma_2}{\mu_2}\right) q_2(t) \\ &< 0,\end{aligned}$$

where the last inequality follows from the facts that $\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} \leq s$, and $\frac{\gamma_2}{\theta_2+\gamma_2} < \frac{\mu_1}{\mu_2} < \frac{\theta_1+\gamma_1}{\gamma_1}$.

Ia.Subcase 2. $\frac{\gamma_2}{\theta_2+\gamma_2} < \frac{\theta_1+\gamma_1}{\gamma_1} < \frac{\mu_1}{\mu_2}$**.**

Consider Lyapunov function of the form

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

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

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

$$dq_1(t) = \lambda_1 - \mu_1 s - \theta_1 q_1(t) - \gamma_1 q_1(t) + \gamma_2 q_2(t)$$
$$dq_2(t) = \lambda_2 - \theta_2 q_2(t) - \gamma_2 q_2(t) + \gamma_1 q_1(t).$$

$$\begin{aligned}\nabla_q V(q)^T f(q) &= (\lambda_1 - \mu_1 s - \theta_1 q_1(t) - \gamma_1 q_1(t) + \gamma_2 q_2(t)) + (\lambda_2 - \theta_2 q_2(t) - \gamma_2 q_2(t) + \gamma_1 q_1(t)) \\ &= \lambda_1 - \mu_1 s + \lambda_2 - \theta_1 q_1(t) - \theta_2 q_2(t) \\ &= \mu_1 \left(\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_1} - s \right) - \theta_1 q_1(t) - \theta_2 q_2(t) \\ &< \mu_1 \left(\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} - s \right) - \theta_1 q_1(t) - \theta_2 q_2(t) \\ &< 0,\end{aligned}$$

where the first inequality follows from the fact that $\mu_1 > \mu_2$ (due to $\frac{\theta_1+\gamma_1}{\gamma_1} < \frac{\mu_1}{\mu_2}$), and the second inequality follows from the facts that $\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} \le s$ and $q_1(t) > 0$.

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

(a) If $\frac{\lambda_1 + \gamma_2 q_2(t)}{\mu_1} \ge s$,

$$dq_1(t) = \lambda_1 - \mu_1 s + \gamma_2 q_2(t)$$
$$dq_2(t) = \lambda_2 - \theta_2 q_2(t) - \gamma_2 q_2(t).$$

$$\begin{aligned}\nabla_q V(q)^T f(q) &= (\lambda_1 - \mu_1 s + \gamma_2 q_2(t)) + (\lambda_2 - \theta_2 q_2(t) - \gamma_2 q_2(t)) \\ &= \lambda_1 - \mu_1 s + \lambda_2 - \theta_2 q_2(t) \\ &= \mu_1 \left(\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_1} - s \right) - \theta_2 q_2(t) \\ &< \mu_1 \left(\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} - s \right) - \theta_2 q_2(t) \\ &< 0,\end{aligned}$$

where the first inequality follows from the fact that $\mu_1 > \mu_2$, and the second inequality follows from the facts that $\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} \le s$ and $q_2(t) > 0$.

(b) If $\frac{\lambda_1 + \gamma_2 q_2(t)}{\mu_1} < s$,

$$dq_1(t) = \lambda_1 - \mu_1 \left(\frac{\lambda_1 + \gamma_2 q_2(t)}{\mu_1} \right) + \gamma_2 q_2(t) = 0$$
$$dq_2(t) = \lambda_2 - \mu_2 \left(s - \frac{\lambda_1 + \gamma_2 q_2(t)}{\mu_1} \right) - \theta_2 q_2(t) - \gamma_2 q_2(t).$$

$$\nabla_q V(q)^T f(q) = \mu_2 \left(\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} - s + \left(\frac{\gamma_2}{\mu_1} - \frac{\theta_2 + \gamma_2}{\mu_2} \right) q_2(t) \right) < 0,$$

where the last inequality follows from the facts that $\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} \le s$, and $\frac{\gamma_2}{\theta_2+\gamma_2} < \frac{\theta_1+\gamma_1}{\gamma_1} < \frac{\mu_1}{\mu_2}$.

IIb. If $\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} < s \le \frac{\lambda_1}{\mu_1} + \frac{\gamma}{\theta_2+\gamma}\frac{\lambda_2}{\mu_1}$.

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.

(i) 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 $0 < \epsilon < \frac{s\mu_1 - \lambda_1}{\gamma_2}$ be such that

$$\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} - s + \left(\frac{\gamma_2}{\mu_1} - \frac{\theta_2 + \gamma_2}{\mu_2}\right)\epsilon < 0. \tag{20}$$

We know such ϵ exists because $\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} - s < 0$ and $\frac{\gamma_2}{\mu_1} - \frac{\theta_2+\gamma_2}{\mu_2} > 0$. Consider states (q_1, q_2) with $q_2 < \epsilon$.

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

$$dq_1(t) = \lambda_1 - \mu_1 s - \theta_1 q_1(t) - \gamma_1 q_1(t) + \gamma_2 q_2(t)$$

$$dq_2(t) = \lambda_2 - \theta_2 q_2(t) - \gamma_2 q_2(t) + \gamma_1 q_1(t).$$

$$\begin{aligned}
\nabla_q V(q)^T f(q) &= \frac{1}{\mu_1}(\lambda_1 - \mu_1 s - \theta_1 q_1(t) - \gamma_1 q_1(t) + \gamma_2 q_2(t)) + \frac{1}{\mu_2}(\lambda_2 - \theta_2 q_2(t) - \gamma q_2(t) + \gamma_1 q_1(t)) \\
&= \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} - s + \left(\frac{\gamma_1}{\mu_2} - \frac{\theta_1 + \gamma_1}{\mu_1}\right) q_1(t) + \left(\frac{\gamma_2}{\mu_1} - \frac{\theta_2 + \gamma_2}{\mu_2}\right) q_2(t) \\
&< \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} - s + \left(\frac{\gamma_2}{\mu_1} - \frac{\theta_2 + \gamma_2}{\mu_2}\right) q_2(t) \\
&< 0,
\end{aligned}$$

where the first inequality follows from the facts that $\frac{\mu_1}{\mu_2} < \frac{\gamma_2}{\theta_2+\gamma_2} < \frac{\theta_1+\gamma_1}{\gamma_1}$ and $q_1(t) > 0$, and the second inequality follows from (20) and the fact that $q_2(t) < \epsilon$.

(b) If $q_1(t) = 0$, $q_2(t) > 0$ (the assumption $q_2(t) < \epsilon$ implies that $\frac{\lambda_1 + \gamma_2 q_2(t)}{\mu_1} < s$),

$$dq_1(t) = \lambda_1 - \mu_1\left(\frac{\lambda_1 + \gamma_2 q_2(t)}{\mu_1}\right) + \gamma_2 q_2(t) = 0$$

$$dq_2(t) = \lambda_2 - \mu_2\left(s - \frac{\lambda_1 + \gamma_2 q_2(t)}{\mu_1}\right) - \theta_2 q_2(t) - \gamma_2 q_2(t).$$

$$\begin{aligned}
\nabla_q V(q)^T f(q) &= \frac{1}{\mu_2}\left(\lambda_2 - \mu_2\left(s - \frac{\lambda_1 + \gamma_2 q_2(t)}{\mu_1}\right) - \theta_2 q_2(t) - \gamma_2 q_2(t)\right) \\
&= \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} - s + \left(\frac{\gamma_2}{\mu_1} - \frac{\theta_2 + \gamma_2}{\mu_2}\right) q_2(t) \\
&< 0,
\end{aligned}$$

where the inequality follows from (20) and the fact that $q_2(t) < \epsilon$.

(ii) Local stability of $(q_1^e, q_2^e) = \left(\frac{\lambda_1 + \frac{\gamma_2}{\theta_2+\gamma_2}\lambda_2 - s\mu_1}{\theta_1 + \gamma_1\frac{\theta_2}{\theta_2+\gamma_2}}, \frac{\lambda_2\theta_1 + \gamma_1(\lambda_1 + \lambda_2 - s\mu_1)}{(\theta_2+\gamma_2)\theta_1 + \gamma_1\theta_2}\right)$: 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 $q_2 > 0$.

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

$$dq_1(t) = \lambda_1 - \mu_1 s - \theta_1 q_1(t) - \gamma_1 q_1(t) + \gamma_2 q_2(t)$$

$$dq_2(t) = \lambda_2 - \theta_2 q_2(t) - \gamma_2 q_2(t) + \gamma_1 q_1(t).$$

$$\begin{aligned}
\nabla_q V(q)^T f(q) &= (\lambda_1 - \mu_1 s - \theta_1 q_1(t) - \gamma_1 q_1(t) + \gamma_2 q_2(t)) + (\lambda_2 - \theta_2 q_2(t) - \gamma_2 q_2(t) + \gamma_1 q_1(t)) \\
&= \lambda_1 - \mu_1 s + \lambda_2 - \theta_1 q_1(t) - \theta_2 q_2(t) \\
&< \lambda_1 - \mu_1 s + \lambda_2 - \theta_1 q_1^e - \theta_2 q_2^e \\
&= 0,
\end{aligned}$$

where the inequality follows from the facts that $q_1(t) \geq q_1^e$, $q_2(t) \geq q_2^e$ and $(q_1(t), q_2(t)) \neq (q_1^e, q_2^e)$.

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

$$dq_1(t) = \lambda_1 - \mu_1 s - \theta_1 q_1(t) - \gamma_1 q_1(t) + \gamma_2 q_2(t)$$

$$dq_2(t) = \lambda_2 - \theta_2 q_2(t) - \gamma_2 q_2(t) + \gamma_1 q_1(t).$$

$$\begin{aligned}
\nabla_q V(q)^T f(q) &= (\lambda_1 - \mu_1 s - \theta_1 q_1(t) - \gamma_1 q_1(t) + \gamma_2 q_2(t)) - (\lambda_2 - \theta_2 q_2(t) - \gamma_2 q_2(t) + \gamma_1 q_1(t)) \\
&= \lambda_1 - \mu_1 s - \lambda_2 - (\theta_1 + 2\gamma_1) q_1(t) + (\theta_2 + 2\gamma_2) q_2(t) \\
&< \lambda_1 - \mu_1 s - \lambda_2 - (\theta_1 + 2\gamma_1) q_1^e + (\theta_2 + 2\gamma_2) q_2^e \\
&= 0,
\end{aligned}$$

where the inequality follows from the facts that $q_1(t) \geq q_1^e$ and $q_2(t) < q_2^e$.

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

$$dq_1(t) = \lambda_1 - \mu_1 s - \theta_1 q_1(t) - \gamma_1 q_1(t) + \gamma_2 q_2(t)$$

$$dq_2(t) = \lambda_2 - \theta_2 q_2(t) - \gamma_2 q_2(t) + \gamma_1 q_1(t).$$

$$\begin{aligned}
\nabla_q V(q)^T f(q) &= -(\lambda_1 - \mu_1 s - \theta_1 q_1(t) - \gamma_1 q_1(t) + \gamma_2 q_2(t)) + (\lambda_2 - \theta_2 q_2(t) - \gamma_2 q_2(t) + \gamma_1 q_1(t)) \\
&= -\lambda_1 + \mu_1 s + \lambda_2 + (\theta_1 + 2\gamma_1) q_1(t) - (\theta_2 + 2\gamma_2) q_2(t) \\
&< -\lambda_1 + \mu_1 s + \lambda_2 + (\theta_1 + 2\gamma_1) q_1^e - (\theta_2 + 2\gamma_2) q_2^e \\
&= 0,
\end{aligned}$$

where the inequality follows from the facts that $q_1(t) < q_1^e$ and $q_2(t) \geq q_2^e$.

(d) If $q_1(t) < q_1^e$ and $q_2(t) < q_2^e$,

$$dq_1(t) = \lambda_1 - \mu_1 s - \theta_1 q_1(t) - \gamma_1 q_1(t) + \gamma_2 q_2(t)$$

$$dq_2(t) = \lambda_2 - \theta_2 q_2(t) - \gamma_2 q_2(t) + \gamma_1 q_1(t).$$

$$\begin{aligned}
\nabla_q V(q)^T f(q) &= -(\lambda_1 - \mu_1 s - \theta_1 q_1(t) - \gamma_1 q_1(t) + \gamma_2 q_2(t)) - (\lambda_2 - \theta_2 q_2(t) - \gamma_2 q_2(t) + \gamma_1 q_1(t)) \\
&= -\lambda_1 + \mu_1 s - \lambda_2 + \theta_1 q_1(t) + \theta_2 q_2(t) \\
&< -\lambda_1 + \mu_1 s - \lambda_2 + \theta_1 q_1^e + \theta_2 q_2^e \\
&= 0,
\end{aligned}$$

where the inequality follows from the facts that $q_1(t) < q_1^e$ and $q_2(t) < q_2^e$. Q.E.D.

A.2. Proof of Theorem 1

PROOF: Consider the equivalent LP formulation (5) for the long-run average cost minimization 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 Class 2. 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 (or P_2 by symmetry) in Appendix A.1. Following similar parameter regimes examined in the stability analysis, we divide the analysis here into different cases. For each case, the tables below list the optimal LP solution (z_1^{e*}, z_2^{e*}), the corresponding (q_1^{e*}, q_2^{e*}), and the static control under which (q_1^{e*}, q_2^{e*}) is a globally asymptotically stable equilibrium.

Case I. $\frac{\mu_1}{\mu_2} < \frac{\gamma_2}{\theta_2 + \gamma_2}$. In this case, the modified $c\mu/\theta$-rule prioritizes Class 2.

	(z_1^{e*}, z_2^{e*})	(q_1^{e*}, q_2^{e*})	Control
$\frac{\lambda_1}{\mu_1}+\frac{\gamma_2}{\theta_2+\gamma_2}\frac{\lambda_2}{\mu_1}<s$	$\left(\frac{\lambda_1}{\mu_1},\frac{\lambda_2}{\mu_2}\right)$	$(0,0)$	P_1, P_2
$\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_2}\le s\le\frac{\lambda_1}{\mu_1}+\frac{\gamma_2}{\theta_2+\gamma_2}\frac{\lambda_2}{\mu_1}$	$\left(\frac{\lambda_1}{\mu_1},\frac{\lambda_2}{\mu_2}\right)$	$(0,0)$	P_2
$\frac{\lambda_2}{\mu_2}+\frac{\gamma_1}{\theta_1+\gamma_1}\frac{\lambda_1}{\mu_2}\le s<\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_2}$	$\left(s-\frac{\lambda_2}{\mu_2}-\frac{\gamma_1\mu_1\left(\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_2}-s\right)}{(\theta_1+\gamma_1)\mu_2-\gamma_1\mu_1},\ \frac{\lambda_2}{\mu_2}+\frac{\gamma_1\mu_1\left(\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_2}-s\right)}{(\theta_1+\gamma_1)\mu_2-\gamma_1\mu_1}\right)$	$\left(\frac{\mu_1\mu_2\left(\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_2}-s\right)}{(\theta_1+\gamma_1)\mu_2-\gamma_1\mu_1},0\right)$	P_2
$s<\frac{\lambda_2}{\mu_2}+\frac{\gamma_1}{\theta_1+\gamma_1}\frac{\lambda_1}{\mu_2}$	$(0,s)$	$\left(\frac{\lambda_1\theta_2+\gamma_2(\lambda_1+\lambda_2-s\mu_2)}{(\theta_1+\gamma_1)\theta_2+\gamma_2\theta_1},\frac{\lambda_2+\frac{\gamma_1}{\theta_1+\gamma_1}\lambda_1-s\mu_2}{\theta_2+\gamma_2\frac{\theta_1}{\theta_1+\gamma_1}}\right)$	P_2

Case II. $\frac{\gamma_2}{\theta_2+\gamma_2}<\frac{\mu_1}{\mu_2}<\frac{\theta_1+\gamma_1}{\gamma_1}$, and $\frac{\lambda_2}{\mu_2}+\frac{\gamma_1}{\theta_1+\gamma_1}\frac{\lambda_1}{\mu_2}\le\frac{\lambda_1}{\mu_1}+\frac{\gamma_2}{\theta_2+\gamma_2}\frac{\lambda_2}{\mu_1}$.

	Modified $c\mu/\theta$-rule prioritizes Class 1		
	(z_1^{e*}, z_2^{e*})	(q_1^{e*}, q_2^{e*})	Control
$\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_2}\le s$	$\left(\frac{\lambda_1}{\mu_1},\frac{\lambda_2}{\mu_2}\right)$	$(0,0)$	P_1, P_2
$\frac{\lambda_1}{\mu_1}+\frac{\gamma_2}{\theta_2+\gamma_2}\frac{\lambda_2}{\mu_1}\le s<\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_2}$	$\left(\frac{\lambda_1}{\mu_1}+\frac{\gamma_2\mu_2\left(\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_2}-s\right)}{(\theta_2+\gamma_2)\mu_1-\gamma_2\mu_2},\ s-\frac{\lambda_1}{\mu_1}-\frac{\gamma_2\mu_2\left(\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_2}-s\right)}{(\theta_2+\gamma_2)\mu_1-\gamma_2\mu_2}\right)$	$\left(0,\frac{\mu_1\mu_2\left(\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_2}-s\right)}{(\theta_2+\gamma_2)\mu_1-\gamma_2\mu_2}\right)$	P_1
$\frac{\lambda_2}{\mu_2}+\frac{\gamma_1}{\theta_1+\gamma_1}\frac{\lambda_1}{\mu_2}\le s\le\frac{\lambda_1}{\mu_1}+\frac{\gamma_2}{\theta_2+\gamma_2}\frac{\lambda_2}{\mu_1}$	$(s,0)$	$\left(\frac{\lambda_1+\frac{\gamma_2}{\theta_2+\gamma_2}\lambda_2-s\mu_1}{\theta_1+\gamma_1\frac{\theta_2}{\theta_2+\gamma_2}},\frac{\lambda_2\theta_1+\gamma_1(\lambda_1+\lambda_2-s\mu_1)}{(\theta_2+\gamma_2)\theta_1+\gamma_1\theta_2}\right)$	P_1
$s<\frac{\lambda_2}{\mu_2}+\frac{\gamma_1}{\theta_1+\gamma_1}\frac{\lambda_1}{\mu_2}$	$(s,0)$	$\left(\frac{\lambda_1+\frac{\gamma_2}{\theta_2+\gamma_2}\lambda_2-s\mu_1}{\theta_1+\gamma_1\frac{\theta_2}{\theta_2+\gamma_2}},\frac{\lambda_2\theta_1+\gamma_1(\lambda_1+\lambda_2-s\mu_1)}{(\theta_2+\gamma_2)\theta_1+\gamma_1\theta_2}\right)$	P_1

	Modified $c\mu/\theta$-rule prioritizes Class 2		
	(z_1^{e*}, z_2^{e*})	(q_1^{e*}, q_2^{e*})	Control
$\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_2}\le s$	$\left(\frac{\lambda_1}{\mu_1},\frac{\lambda_2}{\mu_2}\right)$	$(0,0)$	P_1, P_2
$\frac{\lambda_1}{\mu_1}+\frac{\gamma_2}{\theta_2+\gamma_2}\frac{\lambda_2}{\mu_1}\le s<\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_2}$	$\left(s-\frac{\lambda_2}{\mu_2}-\frac{\gamma_1\mu_1\left(\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_2}-s\right)}{(\theta_1+\gamma_1)\mu_2-\gamma_1\mu_1},\ \frac{\lambda_2}{\mu_2}+\frac{\gamma_1\mu_1\left(\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_2}-s\right)}{(\theta_1+\gamma_1)\mu_2-\gamma_1\mu_1}\right)$	$\left(\frac{\mu_1\mu_2\left(\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_2}-s\right)}{(\theta_1+\gamma_1)\mu_2-\gamma_1\mu_1},0\right)$	P_2
$\frac{\lambda_2}{\mu_2}+\frac{\gamma_1}{\theta_1+\gamma_1}\frac{\lambda_1}{\mu_2}\le s\le\frac{\lambda_1}{\mu_1}+\frac{\gamma_2}{\theta_2+\gamma_2}\frac{\lambda_2}{\mu_1}$	$\left(s-\frac{\lambda_2}{\mu_2}-\frac{\gamma_1\mu_1\left(\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_2}-s\right)}{(\theta_1+\gamma_1)\mu_2-\gamma_1\mu_1},\ \frac{\lambda_2}{\mu_2}+\frac{\gamma_1\mu_1\left(\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_2}-s\right)}{(\theta_1+\gamma_1)\mu_2-\gamma_1\mu_1}\right)$	$\left(\frac{\mu_1\mu_2\left(\frac{\lambda_1}{\mu_1}+\frac{\lambda_2}{\mu_2}-s\right)}{(\theta_1+\gamma_1)\mu_2-\gamma_1\mu_1},0\right)$	P_2
$s<\frac{\lambda_2}{\mu_2}+\frac{\gamma_1}{\theta_1+\gamma_1}\frac{\lambda_1}{\mu_2}$	$(0,s)$	$\left(\frac{\lambda_1\theta_2+\gamma_2(\lambda_1+\lambda_2-s\mu_2)}{(\theta_1+\gamma_1)\theta_2+\gamma_2\theta_1},\frac{\lambda_2+\frac{\gamma_1}{\theta_1+\gamma_1}\lambda_1-s\mu_2}{\theta_2+\gamma_2\frac{\theta_1}{\theta_1+\gamma_1}}\right)$	P_2

Case III. $\frac{\gamma_2}{\theta_2+\gamma_2}<\frac{\mu_1}{\mu_2}<\frac{\theta_1+\gamma_1}{\gamma_1}$, and $\frac{\lambda_2}{\mu_2}+\frac{\gamma_1}{\theta_1+\gamma_1}\frac{\lambda_1}{\mu_2}\ge\frac{\lambda_1}{\mu_1}+\frac{\gamma_2}{\theta_2+\gamma_2}\frac{\lambda_2}{\mu_1}$.

This case follows from Case 2 by symmetry.

Case IV. $\frac{\mu_1}{\mu_2}>\frac{\theta_1+\gamma_1}{\gamma_1}$. In this case, the modified $c\mu/\theta$-rule prioritizes Class 1.

This case follows from Case 1 by symmetry. Q.E.D.

Appendix B: Proofs of the Results in Section 4

The proofs in this section are organized as follows. We start by showing that it is without loss of optimality to restrict our analysis to solutions without 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 Propositions 2 and 3. In particular, we provide more details about Pontryagin's Minimum Principle. We next prove the auxiliary lemmas (Lemmas 2 – 4), which are then used to prove Propositions 2 and 3. Note that we actually prove Proposition 3 first, because the proof of Proposition 2 utilizes the results established in Proposition 3. Putting the results of Propositions 1 – 3 together, we complete the proof of Theorem 2. Lastly, we characterize the policy curve in the special case where $\gamma_1 = 0$, $c_1\mu_1 < c_2\mu_2$, and $r_1 > r_2$ (Proposition 4).

B.1. Proof of Lemma 1

Proof: We prove the lemma by first showing that the cost difference between a chattering trajectory and a properly constructed trajectory without chattering is negligible. This implies that any admissible control policy π that yields a chattering interval can be replaced by a cost-wise equivalent control $\tilde{\pi}$ that does not yield chattering state trajectories. Thus, it is without loss of optimality to consider state trajectories without chattering behavior for the transient optimal control problem (F2′).

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 $I_1 \cup I_2$, i.e. $[0, \epsilon + \epsilon']$.

Over the first interval I_1, the state trajectories evolve as

$$\begin{aligned} q_1(t) &= (q_{20}\gamma_2 + \lambda_1)\, t + o(\epsilon), \quad t \in [0, \epsilon] \\ q_2(t) &= q_{20} + (q_{20}(-\gamma_2 - \theta_2) + \lambda_2 - s\mu_2)\, t + o(\epsilon), \quad t \in [0, \epsilon]\,. \end{aligned}$$

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

$$\begin{aligned} q_1(\epsilon) &= (q_{20}\gamma_2 + \lambda_1)\, \epsilon + o(\epsilon) \\ q_2(\epsilon) &= q_{20} + (q_{20}(-\gamma_2 - \theta_2) + \lambda_2 - s\mu_2)\, \epsilon + o(\epsilon). \end{aligned}$$

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

$$\begin{aligned} q_1(t) &= \epsilon\,(q_{20}\gamma_2 + \lambda_1) + (t - \epsilon)\,[\lambda_1 + \epsilon(-\gamma_1 - \theta_1)(q_{20}\gamma_2 + \lambda_1) - s\mu_1 \\ &\quad + \gamma_2(q_{20} - q_{20}\gamma_2\epsilon - q_{20}\epsilon\theta_2 + \epsilon\lambda_2 - s\epsilon\mu_2)] + o(\epsilon), \quad t \in [\epsilon, \epsilon + \epsilon'] \\ q_2(t) &= q_{20} - q_{20}\gamma_2\epsilon - q_{20}\epsilon\theta_2 + \epsilon\lambda_2 - s\epsilon\mu_2 + (t - \epsilon)\,[\gamma_1\epsilon(q_{20}\gamma_2 + \lambda_1) + \lambda_2 \\ &\quad + (-\gamma_2 - \theta_2)(q_{20} - q_{20}\gamma_2\epsilon - q_{20}\epsilon\theta_2 + \epsilon\lambda_2 - s\epsilon\mu_2)] + o(\epsilon), \quad t \in [\epsilon, \epsilon + \epsilon']\,. \end{aligned}$$

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

$$\epsilon' = \frac{\epsilon\,(q_{20}\gamma_2 + \lambda_1)}{-q_{20}\gamma_2 - \lambda_1 + s\mu_1} + o(\epsilon).$$

The total holding cost over the two intervals I_1 and I_2 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 serve Class 2. Similarly, we characterize the corresponding state trajectory over this interval of length $\epsilon+\epsilon'$ as

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

$$\tilde{q}_2(t)=q_{20}+t\left[q_{20}(-\gamma_2\mu_1-\theta_2\mu_1+\gamma_2\mu_2)+\lambda_2\mu_1+\lambda_1\mu_2-s\mu_1\mu_2\right]/\mu_1+o(\epsilon), \quad t\in[0,\epsilon+\epsilon'],$$

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

$$\begin{aligned}C-\tilde{C}=&\frac{\epsilon^2}{2(q_{20}\gamma_2+\lambda_1-s\mu_1)^2}(q_{20}\gamma_2+\lambda_1)\big(c_2\epsilon(q_{20}\gamma_2+\lambda_1)\left(q_{20}(\gamma_1\gamma_2+(\gamma_2+\theta_2)^2)+\gamma_1\lambda_1-(\gamma_2+\theta_2)\lambda_2\right)\\&+c_2s\left((1+\epsilon(\gamma_2+\theta_2))(q_{20}\gamma_2+\lambda_1)-s\mu_1\right)\mu_2-c_1\big(q_{20}^2\gamma_2^2\epsilon(\gamma_1+\gamma_2+\theta_1+\theta_2)+\gamma_1\epsilon\lambda_1^2+\epsilon\theta_1\lambda_1^2\\&-\gamma_2\epsilon\lambda_1\lambda_2+s\lambda_1\mu_1-s^2\mu_1^2+s\gamma_2\epsilon\lambda_1\mu_2+q_{20}\gamma_2\left(2\gamma_1\epsilon\lambda_1+\epsilon(2\theta_1+\theta_2)\lambda_1+s\mu_1+\gamma_2\epsilon(\lambda_1-\lambda_2+s\mu_2)\right)\big)\big)\\=&\,o(\epsilon).\end{aligned} \tag{21}$$

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

$$q_2(\epsilon+\epsilon')-\tilde{q}_2(\epsilon+\epsilon')=-\frac{\epsilon^2(q_{20}\gamma_2+\lambda_1)\left(q_{20}(\gamma_1\gamma_2+(\gamma_2+\theta_2)^2)+\gamma_1\lambda_1-(\gamma_2+\theta_2)(\lambda_2-s\mu_2)\right)}{q_{20}\gamma_2+\lambda_1-s\mu_1}=o(\epsilon). \tag{22}$$

Importantly, (21) implies that the cost under the policy that has 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 $o(\epsilon)$. From (22), the queue lengths at time $\epsilon+\epsilon'$ under the two policies also differ by $o(\epsilon)$. Now for any interval of length L, suppose we divide it into $O(L/\epsilon)$ small triangles (trajectories where q_1 first increases for ϵ units of time and then decreases to zero). 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 versus non-chattering) is $o(\epsilon)O(L/\epsilon)$, which goes to zero for fixed L as ϵ goes to zero. 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 close enough 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

$$q_1(t) = \epsilon + (-\gamma_1\epsilon + \gamma_2\epsilon - \epsilon\theta_1 + \lambda_1 - s\mu_1)\, t + o(\epsilon), \quad t \in [0, t_1]$$
$$q_2(t) = \epsilon + (\gamma_1\epsilon - \gamma_2\epsilon - \epsilon\theta_2 + \lambda_2)\, t + o(\epsilon), \quad t \in [0, t_1]\,,$$

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 for the second interval, the fluid trajectory over the second interval is

$$\begin{aligned} q_1(t) &= 0, \quad t \in [t_1, t_1 + t_2] \\ q_2(t) &= \frac{[-\epsilon\theta_2\mu_1\,(-\lambda_1 + \lambda_2 + s\mu_1) + \gamma_2\epsilon\,(\lambda_1 - \lambda_2 - s\mu_1)\,(\mu_1 - \mu_2) - (\lambda_1 - s\mu_1)\,(\lambda_2\mu_1 + \lambda_1\mu_2 - s\mu_1\mu_2)]\,(t - t_1)}{\mu_1(-\lambda_1 + s\mu_1)} \\ &\quad + \epsilon + \frac{\epsilon\lambda_2}{-\lambda_1 + s\mu_1} + o(\epsilon), \quad t \in [t_1, t_1 + t_2]\,, \end{aligned}$$

and the value of t_2 is

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

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

$$\begin{aligned} C^{P_1} &= c_1 \int_0^{t_1} [\epsilon + (-\gamma_1\epsilon + \gamma_2\epsilon - \epsilon\theta_1 + \lambda_1 - s\mu_1)\, t]\, dt + c_2 \int_0^{t_1} [\epsilon + (\gamma_1\epsilon - \gamma_2\epsilon - \epsilon\theta_2 + \lambda_2)\, t]\, dt \\ &+ c_2 \int_{t_1}^{t_1+t_2} \Bigg\{ \frac{[-\epsilon\theta_2\mu_1\,(-\lambda_1 + \lambda_2 + s\mu_1) + \gamma_2\epsilon\,(\lambda_1 - \lambda_2 - s\mu_1)\,(\mu_1 - \mu_2) - (\lambda_1 - s\mu_1)\,(\lambda_2\mu_1 + \lambda_1\mu_2 - s\mu_1\mu_2)]\,(t - t_1)}{\mu_1(-\lambda_1 + s\mu_1)} \\ &+ \epsilon + \frac{\epsilon\lambda_2}{-\lambda_1 + s\mu_1} \Bigg\} dt + o(\epsilon^2) \\ &= \frac{\epsilon^2}{2(\lambda_1 - s\mu_1)} \left(-c_1 + \frac{c_2(\lambda_2\mu_2 - \lambda_1(\mu_1 + 2\mu_2) + s\mu_1(\mu_1 + 2\mu_2))}{\lambda_2\mu_1 + (\lambda_1 - s\mu_1)\mu_2} \right) + o(\epsilon^2). \end{aligned}$$

Next, we consider the **strict priority rule to Class 2**. Let C^{P_2} denote the total cost of clearing the fluid queue from initial backlog level $(q_1(0), q_2(0)) = (\epsilon, \epsilon)$. It follows by symmetry that

$$C^{P_2} = \frac{\epsilon^2}{2(\lambda_2 - s\mu_2)} \left(-c_2 + \frac{c_1(\lambda_1\mu_1 - \lambda_2(\mu_2 + 2\mu_1) + s\mu_2(\mu_2 + 2\mu_1))}{\lambda_1\mu_2 + (\lambda_2 - s\mu_2)\mu_1} \right) + o(\epsilon^2).$$

Comparing the total costs under P_1 and P_2, we get

$$\begin{aligned} C^{P_1} - C^{P_2} &= \frac{\epsilon^2(c_1\mu_1 - c_2\mu_2)\,(\lambda_1^2 - \lambda_1(2\lambda_2 + s(\mu_1 - 2\mu_2)) + (\lambda_2 + 2s\mu_1)(\lambda_2 - s\mu_2))}{2(\lambda_1 - s\mu_1)(-\lambda_2 + s\mu_2)(\lambda_2\mu_1 + (\lambda_1 - s\mu_1)\mu_2)} + o(\epsilon^2) \\ &= \frac{\epsilon^2(c_1\mu_1 - c_2\mu_2)\,(\lambda_1(\lambda_1 - s\mu_1) + (\lambda_2 - s\mu_2)(\lambda_2 + 2s\mu_1 - 2\lambda_1))}{2(s\mu_1 - \lambda_1)(s\mu_2 - \lambda_2)(s\mu_1\mu_2 - \lambda_1\mu_2 - \lambda_2\mu_1)} + o(\epsilon^2), \end{aligned} \tag{23}$$

Note that as $s > \lambda_1/\mu_1 + \lambda_2/\mu_2$, in (23), 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

In this section, we provide more details about Pontryagin's Minimum Principle, which will be used in the proof of Lemma 2 – 4 and Propositions 2 – 3. Consider the transient optimization problem (F2′) (also presented below).

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

The pure state constraint $g(q(t)) \leq 0$ 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

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

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

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

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

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

and

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

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, let

$$\zeta := \sqrt{\gamma_1^2 + 2\gamma_1(\gamma_2 + \theta_1 - \theta_2) + (\gamma_2 - \theta_1 + \theta_2)^2} = \sqrt{\gamma_2^2 + 2\gamma_2(\gamma_1 + \theta_2 - \theta_1) + (\gamma_1 - \theta_2 + \theta_1)^2}. \tag{24}$$

Note that ζ is well-defined, because

$$\begin{aligned} \gamma_1^2 + 2\gamma_1(\gamma_2 + \theta_1 - \theta_2) + (\gamma_2 - \theta_1 + \theta_2)^2 &= \gamma_1^2 + 2\gamma_1(\gamma_2 + \theta_1 - \theta_2) + (-\gamma_2 + \theta_1 - \theta_2)^2 \\ &\geq \gamma_1^2 + 2\gamma_1(-\gamma_2 + \theta_1 - \theta_2) + (-\gamma_2 + \theta_1 - \theta_2)^2 \\ &= (\gamma_1 - \gamma_2 + \theta_1 - \theta_2)^2 \geq 0. \end{aligned}$$

Solving the ordinary differential equations in (ADJ) for the dynamic of the adjoint vectors, we get

$$\begin{aligned}
p_1^*(t) = & \frac{1}{\zeta} e^{\frac{1}{2}t(\gamma_1+\gamma_2+\theta_1+\theta_2)} \Bigg\{ \zeta K_1(0) \cosh\left(\frac{t\zeta}{2}\right) + \sinh\left(\frac{t\zeta}{2}\right) \Bigg[(\gamma_1 - \gamma_2 + \theta_1 - \theta_2) K_1(0) - 2\gamma_1 K_2(0) \\
& - 2\gamma_1 \int_0^t \frac{1}{2\zeta} e^{-\frac{1}{2}(\gamma_1+\gamma_2+\theta_1+\theta_2+\zeta)u} \Big(2c_1\gamma_2 + c_2(\gamma_1 - \gamma_2 + \theta_1 - \theta_2 - \zeta) \\
& - e^{\zeta u}(2c_1\gamma_2 + c_2(\gamma_1 - \gamma_2 + \theta_1 - \theta_2 + \zeta)) + 2(-1 + e^{\zeta u})\gamma_2 \eta_1^*(u) \\
& + (-\gamma_1 + \gamma_2 - \theta_1 + \theta_2 + \zeta + e^{\zeta u}(\gamma_1 - \gamma_2 + \theta_1 - \theta_2 + \zeta))\eta_2^*(u) \Big) du \Bigg] \\
& + \left(\zeta \cosh\left(\frac{t\zeta}{2}\right) + (\gamma_1 - \gamma_2 + \theta_1 - \theta_2) \sinh\left(\frac{t\zeta}{2}\right) \right) \int_0^t \frac{1}{2\zeta} e^{-\frac{1}{2}(\gamma_1+\gamma_2+\theta_1+\theta_2+\zeta)u} \\
& \Big(- c_1(\gamma_1 - \gamma_2 + \theta_1 - \theta_2 + \zeta) + 2c_2\gamma_1 + (\gamma_1 - \gamma_2 + \theta_1 - \theta_2 + \zeta + e^{\zeta u}(-\gamma_1 + \gamma_2 - \theta_1 + \theta_2 + \zeta))\eta_1^*(u) \\
& - 2\gamma_1 \eta_2^*(u) - e^{\zeta u}(2c_2\gamma_1 + c_1(-\gamma_1 + \gamma_2 - \theta_1 + \theta_2 + \zeta) - 2\gamma_1 \eta_2^*(u)) \Big) du \Bigg\},
\end{aligned}$$

where $K_1(0), K_2(0)$ are constants that depend on $p_1^*(0)$ and $p_2^*(0)$. The expression for $p_2^*(t)$ follows by symmetry.

The adjoint vector is connected to the value function under the optimal control. 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 (9) - (10) reveal important properties of the optimal control structure. First, observe in (9) - (10) 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 positive 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 $g(q)$ is time invariant, i.e., the function g does not have a time argument, we have $\nabla_t g(q^*(\beta)) = 0$. 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), (H), (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:*

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

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 (9)-(10), 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 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 (9) - (10), 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. We prove the statement for the junction times associated with Class 1; the arguments for Class 2 follow by symmetry. The discussion is divided into three cases based on the relative level of service capacity s.

Case I. $\max\{\frac{\lambda_1}{\mu_1} + \frac{\gamma_2}{\theta_2+\gamma_2}\frac{\lambda_2}{\mu_1}, \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2}\} < s$**.**

(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,

$$dq_1^*(\beta-) = \lambda_1 - \mu_1 z_1^*(\beta-) - \theta_1 q_1^*(\beta-) - \gamma_1 q_1^*(\beta-) + \gamma_2 q_2^*(\beta-) = \lambda_1 - \mu_1 s + \gamma_2 q_2^*(\beta-).$$

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

$$q_2^*(t) < (s\mu_1 - \lambda_1)/\gamma_2 \quad \text{for all } t \in [\beta - \delta, \beta).$$

We next show that $dq_1^*(\beta-) < 0$. Suppose by contradiction $dq_1^*(\beta-) = 0$, then it must be the case that $q_2^*(\beta) = (s\mu_1 - \lambda_1)/\gamma_2$. On the other hand,

$$dq_2^*(t) = \lambda_2 - \mu_2 z_2^*(t) - (\theta_2 + \gamma_2) q_2^*(t) + \gamma_1 q_1^*(t) \le \lambda_2 - (\theta_2 + \gamma_2) q_2^*(t) + \gamma_1 q_1^*(t),$$

which is strictly negative if

$$q_2^*(t) > \lambda_2/(\theta_2 + \gamma_2) + \gamma_1 q_1^*(t)/(\theta_2 + \gamma_2).$$

For $s > \max\{\lambda_1/\mu_1 + \lambda_2/\mu_2,\ \lambda_1/\mu_1 + \lambda_2\gamma_2/((\gamma_2 + \theta_2)\mu_1)\}$, it holds that

$$q_2^*(\beta) = (s\mu_1 - \lambda_1)/\gamma_2 > \lambda_2/(\theta_2 + \gamma_2).$$

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 I(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 in this case.

Case II. $s = \frac{\lambda_1}{\mu_1} + \frac{\gamma_2}{\theta_2+\gamma_2}\frac{\lambda_2}{\mu_1} > \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2}$**.**

(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

$$q_2^*(\beta) = (s\mu_1 - \lambda_1)/\gamma_2 = \lambda_2/(\theta_2 + \gamma_2).$$

Note that the point $(0, \lambda_2/(\theta_2 + \gamma_2))$ 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. This implies that 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,

$$q_2^*(\beta) = (s\mu_1 - \lambda_1)/\gamma_2 = \lambda_2/(\theta_2 + \gamma_2).$$

Following the same reasoning as in Case II(i), since the point $(0, \lambda_2/(\theta_2 + \gamma_2))$ 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. This implies that $z_1^*(\beta+) = 0$ and

$$dq_1^*(\beta+) = \lambda_1 - \mu_1 z_1^*(\beta+) - \gamma_1 q_1^*(\beta+) + \gamma_2 q_2^*(\beta+) > 0,$$

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(i) and Case II(ii), it holds that $q_2^*(\beta) = \lambda_2/(\theta_2 + \gamma_2)$ 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 $dq_1^*(\beta+) = 0$ can be ruled out by exactly the same arguments in Case II(ii).

In the cases where β is an entry or exit point, we show that the trajectories are nontangential. Hence the adjoint vectors $p^*(t)$ are continuous at these junction times associated with Class 1. In the case where β is a contact point, we have established the continuity of the adjoint vectors $p^*(t)$ at β by either showing that the trajectories are nontangetial or using Jump condition (J) (in the case where priority is switched from Class 1 to Class 2 at β)

Case III. $\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} < s < \frac{\lambda_1}{\mu_1} + \frac{\gamma_2}{\theta_2+\gamma_2}\frac{\lambda_2}{\mu_1}$**.**

(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, $q_2^*(\beta) = (s\mu_1 - \lambda_1)/\gamma_2 < \lambda_2/(\theta_2 + \gamma_2)$. Recall that the dynamic of q_2^* follows $dq_2^*(t) = \lambda_2 - \mu_2 z_2^*(t) - (\theta_2 + \gamma_2)q_2^*(t) + \gamma_1 q_1^*(t)$. 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, $q_2^*(\beta) = (s\mu_1 - \lambda_1)/\gamma_2 < \lambda_2/(\theta_2 + \gamma_2)$. 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 $H(q^*(t), z^*(t), p^*(t)) = p_1^*(t)dq_1^*(t) + p_2^*(t)dq_2^*(t) + c_1 q_1^*(t) + c_2 q_2^*(t) > 0$ for $t \in (\beta, \beta+\delta_2)$. However, the 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, for $dq_1^*(\beta-) = 0$, we first 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(iii). Next, 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 III(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 the Hamiltonian condition (H), and thus gives a contradiction.

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

In the cases where β is an entry or exit point, we show that the trajectories are nontangential. Hence the adjoint vectors $p^*(t)$ are continuous at these junction times associated with Class 1. In the case where β is a contact point, we have established the continuity of $p^*(t)$ at β by either showing that the trajectories are nontangetial or using Jump condition (J).

Taking Cases I, II, III 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.6. 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. Thus, the multiplier η^* is equal to zero over any interior arc. Recall from (24) that

$$\zeta = \sqrt{\gamma_1^2 + 2\gamma_1(\gamma_2+\theta_1-\theta_2) + (\gamma_2-\theta_1+\theta_2)^2} = \sqrt{\gamma_2^2 + 2\gamma_2(\gamma_1+\theta_2-\theta_1) + (\gamma_1-\theta_2+\theta_1)^2}.$$

We get from (ADJ) that for $t \in (0, \alpha_j)$,

$$\begin{aligned}
p_1^*(\tau_j - t) &= \frac{c_1}{\theta_1 + \gamma_1\frac{\theta_2}{\gamma_2+\theta_2}} + \frac{c_2\frac{\gamma_1}{\gamma_1+\theta_1}}{\theta_2+\gamma_2\frac{\theta_1}{\gamma_1+\theta_1}} + e^{\frac{1}{2}(\gamma_1+\gamma_2+\theta_1+\theta_2)(\tau_j-t)}\Bigg(K_1(\tau_j)\cosh\left[\frac{1}{2}\zeta(\tau_j-t)\right] \\
&\quad + \frac{1}{\zeta}\left((\gamma_1-\gamma_2+\theta_1-\theta_2)K_1(\tau_j) - 2\gamma_1K_2(\tau_j)\right)\sinh\left[\frac{1}{2}\zeta(\tau_j-t)\right]\Bigg) \\
&= \frac{c_1}{\theta_1 + \gamma_1\frac{\theta_2}{\gamma_2+\theta_2}} + \frac{c_2\frac{\gamma_1}{\gamma_1+\theta_1}}{\theta_2+\gamma_2\frac{\theta_1}{\gamma_1+\theta_1}} + e^{\frac{1}{2}(\gamma_1+\gamma_2+\theta_1+\theta_2)(\tau_j-t)} \\
&\quad \Bigg[\frac{1}{2}\left(K_1(\tau_j) + \frac{1}{\zeta}\left((\gamma_1-\gamma_2+\theta_1-\theta_2)K_1(\tau_j) - 2\gamma_1K_2(\tau_j)\right)\right)e^{\frac{1}{2}\zeta(\tau_j-t)} \\
&\quad - \frac{1}{2}\left(K_1(\tau_j) + \frac{1}{\zeta}\left((\gamma_1-\gamma_2+\theta_1-\theta_2)K_1(\tau_j) - 2\gamma_1K_2(\tau_j)\right)\right)e^{-\frac{1}{2}\zeta(\tau_j-t)}\Bigg] \\
&= \frac{c_1}{\theta_1 + \gamma_1\frac{\theta_2}{\gamma_2+\theta_2}} + \frac{c_2\frac{\gamma_1}{\gamma_1+\theta_1}}{\theta_2+\gamma_2\frac{\theta_1}{\gamma_1+\theta_1}} \\
&\quad + \frac{1}{2}\left(K_1(\tau_j) + \frac{1}{\zeta}\left((\gamma_1-\gamma_2+\theta_1-\theta_2)K_1(\tau_j) - 2\gamma_1K_2(\tau_j)\right)\right)e^{\frac{1}{2}(\gamma_1+\gamma_2+\theta_1+\theta_2+\zeta)(\tau_j-t)} \\
&\quad - \frac{1}{2}\left(K_1(\tau_j) + \frac{1}{\zeta}\left((\gamma_1-\gamma_2+\theta_1-\theta_2)K_1(\tau_j) - 2\gamma_1K_2(\tau_j)\right)\right)e^{\frac{1}{2}(\gamma_1+\gamma_2+\theta_1+\theta_2-\zeta)(\tau_j-t)}.
\end{aligned}$$

where $K_1(\tau_j), K_2(\tau_j)$ are constants that depend on $p_1^*(\tau_j)$ and $p_2^*(\tau_j)$.

Let

$$A_1(\tau_j) := \frac{1}{2}\left(K_1(\tau_j) + \frac{1}{\zeta}\left((\gamma_1-\gamma_2+\theta_1-\theta_2)K_1(\tau_j) - 2\gamma_1K_2(\tau_j)\right)\right). \tag{25}$$

It is immediate that

$$p_1^*(\tau_j - t) = \frac{c_1}{\theta_1 + \gamma_1\frac{\theta_2}{\gamma_2+\theta_2}} + \frac{c_2\frac{\gamma_1}{\gamma_1+\theta_1}}{\theta_2+\gamma_2\frac{\theta_1}{\gamma_1+\theta_1}} + A_1(\tau_j)e^{\frac{1}{2}(\gamma_1+\gamma_2+\theta_1+\theta_2+\zeta)(\tau_j-t)} - A_1(\tau_j)e^{\frac{1}{2}(\gamma_1+\gamma_2+\theta_1+\theta_2-\zeta)(\tau_j-t)}. \tag{26}$$

By symmetry, for

$$A_2(\tau_j) := \frac{1}{2}\left(K_2(\tau_j) + \frac{1}{\zeta}\left((\gamma_2 - \gamma_1 + \theta_2 - \theta_1)K_2(\tau_j) - 2\gamma_2 K_1(\tau_j)\right)\right), \tag{27}$$

we have

$$p_2^*(\tau_j - t) = \frac{c_2}{\theta_2 + \gamma_2 \frac{\theta_1}{\gamma_1+\theta_1}} + \frac{c_1 \frac{\gamma_2}{\theta_2+\gamma_2}}{\theta_1 + \gamma_1 \frac{\theta_2}{\theta_2+\gamma_2}} + A_2(\tau_j) e^{\frac{1}{2}(\gamma_1+\gamma_2+\theta_1+\theta_2+\zeta)(\tau_j - t)} - A_2(\tau_j) e^{\frac{1}{2}(\gamma_1+\gamma_2+\theta_1+\theta_2-\zeta)(\tau_j - t)}. \tag{28}$$

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

$$\begin{aligned}\psi(\tau_j - t) = &\left(\frac{c_1}{\theta_1 + \gamma_1 \frac{\theta_2}{\gamma_2+\theta_2}} + \frac{c_2 \frac{\gamma_1}{\gamma_1+\theta_1}}{\theta_2 + \gamma_2 \frac{\theta_1}{\gamma_1+\theta_1}}\right)\mu_1 - \left(\frac{c_2}{\theta_2 + \gamma_2 \frac{\theta_1}{\gamma_1+\theta_1}} + \frac{c_1 \frac{\gamma_2}{\theta_2+\gamma_2}}{\theta_1 + \gamma_1 \frac{\theta_2}{\theta_2+\gamma_2}}\right)\mu_2 \\ &+ \left(\mu_1 A_1(\tau_j) - \mu_2 A_2(\tau_j)\right) e^{\frac{1}{2}(\gamma_1+\gamma_2+\theta_1+\theta_2+\zeta)(\tau_j - t)} \\ &- \left(\mu_1 A_1(\tau_j) - \mu_2 A_2(\tau_j)\right) e^{\frac{1}{2}(\gamma_1+\gamma_2+\theta_1+\theta_2-\zeta)(\tau_j - t)}.\end{aligned}$$

Lastly, we note that $\gamma_1 + \gamma_2 + \theta_1 + \theta_2 - \zeta > 0$. This is because under Assumption 1, at least one of θ_1 and θ_2 is strictly positive, and then,

$$\begin{aligned}\zeta &= \sqrt{\gamma_1^2 + 2\gamma_1(\gamma_2 + \theta_1 - \theta_2) + (\gamma_2 - \theta_1 + \theta_2)^2} \\ &< \sqrt{\gamma_1^2 + 2\gamma_1(\gamma_2 + \theta_1 + \theta_2) + (\gamma_2 + \theta_1 + \theta_2)^2} \\ &= \gamma_1 + \gamma_2 + \theta_1 + \theta_2.\end{aligned}$$

The statement follows from defining

$$v_1 := \frac{1}{2}(\gamma_1 + \gamma_2 + \theta_1 + \theta_2 + \zeta), \quad v_2 := \frac{1}{2}(\gamma_1 + \gamma_2 + \theta_1 + \theta_2 - \zeta).$$

Q.E.D.

B.7. 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,

$$(c_1\mu_1 - c_2\mu_2)(r_1 - r_2) > 0, \quad \text{for } r_1, r_2 \text{ in (6) and (7).}$$

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 $\alpha_N < \tau_N$,

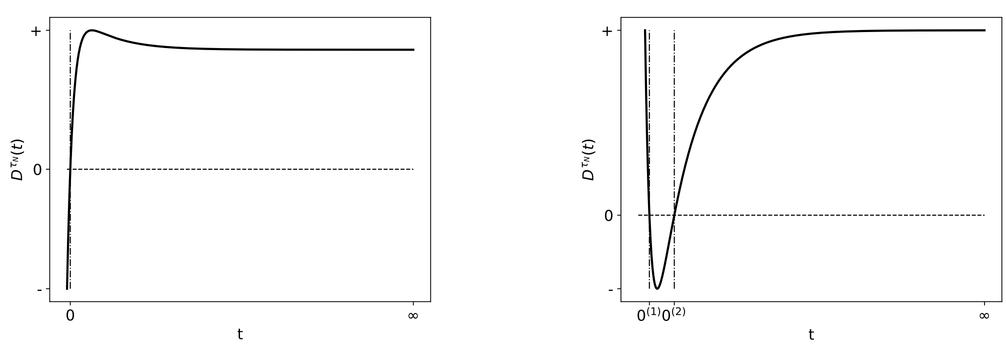
$$\begin{aligned}\psi(\tau_N - t) = r_1 - r_2 &+ \left(\mu_1 A_1(\tau_N) - \mu_2 A_2(\tau_N)\right) e^{\frac{1}{2}(\gamma_1+\gamma_2+\theta_1+\theta_2+\zeta)(\tau_N - t)} \\ &- \left(\mu_1 A_1(\tau_N) - \mu_2 A_2(\tau_N)\right) e^{\frac{1}{2}(\gamma_1+\gamma_2+\theta_1+\theta_2-\zeta)(\tau_N - t)}, \quad t \in (0, \tau_N - \alpha_N),\end{aligned}$$

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

$$\lim_{t\to\infty} D^{\tau_N}(t) = r_1 - r_2 > 0.$$

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 16, with one root in Figure 16(a) and two roots in Figure 16(b). Figure 16 is comprehensive in the sense that any $D^{\tau_N}(t)$ function shares the same behavior in crossing zeros and in the limit as $t \to \infty$. In particular, if $D^{\tau_N}(t)$ has one zero as in Figure 16(a), then it must be that $D^{\tau_N}(t)$ is increasing at the zero point and eventually converges to $r_1 - r_2$. 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 16(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 $r_1 - r_2$. 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 16 **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 16, if $D^{\tau_N}(t)$ has one zero (Figure 16(a)), then $D^{\tau_N}(0)$ is at this unique zero point. If $D^{\tau_N}(t)$ has two zeros (Figure 16(b)), then $D^{\tau_N}(0)$ 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 $\delta>0$ (which can be 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 $0<\beta\le t$. (Naturally, $\beta\le\alpha_N$.) Now, suppose there exists such $\beta>0$, i.e., $\eta_2^*(\tau_N-\beta)>0$. Note that $\eta_1^*(\tau_N-t)=0$ and $\eta_2^*(\tau_N-t)=0$ for all $t\in[0,\beta)$. As $D^{\tau_N}(\beta)>\delta'$ for some $\delta'>0$ and $\eta_2^*(\tau_N-\beta)>0$, it follows that $\psi(\tau_N-\beta)\ge\delta'>0$. 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., $\psi((\tau_N-\beta)-)\le 0$. However, due to the continuity of the switching curve, this contradicts the fact that $\psi(\tau_N-\beta)\ge\delta'>0$. 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,

$$(c_1\mu_1-c_2\mu_2)(r_1-r_2)<0,\quad \text{for } r_1,r_2 \text{ in (6) and (7).}$$

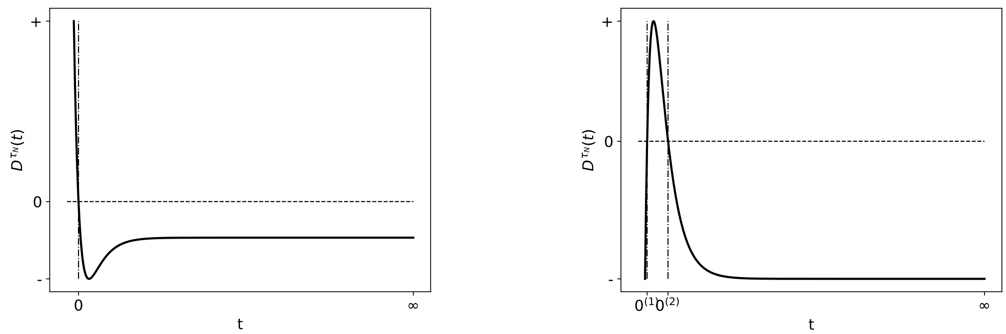
For the moment, suppose Class 1 has a higher $c\mu$-index and Class 2 has a higher modified $c\mu/\theta$-index. Following similar lines of arguments as in Case I, the backward switching curve $\psi(\tau_N-t)$ follows the dynamic of $D^{\tau_N}(t)$ for some non-trivial time interval $t\in(0,\alpha_N)$. 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 17, with Figure 17(a) crossing zero once and Figure 17(b) crossing zero twice. In particular, if $D^{\tau_N}(t)$ has one zero as in Figure 17(a), then it must be that $D^{\tau_N}(t)$ is decreasing at the zero point and eventually converges to $r_1-r_2<0$. 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 17(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 r_1-r_2. Once $D^{\tau_N}(t)$ crosses the second zero point, it will never increase to zero again.

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 17(a). $D_{\tau_N}(0)$ must be at the first zero in Figure 17(b). Now, let time $\beta>0$ denote the second zero in Figure 17(b), i.e., $D^{\tau_N}(\beta)=0$. Then, one of the following three scenarios holds.

Scenario 1. $\tau_N\le\beta$. 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

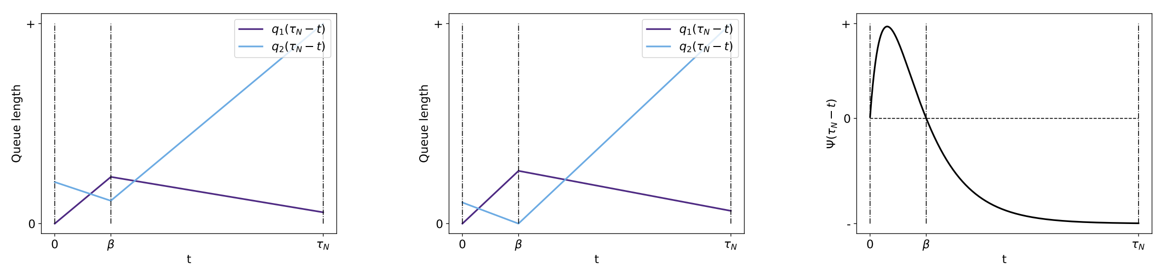
Figure 17 **Possible trajectory of $D^{\tau_1}(t)$ with $c_1\mu_1 > c_2\mu_2$, modified $c_1\mu_1/\theta_1 <$ modified $c_2\mu_2/\theta_2$**



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

(backward in time). In this scenario, we consider the cases where either both queues are strictly positive at $t = \beta$ as in Figure 18(a), or β is a contact point as in Figure 18(b). In either cases, 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 18(c). In this case, the optimal control (forward in time) switches priority once from Class 2 to Class 1.

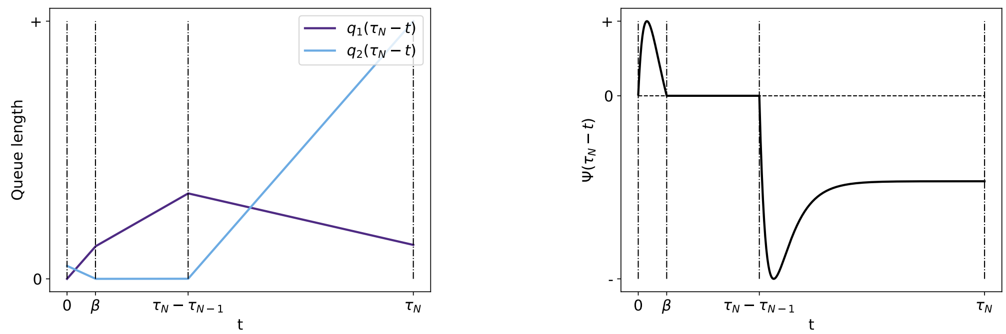
Figure 18 **Backward state trajectory and switching curve in Scenario 2**



(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 Scenario 2, β is an exit point (forward in time) for the trajectory of q_2^*; see Figure 19(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 for $t \in (0, \tau_{N-1})$. 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 19(b).

Figure 19 Backward state trajectory and switching curve in Scenario 3



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

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

B.8. Proof of Proposition 2

Proof: First, as shown in Proposition 3, 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.

Next, consider the case where the $c\mu$-rule and the modified $c\mu/\theta$-rule prioritize different classes, namely,

$$(c_1\mu_1 - c_2\mu_2)(r_1 - r_2) < 0, \quad \text{for } r_1, r_2 \text{ in (6) and (7).}$$

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, the work left is to prove that 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.

For the moment, we consider the case where the $c\mu$-rule prioritizes Class 1 and the modified $c\mu/\theta$-rule prioritizes Class 2. We first note that by the definition of τ_1, both queues are strictly positive for $t < \tau_1$. Thus, the multipliers $\eta_1^*(t) = \eta_2^*(t) = 0$ for $t < \tau_1$. By Lemma 4, the backward switching curve before τ_1 is characterized as follows

$$\begin{aligned}\psi(\tau_1 - t) = r_1 - r_2 &+ (\mu_1 A_1(\tau_1) - \mu_2 A_2(\tau_1))\, e^{\frac{1}{2}(\gamma_1+\gamma_2+\theta_1+\theta_2+\zeta)(\tau_1 - t)} \\ &- (\mu_1 A_1(\tau_1) - \mu_2 A_2(\tau_1))\, e^{\frac{1}{2}(\gamma_1+\gamma_2+\theta_1+\theta_2-\zeta)(\tau_1 - t)},\end{aligned} \tag{29}$$

where $A_1(\tau_1), A_2(\tau_1)$ are constants in $\mathbb{R}$, and $0 \le \zeta < \gamma_1 + \gamma_2 + \theta_1 + \theta_2$.

Note that due to class-transition, when one queue gets emptied, the other queue cannot be arbitrarily large. In the case where Class 1 gets emptied at τ_1, it holds that $q_1^*(\tau_1) = 0$, and for any $\epsilon > 0$,

$$q_2^*(\tau_1) < (s\mu_1 - \lambda_1)/\gamma_2 + \epsilon.$$

Similarly, in the other case where Class 2 gets emptied at τ_1, it holds that $q_2^*(\tau_1) = 0$ and for any $\epsilon > 0$, we have $q_1^*(\tau_1) < (s\mu_2 - \lambda_2)/\gamma_1 + \epsilon$. Since $q_1^*(\tau_1)$ and $q_2^*(\tau_1)$ are uniformly 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.

Now, from the form of $A_1(\tau_1)$ and $A_2(\tau_1)$ in the proof of Lemma 4, in particular, (25) and (27), we see that $A_1(\tau_1)$ and $A_2(\tau_1)$ are bounded if $p_1^*(\tau_1)$ and $p_2^*(\tau_1)$ are bounded, uniformly for any initialization.

Lastly, 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 (29), the sign of the backward switching curve will eventually be governed by $r_1 - r_2$. In other words, for M sufficiently large, the modified $c\mu/\theta$-rule is optimal at time t if $q_1(t) + q_2(t) > M$.

The arguments for the other case where the $c\mu$-rule prioritizes Class 2 and the modified $c\mu/\theta$-rule prioritizes Class 1 follow by symmetry. Q.E.D.

B.9. Proof of Theorem 2

Proof: The statement of Theorem 2 follows directly from Propositions 1, 2, and 3. Q.E.D.

B.10. Proof of Proposition 4

Proof: For $c_1\mu_1 < c_2\mu_2$ and $r_1 > r_2$, Theorem 2 indicates that a one-time switch in priority from Class 1 to Class 2 will take place if the system is initialized far enough from the origin. To derive the policy curve at which (state) the switching takes place, we apply the 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.

Starting from time t_1, the dynamic of the adjoint vector for $p^*(t)$ is specified by (ADJ) as

$$\begin{aligned}
p_1^*(t) &= K_1(t_1)e^{t\theta_1} + e^{t\theta_1}\int_0^t e^{-s\theta_1}\left(-c_1 + \eta_1^*(s)\right)ds \\
p_2^*(t) &= K_2(t_1)e^{t(\theta_2+\gamma_2)} + \frac{K_1\gamma_2}{\gamma_2 - \theta_1 + \theta_2}\left(e^{t\theta_1} - e^{t(\theta_2+\gamma_2)}\right) + \frac{\gamma_2}{\gamma_2 - \theta_1 + \theta_2}e^{t\theta_1}\int_0^t e^{-s\theta_1}\left(-c_1 + \eta_1^*(s)\right)ds \\
&\quad - \frac{\gamma_2}{\gamma_2 - \theta_1 + \theta_2}e^{t(\theta_2+\gamma_2)}\int_0^t e^{s(-\gamma_2-\theta_2)}\left(-c_1 + \eta_1^*(s)\right)ds + e^{t(\theta_2+\gamma_2)}\int_0^t e^{s(-\gamma_2-\theta_2)}\left(-c_2 + \eta_2^*(s)\right)ds,
\end{aligned} \tag{30}$$

where $K_1(t_1)$ and $K_2(t_1)$ are constants that depends on $p_1^*(t_1)$ and $p_2^*(t_1)$. Since there is no other switch in priority (Proposition 3) after t_1, $q_1(t) > 0$ for $t \in (t_1, t_3)$, and $q_2(t) > 0$ for $t \in (t_1, t_2)$. Then, (30) reduces to

$$\begin{aligned}
p_1^*(t) &= \frac{c_1}{\theta_1} + e^{\theta_1 t}K_1(t_1) \quad \text{for } t \in [t_1, t_3] \\
p_2^*(t) &= \frac{c_2}{\theta_2+\gamma_2} + \frac{c_1\gamma_2}{\theta_1(\theta_2+\gamma_2)} + \frac{e^{\theta_1 t}\gamma_2 K_1(t_1) + e^{(\theta_2+\gamma_2)t}(-\gamma_2 K_1(t_1) + (\gamma_2 - \theta_1 + \theta_2)K_2(t_1))}{\gamma_2 - \theta_j + \theta_2} \quad \text{for } t \in [t_1, t_2].
\end{aligned} \tag{31}$$

The rest of the analysis is divided into three time 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 the Hamiltonian condition (H), we are able to characterize the constants $K_1(t_1), K_2(t_1)$ in (31) as well as the policy curve. These steps will become self-explanatory as the proof proceeds.

Case I. 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.

$$\begin{aligned}
q_1^*(t_1-) &= a_1\\
q_2^*(t_1-) &= a_2\\
H(q^*(t_1-), z^*(t_1-), p^*(t_1-)) &= c_1a_1 + c_2a_2 + (a_2\gamma_2 - a_1\theta_1 + \lambda_1 - s\mu_1)\left(\frac{c_1}{\theta_1} + K_1(t_1)\right)\\
&\quad + (-a_2(\theta_2+\gamma_2) + \lambda_2)\left(\frac{c_1\gamma_2 + c_2\theta_1}{\gamma_2\theta_1 + \theta_1\theta_2} + K_2(t_1)\right).
\end{aligned}$$

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

$$\begin{aligned}
q_1^*(t) = &-\frac{1}{\theta_1(\theta_2+\gamma_2)(\gamma_2-\theta_1+\theta_2)} e^{-(t-t_1)(\gamma_2+\theta_1+\theta_2)}\Big(e^{(t-t_1)\theta_1}\gamma_2\theta_1(a_2(\theta_2+\gamma_2) - \lambda_2 + s\mu_2)\\
&- e^{(t-t_1)(\theta_2+\gamma_2)}(\theta_2+\gamma_2)(a_2\gamma_2\theta_1 + a_1\theta_1(\gamma_2-\theta_1+\theta_2) - \gamma_2\lambda_1 + \theta_1\lambda_1 - \theta_2\lambda_1 - \gamma_2\lambda_2 + s\gamma_2\mu_2)\\
&- e^{(t-t_1)(\gamma_2+\theta_1+\theta_2)}(\gamma_2-\theta_1+\theta_2)(\theta_2\lambda_1 + \gamma_2(\lambda_1+\lambda_2-s\mu_2))\Big),\\
q_2^*(t) = &\frac{1}{\theta_2+\gamma_2} e^{-(t-t_1)(\theta_2+\gamma_2)}\left(a_2(\theta_2+\gamma_2) + (-1+e^{(t-t_1)(\theta_2+\gamma_2)})(\lambda_2 - s\mu_2)\right),\\
t_2 - t_1 = &\frac{1}{\theta_2+\gamma_2}\log\left(\frac{-a_2\gamma_2 - a_2\theta_2 + \lambda_2 - s\mu_2}{\lambda_2 - s\mu_2}\right)\\
H(q^*(t), z^*(t), p^*(t)) = &\frac{1}{\theta_1(\theta_2+\gamma_2)}\big\{c_1\theta_2\lambda_1 + c_2\theta_1(\lambda_2 - s\mu_2) + c_1\gamma_2(\lambda_1+\lambda_2-s\mu_2)\\
&- \theta_1(\theta_2+\gamma_2)\big[a_1\theta_1K_1(t_1) - \lambda_1K_1(t_1) + a_2\theta_2K_2(t_1) - \lambda_2K_2(t_1)\\
&+ s\mu_2K_2(t_1) + a_2\gamma_2(-K_1(t_1) + K_2(t_1))\big]\big\}.
\end{aligned}$$

Putting the analysis for Interval 1 and Interval 2 together, we can solve for $K_1(t_1)$ and $K_2(t_1)$ from the system of equations $H(q^*(t_1), z^*(t_1), p^*(t_1)) = 0$ and $H(q^*(t), z^*(t), p^*(t)) = 0$ for $t \in [t_1, t_2)$. In particular,

$$\begin{aligned}
K_1(t_1) &= \frac{c_1(-a_2(\theta_2+\gamma_2)+\lambda_2)\mu_1 + c_2a_2\theta_1\mu_2 + c_1(a_2\gamma_2 + \lambda_1 - s\mu_1)\mu_2}{\theta_1(a_2(\theta_2+\gamma_2) - \lambda_2)\mu_1 + \theta_1(-a_2\gamma_2 + a_1\theta_1 - \lambda_1 + s\mu_1)\mu_2}\\
K_2(t_1) &= -\frac{c_1\gamma_2 + c_2\theta_1}{\gamma_2\theta_1 + \theta_1\theta_2} + \frac{(c_1a_1 + c_2a_2)\mu_1}{a_2(\theta_2+\gamma_2)\mu_1 - \lambda_2\mu_1 - a_2\gamma_2\mu_2 + (a_1\theta_1 - \lambda_1 + s\mu_1)\mu_2}.
\end{aligned} \tag{32}$$

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

$$\begin{aligned}
q_1^*(t) = &\frac{e^{-(t-t_2)\theta_1}}{\theta_1(-\gamma_2+\theta_1-\theta_2)}\Big(-(a_2\gamma_2\theta_1 + a_1\theta_1(\gamma_2-\theta_1+\theta_2) - \gamma_2\lambda_1 + \theta_1\lambda_1 - \theta_2\lambda_1 - \gamma_2\lambda_2 + s\gamma_2\mu_2)\\
&\left(1 + \frac{a_2(\theta_2+\gamma_2)}{-\lambda_2 + s\mu_2}\right)^{-\frac{\theta_1}{\theta_2+\gamma_2}} - \frac{1}{\mu_2}\big((\lambda_2 - s\mu_2)(-(\gamma_2-\theta_1+\theta_2)\mu_1 + \gamma_2\mu_2)\\
&+ e^{(t-t_2)\theta_1}(\gamma_2-\theta_1+\theta_2)(\lambda_2\mu_1 + (\lambda_1 - s\mu_1)\mu_2)\big)\Big),\\
q_2^*(t) = &\,0,\\
t_3 - t_2 = &\frac{1}{\theta_1}\log\left(\frac{1}{(\gamma_2-\theta_1+\theta_2)(\lambda_2\mu_1 + (\lambda_1 - s\mu_1)\mu_2)}\Big((\lambda_2 - s\mu_2)((\gamma_2-\theta_1+\theta_2)\mu_1 - \gamma_2\mu_2)\right.\\
&- \mu_2\left(a_2\gamma_2\theta_1 + a_1\theta_1(\gamma_2-\theta_1+\theta_2) - \gamma_2\lambda_1 + \theta_1\lambda_1 - \theta_2\lambda_1 - \gamma_2\lambda_2 + s\gamma_2\mu_2\right)\\
&\left.\left(1 + \frac{a_2(\theta_2+\gamma_2)}{-\lambda_2 + s\mu_2}\right)^{-\frac{\theta_1}{\theta_2+\gamma_2}}\Big)\right).
\end{aligned}$$

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

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

Then, plugging the expression of $q_1^*(t)$, (31), into $H(q^*(t),z^*(t),p^*(t))$, we get

$$\begin{aligned}H(q^*(t),z^*(t),p^*(t)) = {} & \frac{K_1(t_1)}{\mu_2(\gamma_2-\theta_1+\theta_2)}\Bigg(+ (\lambda_2 - s\mu_2)((\gamma_2-\theta_1+\theta_2)\mu_1 - \gamma_2\mu_2)\left(1+\frac{a_2(\theta_2+\gamma_2)}{-\lambda_2+s\mu_2}\right)^{\frac{\theta_1}{\theta_2+\gamma_2}} \\ & - \mu_2\big(a_2\gamma_2\theta_1 + a_1\theta_1(\gamma_2-\theta_1+\theta_2) - \gamma_2\lambda_1 + \theta_1\lambda_1 - \theta_2\lambda_1 - \gamma_2\lambda_2 + s\gamma_2\mu_2\big)\Bigg) \\ & + \frac{c_1(\lambda_2\mu_1 + (\lambda_1 - s\mu_1)\mu_2)}{\theta_1\mu_2}.\end{aligned}$$

Plugging the value of $K_1(t_1)$, (32), into the equality $H(q^*(t),z^*(t),p^*(t))=0$ for $t\in[t_2,t_3)$ establishes the relationship (a_1,a_2) must satisfy. This gives the policy curve in Proposition 4.

Case II. 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 $(\lambda_1+\gamma_2q_2^*(t_1-))/\mu_1$ servers to Class 1 and the rest of the servers to Class 2.

$$\begin{aligned}q_1^*(t_1-) &= 0 \\ q_2^*(t_1-) &= a_2 \\ H(q^*(t_1-),z^*(t_1-),p^*(t_1-)) &= c_2a_2 + \left(-a_2(\gamma_2+\theta_2) + \lambda_2 - s\mu_2 + \frac{(a_2\gamma_2+\lambda_1)\mu_2}{\mu_1}\right)\left(\frac{c_1\gamma_2+c_2\theta_1}{\gamma_2\theta_1+\theta_1\theta_2} + K_2(t_1)\right)\end{aligned}$$

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

$$\begin{aligned}q_1^*(t) = {} & -\frac{1}{\theta_1(\theta_2+\gamma_2)(\gamma_2-\theta_1+\theta_2)}e^{-(t-t_1)(\gamma_2+\theta_1+\theta_2)}\Big(e^{(t-t_1)\theta_1}\gamma_2\theta_1(a_2(\theta_2+\gamma_2)-\lambda_2+s\mu_2) \\ & - e^{(t-t_1)(\theta_2+\gamma_2)}(\theta_2+\gamma_2)(a_2\gamma_2\theta_1 - \gamma_2\lambda_1 + \theta_1\lambda_1 - \theta_2\lambda_1 - \gamma_2\lambda_2 + s\gamma_2\mu_2) \\ & - e^{(t-t_1)(\gamma_2+\theta_1+\theta_2)}(\gamma_2-\theta_1+\theta_2)(\theta_2\lambda_1 + \gamma_2(\lambda_1+\lambda_2-s\mu_2))\Big), \\ q_2^*(t) = {} & \frac{1}{\theta_2+\gamma_2}e^{-(t-t_1)(\theta_2+\gamma_2)}\left(a_2(\theta_2+\gamma_2) + (-1+e^{(t-t_1)(\theta_2+\gamma_2)})(\lambda_2-s\mu_2)\right), \\ t_2 - t_1 = {} & \frac{1}{\theta_2+\gamma_2}\log\left(\frac{-a_2\gamma_2 - a_2\theta_2 + \lambda_2 - s\mu_2}{\lambda_2 - s\mu_2}\right) \\ H(q^*(t),z^*(t),p^*(t)) = {} & \frac{1}{\theta_1(\theta_2+\gamma_2)}\big\{c_1\theta_2\lambda_1 + c_2\theta_1(\lambda_2-s\mu_2) + c_1\gamma_2(\lambda_1+\lambda_2-s\mu_2) \\ & - \theta_1(\theta_2+\gamma_2)\big[-\lambda_1K_1(t_1) + a_2\theta_2K_2(t_1) - \lambda_2K_2(t_1) \\ & + s\mu_2K_2(t_1) + a_2\gamma_2(-K_1(t_1)+K_2(t_1))\big]\big\}.\end{aligned}$$

Putting the analysis for Interval 1 and Interval 2 together, we can solve for $K_1(t_1)$ and $K_2(t_1)$ from the system of equations $H(q^*(t_1),z^*(t_1),p^*(t_1))=0$ and $H(q^*(t),z^*(t),p^*(t))=0$ for $t\in[t_1,t_2)$. In particular, we get

$$\begin{aligned}K_1(t_1) &= \frac{c_1(-a_2(\theta_2+\gamma_2)+\lambda_2)\mu_1 + c_2a_2\theta_1\mu_2 + c_1(a_2\gamma_2+\lambda_1-s\mu_1)\mu_2}{\theta_1(a_2(\theta_2+\gamma_2)-\lambda_2)\mu_1 + \theta_1(-a_2\gamma_2-\lambda_1+s\mu_1)\mu_2} \\ K_2(t_1) &= -\frac{c_1\gamma_2+c_2\theta_1}{\gamma_2\theta_1+\theta_1\theta_2} + \frac{c_2a_2\mu_1}{a_2(\theta_2+\gamma_2)\mu_1 - \lambda_2\mu_1 - a_2\gamma_2\mu_2 + (-\lambda_1+s\mu_1)\mu_2}.\end{aligned} \tag{33}$$

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

$$\begin{aligned}
q_1^*(t) = & \frac{e^{-(t-t_2)\theta_1}}{\theta_1(-\gamma_2+\theta_1-\theta_2)}\Bigg(-(a_2\gamma_2\theta_1 - \gamma_2\lambda_1 + \theta_1\lambda_1 - \theta_2\lambda_1 - \gamma_2\lambda_2 + s\gamma_2\mu_2) \\
& \left(1+\frac{a_2(\theta_2+\gamma_2)}{-\lambda_2+s\mu_2}\right)^{-\frac{\theta_1}{\theta_2+\gamma_2}} - \frac{1}{\mu_2}\big((\lambda_2-s\mu_2)(-(\gamma_2-\theta_1+\theta_2)\mu_1+\gamma_2\mu_2) \\
& + e^{(t-t_2)\theta_1}(\gamma_2-\theta_1+\theta_2)(\lambda_2\mu_1+(\lambda_1-s\mu_1)\mu_2)\big)\Bigg), \\
q_2^*(t) = & \, 0, \\
t_3 - t_2 = & \frac{1}{\theta_1}\log\Bigg(\frac{1}{(\gamma_2-\theta_1+\theta_2)(\lambda_2\mu_1+(\lambda_1-s\mu_1)\mu_2)}\Bigg((\lambda_2-s\mu_2)((\gamma_2-\theta_1+\theta_2)\mu_1-\gamma_2\mu_2) \\
& - \mu_2\left(a_2\gamma_2\theta_1 - \gamma_2\lambda_1 + \theta_1\lambda_1 - \theta_2\lambda_1 - \gamma_2\lambda_2 + s\gamma_2\mu_2\right)\left(1+\frac{a_2(\theta_2+\gamma_2)}{-\lambda_2+s\mu_2}\right)^{-\frac{\theta_1}{\theta_2+\gamma_2}}\Bigg)\Bigg).
\end{aligned}$$

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

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

Then, plugging the expression of $q_1^*(t)$, (31) into $H(q^*(t), z^*(t), p^*(t))$, we get

$$\begin{aligned}
H(q^*(t), z^*(t), p^*(t)) = & \frac{K_1(t_1)}{\mu_2(\gamma_2-\theta_1+\theta_2)}\Bigg(-\mu_2\big(a_2\gamma_2\theta_1 - \gamma_2\lambda_1 + \theta_1\lambda_1 - \theta_2\lambda_1 - \gamma_2\lambda_2 + s\gamma_2\mu_2\big) \\
& + (\lambda_2-s\mu_2)((\gamma_2-\theta_1+\theta_2)\mu_1-\gamma_2\mu_2)\left(1+\frac{a_2(\theta_2+\gamma_2)}{-\lambda_2+s\mu_2}\right)^{\frac{\theta_1}{\theta_2+\gamma_2}}\Bigg) \\
& + \frac{c_1(\lambda_2\mu_1+(\lambda_1-s\mu_1)\mu_2)}{\theta_1\mu_2}.
\end{aligned}$$

Plugging the value of $K_1(t_1)$, (33), into the equality $H(q^*(t), z^*(t), p^*(t)) = 0$ for $t \in [t_2, t_3)$ 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 $H(q^*(t), z^*(t), p^*(t)) = 0$ 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 verifies this. Q.E.D.

Appendix C: Proof of Theorem 3

Proof: We dissect the transient optimization problem over the entire time horizon $[0, T+\tau^*]$ into a two-stage optimal control problem. The first-stage problem (12) is over the time interval $[0, T)$. The second problem (13) is over the time interval $[T, T+\tau^*]$ and its initial condition is equal to the terminal state in problem (12). We also note that (13) over $[T, T+\tau^*]$ is equivalent to (F2′) over $[0, \tau^*]$ with the appropriate initial condition. In what follows, to distinguish problems (12) and (F2′), we will append superscripts [1] and [2] to the queue length processes, dual vectors, etc., associated with problems (12) and (F2′), respectively.

For example, we will write the time horizon for (12) as $[0^{[1]}, T^{[1]})$ and the time horizon for (F2′) as $[0^{[2]}, \tau^{*[2]}]$, where $0^{[1]} = 0, T^{[1]} = T, 0^{[2]} = 0$, and $\tau^{*[2]} = \tau^*$.

We first note that for the second-stage problem (F2′) over $[0^{[2]}, \tau^{*[2]}]$, it follows directly from Theorem 2 that the optimal scheduling policy follows the $c\mu$-rule when the states are sufficiently small, and follows the modified $c\mu/\theta$-rule when the states are sufficiently large. The work left is to show that the optimal scheduling policy switches priority at most once over the entire transient time horizon $[0, T+\tau^*]$. To do this, we establish an analogous version of Proposition 3 below.

Claim A. Under Assumptions 1 and 4, for the transient optimal control problem (12) and (F2′), if the $c\mu$-rule and the modified $c\mu/\theta$-rule prioritize the same class, then the optimal transient scheduling policy does not switch priority. If the two index rules prioritize different classes, then the optimal transient scheduling policy switches priority at most once over the transient time horizon $[0, T+\tau^*]$.

To establish Claim A, we observe that problem (12) over the initial period $[0^{[1]}, T^{[1]})$ is an optimal control problem with fixed time, free terminal state, terminal cost, and no state constraints. For this type of problems, the following version of Pontryagin's Minimum Principle applies.

Lemma 5 (Theorem 3.4 in Grass et al. (2008)) *Under Assumption 4, let $z^{*[1]}$ be an optimal solution to* (12)*, and $q^{*[1]}$ be the corresponding state trajectory. There exists a continuous and piecewise continuously differentiable adjoint vector $p^{*[1]} : [0^{[1]}, T^{[1]}] \to \mathbb{R}^2_+$ satisfying for all $t \in [0^{[1]}, T^{[1]}]$:*

1. *Ordinary Differential Equation condition (ODE):*

$$q^{*[1]}(0) = q_0, \quad dq^{*[1]}(t) = f^{[1]}\left(q^{*[1]}(t), z^{*[1]}(t), t\right)$$

2. *Adjoint Vector condition (ADJ):*

$$dp^{*[1]}(t) = -\nabla_q H^{[1]}(q^{*[1]}(t), z^{*[1]}(t), p^{*[1]}(t), t)$$

3. *Minimization condition (M):*

$$H^{[1]}(q^{*[1]}(t), z^{*[1]}(t), p^{*[1]}(t), t) = \min_z \{H^{[1]}(q^{*[1]}(t), z^{[1]}(t), p^{*[1]}(t), t)\}$$

4. *Transversality condition (T):*

$$p^{*[1]}(T^{[1]}) = \nabla_q \Xi(q^{*[1]}(T^{[1]})). \tag{34}$$

Note that as we allow for time-varying arrival rates on $[0^{[1]}, T^{[1]}]$, $f^{[1]}$ and $H^{[1]}$ have an explicit time component. We draw several connections between the two versions of Pontryagin's Minimum Principles in Lemma 5 and Theorem 5. First, the construction of the Hamiltonian and conditions (ODE) and (M) are essentially the same in the two versions, except that fluid dynamic $f^{[1]}$ and the Hamiltonian $H^{[1]}$ in Lemma 5 are time-dependent through $\lambda^{[1]}(t)$. Second, Minimization condition (M) in Lemma 5 specifies the dynamics of the adjoint vector for problem (12)

$$dp_1^{*[1]}(t) = (\theta_1 + \gamma_1)p_1^{*[1]}(t) - \gamma_1 p_2^{*[1]}(t) - c_1, \quad dp_2^{*[1]}(t) = (\theta_2 + \gamma_2)p_2^{*[1]}(t) - \gamma_2 p_1^{*[1]}(t) - c_2, \tag{35}$$

while the Minimization condition (M) in Theorem 5 gives that for problem (F2′)

$$dp_1^{*[2]}(t) = (\theta_1+\gamma_1)p_1^{*[2]}(t) - \gamma_1 p_2^{*[2]}(t) - c_1 + \eta_1^{*[2]}(t),\ dp_2^{*[2]}(t) = (\theta_2+\gamma_2)p_2^{*[2]}(t) - \gamma_2 p_1^{*[2]}(t) - c_2 + \eta_2^{*[2]}(t). \tag{36}$$

Comparing (35) with (36), we note that the adjoint vectors for problems (12) and (F2′) follow the same dynamic when the state constraints in (F2′) are not active, namely, when both queues are strictly positive. Third, Transversality condition (34) holds exclusively for the first-stage problem (12) which has fixed terminal time and no terminal (state) constraint.

To this end, for the second-stage problem (F2′), let $\tau_1^{[2]}$ denote the first time one of the two queues hits zero. The pseudo switching curve associated with $\tau_1^{[2]}$ is given by

$$\begin{aligned} D^{\tau_1^{[2]}}(t) = r_1 - r_2 &+ \left(\mu_1 A_1^{[2]}(\tau_1^{[2]}) - \mu_2 A_2^{[2]}(\tau_1^{[2]})\right) e^{\frac{1}{2}(\gamma_1+\gamma_2+\theta_1+\theta_2+\zeta)(\tau_1^{[2]}-t)} \\ &- \left(\mu_1 A_1^{[2]}(\tau_1^{[2]}) - \mu_2 A_2^{[2]}(\tau_1^{[2]})\right) e^{\frac{1}{2}(\gamma_1+\gamma_2+\theta_1+\theta_2-\zeta)(\tau_1^{[2]}-t)}, \quad \text{for all } t \geq 0. \end{aligned}$$

Since both queues are strictly positive for $t \in [0^{[2]}, \tau_1^{[2]})$, the multiplies $\eta_1^{*[2]}(t) = \eta_2^{*[2]}(t) = 0$ for $t \in [0^{[2]}, \tau_1^{[2]})$. It follows that the switching curve for problem (F2′) backward from time $\tau_1^{[2]}$ agrees with $D^{\tau_1^{[2]}}(t)$, namely,

$$\psi^{[2]}(\tau_1^{[2]} - t) = D^{\tau_1^{[2]}}(t) \quad \text{for all } t \in (0^{[2]}, \tau_1^{[2]}].$$

Now, recall that $\Xi(q_0)$ is the value function from state q_0 in the second-stage problem (F2′). Thus, it follows from Transversality condition (34) in Lemma 5 that

$$p^{*[1]}(T^{[1]}) = \nabla_q \Xi(q^{*[1]}(T^{[1]})) = \nabla_q \Xi(q^{*[2]}(0^{[2]})) = p^{*[2]}(0^{[2]}). \tag{37}$$

By (37), together with the fact that the adjoint vectors for problems (12) and (F2′) follow the same dynamic when both queues are strictly positive, it is easy to see that the backward switching curve for the first-stage problem $\psi^{[1]}$ is connected to the pseudo switching curve for the second-stage problem $D^{\tau_1^{[2]}}$ via

$$\psi^{[1]}(T^{[1]} - t) = D^{\tau_1^{[2]}}(\tau_1^{[2]} + t), \quad \text{for all } t \in [0^{[1]}, T^{[1]}]. \tag{38}$$

It follows from (38) that analyzing the first-stage backward switching curve $\psi^{[1]}(T^{[1]} - t)$ is equivalent to analyzing the second-stage pseudo switching curve $D^{\tau_1^{[2]}}(\tau_1^{[2]} + t)$ extended beyond the beginning epoch of the second-stage problem for another $T^{[1]}$ time units. It is then straightforward to see that the arguments in the proof of Proposition 3 extend to the first-stage problem (12) and Claim A follows. Q.E.D.

Appendix D: The Special Cases with No Class-Transition and Abandonment

The special case where $\gamma_1 = \gamma_2 = \theta_1 = \theta_2 = 0$ is not covered in Theorem 2, as Assumption 1 does not hold in this case. However, the same lines of argument, utilizing the Pontryagin's Minimum Principle, can be use in this case to establish the optimality of the $c\mu$-rule. Indeed, the proof is more concise here and nicely illustrates the main idea behind our proof strategy.

Corollary 2 *If $\gamma_1 = \gamma_2 = \theta_1 = \theta_2 = 0$, and $s > \lambda_1/\mu_1 + \lambda_2/\lambda_2$, the $c\mu$-rule is optimal for the transient fluid optimal control problem (F2′).*

Proof: 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

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

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

$$dp_1^*(t) = -c_1 + \eta_1^*(t) \quad \text{and} \quad dp_2^*(t) = -c_2 + \eta_2^*(t). \tag{39}$$

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) when $q_1^*(t)$ hits zero, i.e.,

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

Following the same lines of 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

$$\psi(\tau_N - t) = \psi(\tau_N) + (c_1\mu_1 - c_2\mu_2)t + \left(\mu_2 \int_{\tau_N - t}^{\tau_N} \eta_2^*(s) ds - \mu_1 \int_{\tau_N - t}^{\tau_N} \eta_1^*(s) ds\right) = (c_1\mu_1 - c_2\mu_2)t. \tag{40}$$

Since $c_1\mu_1 > c_2\mu_2$, the significance of (40) 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 (40) indeed holds for all $t \in [0, \tau_N]$. Namely, strict priority to Class 1 is optimal throughout $[0, \tau^*]$. Q.E.D.

D.1. Full Characterization of the Dual Vectors When $\gamma_1 = \gamma_2 = \theta_1 = \theta_2 = 0$

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 class-transition. On the other hand, for simplified systems without abandonment or class-transition, we can provide a full characterization of the dual vectors. We next illustrate the derivation.

By Corollary 2, the $c\mu$-rule is optimal at all time for systems without abandonment and without class-transition. 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

$$\begin{aligned} p_1^*(t) &= \frac{1}{\lambda_1 - s\mu_1}\left(-c_1 q_1^*(t) + \frac{c_2(-q_2^*(t)\lambda_1 + q_1^*(t)\lambda_2 + sq_2^*(t)\mu_1)\mu_2}{\lambda_2\mu_1 + (\lambda_1 - s\mu_1)\mu_2}\right), \quad t \in [0, \tau^*] \\ p_2^*(t) &= \frac{c_2(q_2^*(t)\mu_1 + q_1^*(t)\mu_2)}{-\lambda_2\mu_1 - \lambda_1\mu_2 + s\mu_1\mu_2}, \quad t \in [0, \tau^*]. \end{aligned} \tag{41}$$

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 (41).

In addition, it follows from (39) 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,

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

$$\eta_2^* = 0, \quad t \in [0, \tau^*].$$

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

$$\xi_1^*(t) = \mu_1 p_1^*(t), \quad t \in [0, \tau^*]$$

$$\xi_2^*(t) = 0, \quad t \in [0, \tau^*]$$

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

We comment that similar analysis to delineate the dual vectors is not replicable for the general system with both abandonment and class-transition. We shall illustrate the difficulty for a simplified system with one-way class-transition, namely, $\gamma_1 = 0$. Consider the scenario where the $c\mu$-rule prioritizes Class 2 and the modified $c\mu/\theta$-rule prioritizes Class 1 ($\gamma_1 = 0$). 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 class-transition, 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 ($\gamma_1 = 0$), the analysis is hindered by not being able to characterize the policy curve as well as the optimal state trajectory.

Appendix E: MDP Solutions in Section 4.4

In this section, we provide details about how we solve the transient scheduling problem (S2) to derive the MDP policy in Figure 14. In addition, we elaborate on the initialization for the simulation experiments in Table 1.

We use the uniformization approach with truncation to solve the MDP (S2). Let $\Lambda := \lambda_1 + \lambda_2 + (\mu_1 + \mu_2)s + (\theta_1 + \theta_2 + \gamma_1 + \gamma_2)X_{max}$, where $s = 3$ in the small system we consider, and the maximum number in system after truncation is $X_{max} = 40$. To truncate the infinite state space Markov process, the transition rates are modified such that the number-in-system does not exceed X_{max} for each class. In our setting, if $X_1 = 40$, the arrival rate to Class 1 is set to $\lambda_1 = 0$, and the deterioration rate from Class 2 to Class 1 is set to $\gamma_2 = 0$. Similar treatment is applied to Class 2 when $X_2 = 40$.

Define the set of feasible server allocations as

$$\mathcal{Z}(X_1, X_2) := \{(Z_1, Z_2) \in \mathbb{Z}_+ \times \mathbb{Z}_+ : Z_1 \le X_1,\ Z_2 \le X_2,\ Z_1 + Z_2 \le s\}.$$

The bellman operator for the MDP takes the form

$$\begin{aligned}\Xi(X_1, X_2) = \frac{1}{\Lambda}\Bigg[& c_1(X_1 - Z_1) + c_2(X_2 - Z_2) + \min_{(Z_1,Z_2)\in\mathcal{Z}(X_1,X_2)} \Bigg\{ \lambda_1 \Xi(X_1 + 1, X_2) + \lambda_2 \Xi(X_1, X_2 + 1) \\ & + (Z_1\mu_1 + \theta_1(X_1 - Z_1))\Xi(X_1 - 1, X_2) + (Z_2\mu_2 + \theta_2(X_2 - Z_2))\Xi(X_1, X_2 - 1) \\ & + \gamma_1(X_1 - Z_1)\Xi(X_1 - 1, X_2 + 1) + \gamma_2(X_2 - Z_2)\Xi(X_1 + 1, X_2 - 1) \\ & + (\Lambda - \lambda_1 - \lambda_2 - Z_1\mu_1 - \theta_1(X_1 - Z_1) - Z_2\mu_2 - \theta_2(X_2 - Z_2) - \gamma_1(X_1 - Z_1) \\ & - \gamma_2(X_2 - Z_2))\Xi(X_1, X_2) \Bigg\} \Bigg] \quad \text{if } X_1 + X_2 > s,\end{aligned}$$

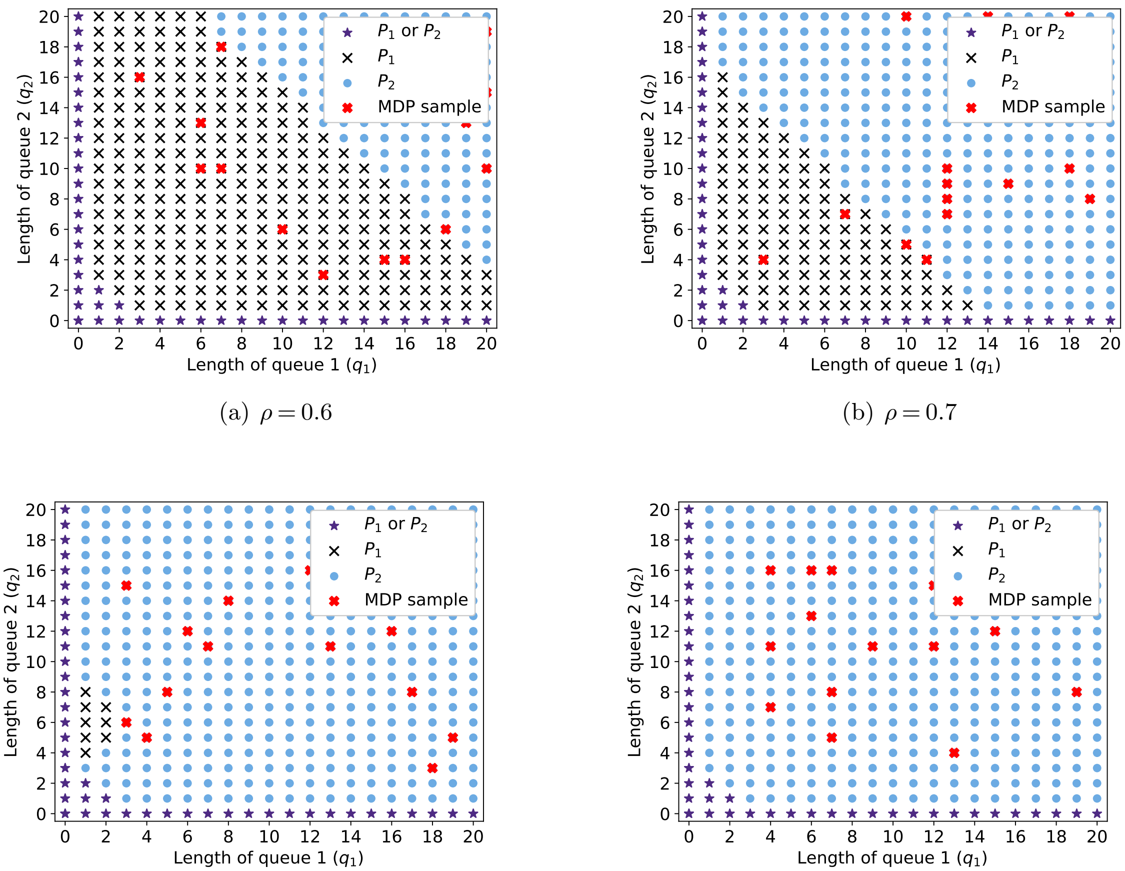
and

$$\Xi(X_1, X_2) = 0 \quad \text{if } X_1 + X_2 \le s. \tag{42}$$

Note that (42) reflects the terminal cost 0 when the system reaches 0 queue (absorbing states) in the transient control problem (S2).

In Table 1, when simulating the system dynamics under different policies, we select $J = 15$ initial conditions by sampling X_1 and X_2 independently and uniformly from 3 to 20. Since the small system in consideration has 3 servers, the lower bound is set so that there is positive queue at initialization under any server allocation. Figure 20 illustrates the selected initial points as red crosses.

Figure 20 Initialization (red crosses) for the simulation in Table 1 and the corresponding optimal MDP solutions



(a) $\rho = 0.6$ (b) $\rho = 0.7$

(c) $\rho = 0.8$ (d) $\rho = 0.9$