Essays on Skills-Based Routing

Jinsheng Chen

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2022

Abstract

Essays on Skills-Based Routing

Jinsheng Chen

Service systems such as call centers and hospital inpatient wards typically feature multiple classes of customers and multiple types of servers. Not all customer-server pairs are compatible, and some types of servers may be more efficient at serving some classes of customers than others. In the queueing literature, the problem of matching customers and servers is known as skills-based routing. This thesis consists of two works I have done in this area. The first work, which is done jointly with Jing Dong and Pengyi Shi, considers the routing problem in the face of a demand surge such as a pandemic. It shows how future arrival rate information, which is often available through demand forecast models, can be used to route near-optimally, even when there may be prediction errors. The methods used involve fluid approximations and optimal control theory, and the policies obtained are intuitive and easy to implement. The second work, which is done jointly with Jing Dong, incorporates a staffing element in addition to routing. Asymptotically optimal staffing and scheduling policies are derived for an M-model, both with and without demand uncertainty. The methods used involve diffusion approximations and stochastic-fluid approximations.

Table of Contents

Acknowledgements

First of all, I must thank Jing Dong for her continuous support and guidance throughout my PhD study. Both works that comprise this thesis were done jointly with her, and they would not have been possible without her constant patient encouragement and suggestions as well as her immense knowledge of the area.

I must also thank Pengyi Shi for collaborating extensively on the first work of this thesis (on optimal routing under demand surges). The work certainly owes much to her deep insights and support.

I would also like to thank the Columbia faculty for teaching me much about operations research. In particular, I am indebted to Ward Whitt, David Yao, and Henry Lam for sharing with me their profound insights on queueing and stochastic processes, serving on my thesis committee, and making suggestions that have no doubt improved this thesis.

Finally, I thank my family and friends for their unwavering support throughout the years.

Chapter 1: Introduction

1.1 Introduction

Resource flexibility is a topic that has been much studied over the years in many contexts, such as in manufacturing systems and in supply chains. In service systems, there are typically multiple types of customers (or jobs) arriving at the system, and they need to be processed by a server. There are different types of servers, who may be described by the types of customers they can serve. Typical examples are call centers, where customers may request different types of services or services in different languages, and hospitals, where patients may be grouped according to the type of condition they have. A server who can only serve a single type of customer is said to be *dedicated*, while a server who can serve multiple types of customers is said to be *flexible*. The processing time of a customer may depend on the type of server he or she is assigned to. For example, patients at a hospital may be discharged faster if they are sent to a ward that is staffed with nurses specialized in handling their condition.

Precisely because there are multiple types of customers and servers, it is highly non-trivial to match customers with servers in real-time. Servers who complete service and find that there are multiple classes of customers waiting in queue may be assigned new customers to serve, or be left to idle. Such decisions are subject to the real-time availability of servers and the real-time queue lengths, and are made to optimize certain performance metrics related to operating cost or service quality. This problem of matching customers with servers to optimize some performance metric is referred to as skills-based routing [1]. The recent survey [2] describes many important papers in the area.

Despite the large and growing body of literature on skills-based routing, there remains much to study in this area. Due to the complexity of analyzing scheduling decisions, most of the work in

the literature imposes rather restrictive assumptions to ensure mathematical tractability. However, many such conditions are unrealistic in practice, and it is important to understand what happens when they are not met. For example, resources are often assumed to be able to costlessly and instantaneously switch between different processing activities, and flexible resources may be more costly to staff than dedicated ones. Such disadvantages of resource flexibility can greatly affect the structure of a near-optimal scheduling policy.

This thesis consists of two works that each examines how to do near-optimal skills-based routing when there are certain pitfalls of resource flexibility in service systems. The first work, on optimal routing under demand surges, studies skills-based routing in the face of demand surges. It studies how to make use of future arrival rate information when make routing decisions. This is important because of the growing availability of data and demand forecast models. The second work, on optimal sizing and scheduling of flexible servers, considers in addition to skills-based routing the problem of deciding how many of each type of server to staff. It shows that the joint staffing and scheduling problem can be decomposed into individual problems under certain conditions, and uses that to obtain near-optimal prescriptions. A brief summary of each of these two works is given below. Even more details on the motivation of each work, as well as a closer review of the relevant literature, can be found in their respective chapters.

Optimal routing under demand surges: the value of future arrival rates

Demand surges are an operational challenge faced by many service systems. Common examples are mass casualty incidents or pandemics leading to an influx of patients to a hospital, and major weather events leading to widespread flight cancellations and a large increase in calls to an airline call center. When facing demand surges, the key challenge is to do scheduling to return the system to normality as efficiently as possible.

In recent years, because of the advancement of statistical learning techniques and the increase in the availability of data, there has been a growing number of demand forecasting models. It is thus important to know how to make use of such demand forecasts in skills-based routing, and

account for the fact that forecasts may have errors.

Motivated by the pandemic, we study the problem in a setting that is especially relevant to healthcare. In a hospital, inpatients are best warded at the ward appropriate for their condition, because the nurses there are best trained to handle their condition. However, it is possible that the appropriate ward is full, and the patient therefore may be sent to a different ward. Such a routing decision is known as *overflow*, and it may come with costs such as a lower quality of care (e.g. a longer length-of-stay) or other inconvenience costs for staff [3]. The work thus considers scheduling in a multi-class multi-pool queueing network with the following three important features:

1. Time-varying arrival rates: One or more customer classes may experience demand surge for certain periods of time. These demand surges are forecasted, and may be subject to prediction errors.

2. Service slowdown: To capture the potential efficiency loss when customers are served by servers from the non-primary pool, the service rates are both class- and pool-dependent. For a particular class of customers, the service rate in the primary pool is higher than the service rates in other non-primary pools.

3. Overflow cost: Assigning customers to non-primary pools may not only increase the service duration, but also impose other inconvenience costs or service quality cost. These are modeled through the overflow costs. That is, a penalty is charged for a customer that is placed with a non-primary server. Overflow costs can be both class-and-pool-dependent, reflecting the heterogeneous levels of "utility" or inconvenience cost when assigning customers from different classes to different server pools.

We exploit information about the demand surge (that is, information about the future arrival rates) to optimally deal with it. The main results and contributions are as follows:

1. *Modeling and solution framework*: The models incorporate general time-varying arrival rates, class-and-pool-dependent service rates, and overflow costs. These key features arise

in many service systems, especially healthcare delivery systems, and are crucial to take into account when studying demand surge with partial flexibility. Due to the non-stationarity and high dimensionality, the corresponding routing problem has not been well-studied in the literature. We leverage fluid approximation and optimal control theory to develop near-optimal transient controls for these systems.

2. *The value of future information:* The near-optimal control that we develop is interpretable and easy to implement. One follows an index that takes into account the holding costs, the service rates, the overflow costs, and the time it takes to empty each class using primary resources only. Knowledge of future arrival rates helps us to choose the best time to do overflow. For example, whereas common policies such as the maximum pressure policy react to large queue lengths and will only prioritize a class once it is sufficiently congested, our policy can prioritize a class even when it is lightly congested, if it is about to experience a demand surge. That is, our policy makes use of future information to proactively initiate overflow. Importantly, we show that even in the presence of forecast errors, our policy can perform very well.

3. *Pontryagin's minimum principle*: We demonstrate how optimal control theory can be applied to develop structural insights into the optimal routing policies in transient queueing control problems. In particular, the index-based policies that utilize future arrival rate information are derived using the Pontryagin's minimum principle for the corresponding fluid optimal control problems. For the main model, we further establish that the optimal control derived from the fluid control problem is asymptotically optimal for a properly scaled sequence of transient stochastic queueing control problems. The development involves non-trivial applications of the Pontryagin's minimum principle due to the state constraints. These state constraints are further complicated by the fact that there are multiple customer classes and general time-varying arrival rates. For example, the queue length of one class can be at zero while overflow remains in place. More importantly, the obtained optimal controls which are

highly interpretable and easy to implement.

Optimal sizing and scheduling of flexible servers

In practice, flexible servers are typically more expensive to hire due to their added capabilities. Also, due to multi-tasking, flexible servers may be less efficient than dedicated servers. For example, a multilingual call center agent would typically be paid more than a monolingual one, but may still have a preferred language to use, and only use other languages when required by operational demands. In this paper, we study how to jointly staff and schedule service systems when these shortcomings of resource flexibility are present.

Two demand scenarios are considered. One has deterministic arrival rates, which occurs when we have a very accurate estimate of customer demand. In this case, the flexible pool can be used to hedge against stochasticity, i.e., the stochastic fluctuation of interarrival times and service times. In particular, due to the stochasticity in system dynamics, one queue may incur an above average load while the other is at or below its normal load from time to time. In such situations, the flexible pool can be used to help the class with a heavier load, and thus balance the load between the two classes. The other scenario has random arrival rates, which occurs when there is a high degree of uncertainty in customer demand. In this case, the flexible pool is mainly used to hedge against parameter uncertainty. In particular, when the realized arrival rate of one class is higher than average while the realized arrival rate of the other class is at or below average, the flexible pool can be used to help the class with a higher realized arrival rate, and thus balance the load.

Because staffing and scheduling decisions interact, the joint optimization problem is very challenging. The approach we use is to show that the two decisions can be decoupled, and to obtain the optimal staffing levels through asymptotic analysis when the scheduling policy is fixed. When arrival rates are deterministic and symmetric, we use a coupling construction to derive the optimal scheduling policy for any staffing level. The scheduling policy prioritizes the dedicated servers (faster servers) when routing customers to servers, and prioritizes the class with more customers in the system when scheduling flexible servers, assuming the abandonment rate is less than the

service rates. Given the optimal scheduling policy, we then optimize the staffing policy. To derive structural insights into the size of the flexible pool, we employ a heavy-traffic asymptotic approach, where we send the arrival rate to infinity and study how the size of the flexible pool scales with the arrival rate. The result provides necessary and sufficient conditions for staffing rules to be asymptotically optimal. The key insight is that when flexibility comes at a cost, the optimal size of the flexible pool only leads to partial resource pooling: The flexible pool helps with load-balancing, but the effect is not large enough to equalize the two queues asymptotically.

When arrival rates are random and the magnitude of the parameter uncertainty dominates the system stochasticity, we use a stochastic-fluid relaxation of the optimal staffing problem. In this relaxation, we ignore the stochasticity of the queueing dynamics and focus on the parameter uncertainty only. The stochastic-fluid optimization problem is a special case of the single-period multi-product inventory problem with demand substitution, for which we can characterize the optimal solution explicitly. The relaxation also motivates a simple scheduling rule that essentially decomposes the M-model into two independent inverted-V models for any realization of the arrival rates. As the arrival rates (demands) increase, we show that the staffing and scheduling rules derived based on the stochastic-fluid relaxation are asymptotically optimal. The key insight is that when facing both parameter uncertainty and cost of flexibility, the optimal size of the flexible pool provides some hedging against the parameter uncertainty, and the cost saving, compared to the no-flexible resource case, is increasing with the magnitude of the uncertainty.

In addition to providing prescriptive solutions to managing flexibility, our work also makes the following contributions to the queueing literature.

1. When the arrival rates are symmetric and deterministic, we construct the optimal scheduling policy for any arrival rates and staffing levels. In contrast to most of the optimal scheduling literature for multi-class queues, the results do not rely on any asymptotic argument (see, for example, [4]). Instead, the proof uses a coupling argument that can be of interest when analyzing other Markovian queueing systems. The coupling technique also allows us to establish the optimality of a non-standard scheduling policy when the abandonment rate is

larger than the service rates.

2. When the arrival rates are deterministic and the flexible pool is of the optimal order, we derive the diffusion limit of the M-model under heavy-traffic. The limit is a two-dimensional diffusion process. In particular, the complete resource pooling condition is not satisfied when the flexible pool is optimally sized, i.e., the flexible pool size is not large enough to instantaneously balance the queue lengths between the two classes. Thus, there is no state space collapse in the limit, i.e., the two-dimensional queue length process does not reduce to a one-dimensional process in the limit. This is in contrast to most of the optimal scheduling literature (see, for example, [5, 6]). On the other hand, the limiting process cannot be fully decomposed along each dimension, i.e., the drift terms of the two component diffusion processes are interconnected. Thus, partial resource pooling is achieved.

3. When the arrival rates are random and the parameter uncertainty is of a larger order than the stochasticity of the queueing dynamics, we quantify the optimality gap for policies derived based on the stochastic fluid approximation. This extends the results in [7] from a multi-server queue with a single class of customers and a single pool of servers to a multi-class queue with multiple server types.

1.2 Organization

This thesis is organized as follows. The first work, on optimal routing under demand surges, occupies Chapter 2. The second work, on optimal staffing and scheduling of flexible servers, occupies Chapter 3. To make for easier reading, all proofs appear in the Appendix.

Chapter 2: Optimal Routing under Demand Surges: The Value of Future Arrival Rates

2.1 Introduction

In service systems, there are typically multiple classes of customers with different service needs. It is of critical importance for service operations management to allocate the proper amount of resources to meet the needs of each class of customers. The resource allocation problem is particularly relevant and challenging when certain classes of customers experience demand surges, and yet their dedicated capacity cannot be scaled up quickly. With the recent advancement of statistical learning tools and the growing availability of data, many advanced demand forecasting models have been developed to accurately predict demand surges. Take the COVID-19 pandemic as an example: researchers from different fields have worked together to develop prediction models for demand surges of different types of hospital resources (e.g., ICU beds and ventilators). Yet, the majority of these prediction models do not provide prescriptive solutions for effective allocation of resources. To maximize the practical impact, hospital management needs more concrete decision support on how to translate the demand forecast to determining whether they have enough beds to accommodate the COVID-19 patients, and if not, whether they should use beds from other specialties by cancelling or delaying elective surgeries. In this work, we study how to utilize future demand information, even when there are certain prediction errors, to design optimal routing strategies under demand surges. We explicitly characterize how to incorporate future demand into routing decisions and quantify the benefit of doing so.

We first elaborate on our modeling framework. In recent years, customer specialization has become increasingly sophisticated with the trend of personalized service. To better serve customers' different needs, it is common for service systems to have primary server pools, each dedicated to

serving a specific class of customers, with "servers" in that pool trained in skill sets tailored to the primary customer class. For example, hospitals usually partition inpatient beds into different specialty units and assign nurses trained to care for patients in that specialty. Call centers may hire agents who speak different languages to serve customers with different language needs. In the rest of the work, we use servers to refer to the critical resources in service systems, such as hospital beds, call center agents, etc. While specialization for each class of customers has the obvious benefits of delivering high-quality services and improving customer experiences, it is also common practice for service systems to cross-train servers so that they can serve non-primary customers when necessary. The primary reason is the presence of uncertainty and variability in demand. In particular, demand fluctuates quickly over time, but the capacity in each server pool is rather static since it takes time to train new staff or build new facilities.

We distinguish between two type of variabilities in demand: one is normal stochastic fluctuations due to the randomness in customer arrivals and service requirements; the other is the unusual increase in demand that can cause prolonged congestion in the system. The latter is often referred to as demand surge. To deal with the normal range of stochastic fluctuations in demand, a certain degree of slackness is usually added in the capacity, e.g., employing the square-root staffing rule, where the staffing level is set to meet the mean demand plus an uncertainty hedging term that is of the square root order of the mean demand. This staffing principle is shown to be near-optimal for systems operating in a stationary environment [8].

Demand surges happen infrequently, but can lead to substantial congestion in the system and service quality deterioration. For example, the ongoing COVID-19 pandemic has put an enormous amount of stress on healthcare delivery systems. A bad flu season or a mass casualty incident can cause a sudden increase in certain types of patients arriving at hospitals. Advancements in technology have significantly increased our ability to forecast such surges. With this future demand information, system managers may proactively leverage partial flexibility to effectively mitigate the effect of demand surges. However, concrete decision rules are still lacking, especially when we have to carefully balance between the benefits and costs associated with partial flexibility. For

example, in the hospital inpatient flow setting, partial flexibility comes in the form of off-service placement – sending patients of a particular specialty to a non-primary ward that is designated to treat a different specialty. While off-service placement can help achieve better resource utilization, it can also lead to worse patient outcomes, including longer length of stay and higher readmission rate [3]. In addition, it may generate a greater workload for nurses in the off-service ward due to multi-tasking [9], and can create a sense of unfairness among staff [10]. Similar tradeoffs between the benefits and the costs of flexibility are also pervasive in other service systems. Examples include call centers [11], bike-sharing [12], and emergency departments [13].

In this work, we take an important first step to study how to leverage demand forecasts to deal with temporary demand surges with (partial) flexibility. The goal is to design routing policies that optimally balance the benefits and costs of flexibility. To model the typical service setting, we consider a queueing system with multiple classes of customers and multiple servers pools. Each class has its own dedicated pool of servers, which we refer to as its primary pool. When the system is congested, customers can be assigned to non-primary pools. Such routing is referred to as "overflow." We consider the following costs of overflow:

1. Service slowdown: To capture the potential efficiency loss when customers are served in a non-primary pool, the service rates are both class- and pool-dependent. For each class of customers, the service rates in the non-primary pools are slower than in the primary pool.

2. Overflow cost: Assigning customers to non-primary pools not only increases the service time, but also imposes other inconvenience costs and/or costs caused by a compromised service quality. We model these costs through overflow costs. That is, a penalty is charged for a customer that is placed in a non-primary pool. We allow the overflow costs to be both class- and pool-dependent, which can reflect heterogeneous levels of "utility" (or inconvenience cost) when assigning customers from different classes to different server pools.

The routing policy refers to whether to assign a customer to a non-primary pool or to keep her waiting to be served by a primary pool server. To capture the demand surge, we allow general

time-varying arrival rates whereby one or more customer classes may experience one or multiple demand surges within certain time periods. Scenarios with multiple surges are of particular interest in light of the recent COVID-19 pandemic. To understand the value of demand forecasts, we focus on the scenario where we have access to the future arrival rate information, which may potentially be subject to prediction errors.

Our objective is to minimize the cumulative holding (waiting) costs and overflow costs until the demand surge is fully absorbed. When facing demand surges, overflow should be utilized to efficiently reduce congestion that is captured via the holding cost. However, due to the time-varying arrivals, service slowdowns, and overflow costs, deriving the optimal overflow strategy faces significant challenges and can render existing well-known routing policies, such as the $c\mu$-rule or the maximum pressure policy, highly suboptimal. Specifically, these challenges include:

1. The general time-varying arrival rate complicates the decision of when to initiate or stop overflow: it could be optimal to initiate overflow before the queue builds up in anticipation of a demand surge; or to end the overflow earlier, before the queue length is depleted, in anticipation of the end of a surge. The prediction error and multiple surges present further complications.

2. Because of service slowdowns, a too aggressive overflow strategy may generate more holding cost than having no overflow.

3. In the presence of overflow costs, the optimal routing policy may not be workload-conserving. That is, when a class has a positive queue, even when the non-primary pools have extra capacity, it may be better to keep these non-primary servers idle to avoid overflow costs.

The non-stationarity and high-dimensionality of the problem can render many existing analytical and numerical tools inapplicable. To derive structural insights on the optimal routing strategy, we take the fluid approximation approach. In particular, we formulate a transient fluid optimal control problem and derive closed-form solutions that leverage the future arrival rates. We then translate the fluid-based policy to the stochastic system and show that the policy achieves near-

optimal performance even when there are certain prediction errors. Our main contributions are summarized as follows:

I. Routing with demand forecast. As far as we know, we are one of the first to explicitly prescribe how to incorporate demand forecast of time-varying arrival rates into real-time routing decisions in a multi-class, multi-server system. Most existing papers on time-varying queues either assume that the time-varying arrival rates are known with some periodic pattern [14], or they focus on capacity planning to hedge against the uncertain arrival rates [15]. On the other hand, most demand forecasting work focus on the prediction side only (see, for example, [16] for call center arrivals and [17] for hospital impatient occupancy). Our work bridges demand forecasts with a key operational decision: customer routing. We explicitly characterize how to incorporate the demand forecast in the routing policy by solving a corresponding fluid transient control problem. We also show that the routing policy is asymptotically optimal for a properly scaled sequence of systems.

II. Two-stage index-based policy using future information. The focus of our main development is the N-model, a two-class model in which the primary pool for class 2 can provide help to class 1 but not the other way around. The near-optimal control policy that we develop can be summarized as a two-stage index-based look-ahead policy, which is highly interpretable and easy to implement. In the first stage, we compare the $h\mu$ index, where h is the holding cost and μ is the service rate, to decide which class can be prioritized. In the second stage, we look at another index that combines the $h\mu$ index, the time it takes to empty the queues with a proper set of resources, and the overflow costs to decide how long the overflow (if any) should last. The calculation of the time to empty the queues is where the future arrival rate information is utilized. The actual policy will be made precise in Section 2.3.

Interpreting our two-stage index-based policy provides insights into the value of future demand information and how to prioritize different customers under demand surges. In particular, based on the second-stage index, our policy suggests that other server pools may start prioritizing the customer class that is about to experience a demand surge, even though this class is not very congested yet. Similarly, when a customer class has a large queue, but the demand surge is about to

dissipate, other server pools may stop serving this customer class in anticipation of the upcoming drop in demand. This proactive nature of our policy is distinct from common non-look-ahead policies such as the $c\mu$ rule and the maximum pressure policy, and is why our policy outperforms these benchmark policies.

We stress that although future demand information being beneficial is somewhat expected, it is highly nontrivial to identify the proper form of incorporating it in the routing decision. Our results show that one needs to compare the holding cost and overflow cost in proper time-scales by accounting for the future impact of the overflow action. This is in contrast to comparing the instantaneous cost reduction, as in the $c\mu$ rule or its adjusted version that accounts the instantaneous change in overflow cost.

III. Pontryagin's minimum principle. Due to the time-nonstationarity and high dimensionality of our model, the corresponding routing problem has not been well-studied in the literature. We derive structural insights by studying the corresponding fluid transient control problems. In particular, our index-based policies that utilize future arrival rate information are derived using Pontryagin's minimum principle for the corresponding transient fluid control problems. Our derivation of the optimal scheduling policy involves non-trivial applications of Pontryagin's minimum principle due to the state constraints. A main difficulty lies in coming up with the right dual variables satisfying these constraints. The problem is further complicated by the fact that there are multiple customer classes and general time-varying arrival rates. We explicitly characterize all the dual variables, which lead to the closed-form characterization of the optimal control. These developments could shed light on other transient queueing control problems.

IV. Practical applicability to complex systems. For a routing policy to be useful in practice, it needs to be adaptable to (i) complicated network structures and (ii) demand forecasts with errors or limited look-ahead time windows. For (i), we extend the fluid-control analysis beyond the N-model, and study the X-model, the many-help-one extended N-model, and the one-help-many extended N-model. These extensions provide insights into designing good routing policies for more general systems. Based on the structure of optimal fluid control in these models, we pro-

pose a two-stage index-based look-ahead policy for general multi-class, multi-pool systems. We evaluate the performance of this look-ahead policy in stochastic systems via simulation. We compare the performance of this heuristic policy to other benchmark policies, such as the $c\mu$ rule and the maximum pressure policy, and show that our policy achieves superior performance for a wide range of system parameters.

For (ii), we substantiate our theoretical analysis with numerical evaluation in scenarios where demand forecast has errors of different magnitude and the forecast is restricted to a limited time window. In these scenarios, our numerical results suggest that the proposed policy continues to perform well. The adaptiveness of our policy to complex systems, its robustness to noisy arrival rate information, together with its simplicity, make it very appealing for implementation in real service systems when demand surges are present. We also identify factors that drive our policy to perform better than other benchmark policies, providing useful insights on managing systems in time-nonstationary environments.

2.1.1 Organization

The rest of the chapter is organized as follows. We conclude this section with a brief review of the literature. In Section 2.2, we introduce our main model, which is the N-model, and our main problem, which is designing an optimal routing policy under demand surge. To gain structural insights into the optimal policy, we study a deterministic fluid control problem in Section 2.3. We start with a single demand surge and perfect future arrival rate information, and then introduce adaptations for estimation errors, limited look-ahead time windows, and multiple surges. The fluid control problem can be viewed as an approximation to the original stochastic problem. We establish the asymptotic optimality of the policy derived from the fluid control for a sequence of stochastic systems in Section 2.4. To extend the routing strategy to more general network structures, we study the optimal fluid control problem for several extended models in Section 2.5. Based on the similar policy structure from these model extensions, we propose a two-stage index-based look-ahead policy for general parallel-server systems. We substantiate our theoretical

analysis with extensive numerical experiments in Section 2.6. All proofs are left to the Appendix.

2.1.2 Literature Review

Our work is related mainly to four streams of literature: flexibility in service systems; skill-based routing; optimal control theory in queueing; and scheduling with future demand.

Flexibility in service systems. In the operations management literature, it is well-known that resource pooling, sometimes through creating flexible resources, can drastically improve system performance [18, 19, 20, 21, 22]. [23] and [24] show that in a stationary environment, even a little flexibility can lead to substantial performance gain. However, in recent years, a growing amount of research has also studied situations in which pooling may not be as beneficial. This can be due to system architectures [25], different priorities among different classes of jobs [26], efficiency loss due to multi-tasking [27], and agent incentives [13], to name a few. Our work contributes to this line of literature by analyzing how resource pooling should be utilized when overflow assignment is associated with a slowdown effect and overflow costs in a time-nonhomogeneous environment.

Skill-based routing (SBR). There is a rich literature on SBR [28]. An exact analysis of SBR is usually analytically intractable due to the large state space and policy space. Much of the SBR literature utilizes a heavy-traffic asymptotic framework to gain analytical tractability. Our work relates to conventional heavy-traffic scaling. In this regime, [29] studies the scheduling problem of multi-class $G/G/1$ queues with convex holding costs and establishes the asymptotic optimality of the generalized $c\mu$ rule. In the N-model setting with preemption, [30] show that a threshold-based priority rule is asymptotically optimal. [31] consider a general service system with multiple customer classes and multiple types of flexible servers. They show that the generalized $c\mu$-rule is asymptotically optimal over all scheduling disciplines (preemptive and non-preemptive).

Apart from the $c\mu$ rule, the maximum pressure policy is another commonly used policy in SBR. The maximum pressure policy takes the same form as the MaxWeight policy in parallel server systems. [32] and [33] show that the maximum pressure policy is throughput optimal. [34] further prove that with quadratic holding cost, the maximum pressure policy is asymptotically opti-

mal under the conventional heavy-traffic scaling for some models. [35] establishes the asymptotic optimality of a general class of MaxWeight policies with strongly convex holding costs. There is also a rich literature on SBR in the many-server asymptotic regime; see [2] for a survey. Our work focuses on optimal routing to deal with demand surge, and our proposed policy is fundamentally different from the policies derived in the literature. In particular, we explicitly characterize how future arrival rates and overflow costs should be properly considered when making routing decisions.

Fluid transient control. Fluid approximation, which can capture first-order system dynamics well [36, 14], is often used to analyze transient queueing behavior. [37] and [38] detail how to develop effective queueing control policies based on fluid approximations. We employ optimal control theory to characterize the optimal fluid control policy; see [39] and [40] for an overview of optimal control theory. In particular, we leverage Pontryagin's Minimum Principle [41]. [42] review several applications of optimal control theory to dynamic rate queues. In a recent work, [43] apply Pontryagin's Minimum Principle to study the optimal scheduling of proactive service in systems with customer deterioration. [44] leverage optimal control theory to study the optimal call-back scheduling policy in call centers. The policy developed in their paper also has a look-ahead structure that takes the future arrival rate into account. However, they focus on a single class of customers and, thus, need only a single index. In contrast, we identify a two-stage index structure when dealing with multiple classes of customers. The work most relevant to ours is [45]. The authors study routing policies in a two-class, single-server system. However, they do not incorporate the overflow cost, and their analysis is limited to simple time-varying arrival patterns (high/low constant arrivals). Our analytical framework allows us to study very general time-varying arrival rates and the tradeoff between holding and overflow costs.

The value of future demand. Our analysis highlights the value of future arrival rate information in transient control problems. A few recent works demonstrate the value of future demand information in developing effective admission control or scheduling policies [46, 44, 47]. These works require detailed future demand information, including the actual/predicted arrival times and service

times of customers. In contrast, our policy requires only the average future demand (i.e., arrival rate), which can be estimated more easily in practice. More importantly, we prove the asymptotic optimality of the fluid-based policy in the N-model even when there are certain prediction errors. Predicted demand has been utilized to optimize staffing decisions (see, e.g., [15, 48] for call center staffing and [49] for emergency department nurse staffing). Our work is different from the above works in two main aspects. First, the above works study stationary performance metrics while we focus on transient system dynamics under demand surge. Second, routing decisions are fundamentally different from staffing decisions, and the two can happen at very different time scales.

2.2 Problem Formulation

To demonstrate our methodology and key insights, we use the N-model as our main model; other network structures are studied in Section 2.5. The N-model consists of two customer classes and two server pools. Customers in class i, $i = 1, 2$, arrive at the system according to a time-varying Poisson process with rate $(\lambda_i(t))_{t\geq 0}$. Class 1 customers can be served by both pool 1 and pool 2 servers, while class 2 customers can be served only by pool 2 servers. The number of servers in pool i is s_i, $i = 1, 2$. The service times are exponentially distributed with class-and-pool-dependent service rates. In particular, if a class i customer is served by a server in pool j, the service rate is μ_{ij}. We assume that $\mu_{11} > \mu_{12}$ to capture the efficiency loss of non-primary service. We also define $\mu_{21} = 0$ to capture the service non-compatibility.

Let $X_i(t)$ denote the number of class i customers in the system at time t; $Z_{ij}(t)$ denote the number of class i customers in service in pool j at time t; and $Q_i(t)$ denote the number of class i customers waiting in the queue at time t. Note that

$$X_1(t) = Q_1(t) + Z_{11}(t) + Z_{12}(t) \text{ and } X_2(t) = Q_2(t) + Z_{22}(t).$$

Let A_i and S_{ij} denote rate-1 Poisson processes modeling the arrival and service processes, respec-

tively. Then, the system dynamics can be characterized via

$$X_1(t) = X_1(0) + A_1\left(\int_0^t \lambda_1(s)ds\right) - S_{11}\left(\mu_{11}\int_0^t Z_{11}(s)ds\right) - S_{12}\left(\mu_{12}\int_0^t Z_{12}(s)ds\right),$$
$$X_2(t) = X_2(0) + A_2\left(\int_0^t \lambda_2(s)ds\right) - S_{22}\left(\mu_{22}\int_0^t Z_{22}(s)ds\right),$$

where $Z(t) = (Z_{11}(t), Z_{12}(t), Z_{22}(t))$, $t \geq 0$, is determined by the scheduling policy. We consider the class of preemptive Markovian policies, which can be viewed as a mapping from $X(t) = (X_1(t), X_2(t))$ to $Z(t) = (Z_{11}(t), Z_{12}(t), Z_{22}(t))$, where $Z(t) \in \mathbb{N}_0^3$ satisfies

$$Z_{11}(t) \leq s_1, Z_{12} + Z_{22}(t) \leq s_2, Z_{11}(t) + Z_{12}(t) \leq X_1(t), Z_{22}(t) \leq X_2(t).$$

We consider non-anticipative policies that do not know realized customer demand in the future, but we allow the policies to take future arrival rates into account. Note that the arrival rates can be viewed as part of the system parameters. Let π denote a scheduling policy (i.e., a routing policy). We use the superscript π to denote the dependence of the system dynamics on the policy – e.g., X^π and Z^π. We occasionally suppress the superscript when it is clear from the context.

Focusing on planning under demand surges, we consider time-varying arrival rates that can cause one or more customer classes to experience surge in demand (arrivals). Without loss of generality, we assume that time 0 is the beginning of the demand surge and the demand surge will last for a finite amount of time. In addition, the demand surge is sufficiently large such that the total demand exceeds the total processing capacity during the surge period. (This will be made precise in Assumption 1 in the following section.) To illustrate the main idea, Figure 2.1 shows the realized and projected demand for Intensive Care Unit (ICU) beds by COVID-19 patients in the US [50], where one demand peak is from November 2020 to February 2021. During this peak time, hospitals had to cancel elective surgeries to accommodate the surge of demand from COVID-19 patients with severe respiratory symptoms.

Our goal is to operate the system in the most cost-effective way so that it returns to the normal state of operation after the demand surge. For the objective function, we consider two types of

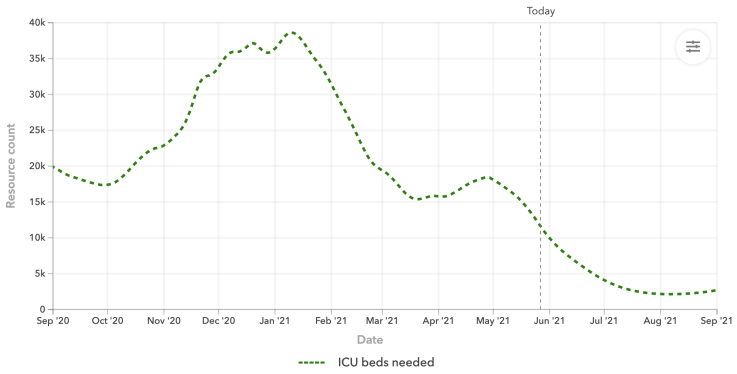


Figure 2.1: Demand for ICU beds for COVID-19 patients in the US, based on IMHE projection

costs related to the routing decisions: holding cost and overflow cost. We aim to find a scheduling (routing) policy that minimizes the total cost of holding customers in the queue and overflowing class 1 customers to pool 2 over a properly defined finite time horizon. Mathematically, we denote the holding cost for class i customers as h_i, $i = 1, 2$, and the overflow cost for each class 1 customer in pool 2 as ϕ_{12}. The optimal scheduling problem is formulated as finding a policy π that minimizes

$$V^{\pi}(x) = \mathbb{E}\left[\int_0^T h_1 X_1^{\pi}(t) + h_2 X_2^{\pi}(t) + \phi_{12} Z_{12}^{\pi}(t) dt \;\middle|\; X(0) = x\right]. \tag{2.1}$$

Here, the planning horizon T is a long enough time such that the system can fully absorb the demand surge by time T. (This will be made more precise in Section 2.4.)

Solving (2.1) – i.e., finding a policy π to minimize $V^{\pi}(x)$ – analytically is intractable. Even solving it numerically can be computationally prohibitive due to the large state space and policy space. Thus, we take the approach of studying a corresponding deterministic fluid control problem, which serves as a good approximation to (2.1).

2.3 Fluid Optimal Control

We first specify the deterministic fluid model $q(t) = (q_1(t), q_2(t))$ that corresponds to the stochastic system described in Section 2.2. The arrival rates and service rates in the fluid model are the same as the stochastic system introduced in Section 2.2. The dynamics of the fluid model

are characterized via

$$\dot{q}_1(t) = \lambda_1(t) - \mu_{11} z_{11}(t) - \mu_{12} z_{12}(t),$$
$$\dot{q}_2(t) = \lambda_2(t) - \mu_{22} z_{22}(t),$$

where $\dot{q}_i(t) := dq_i(t)/dt$. The scheduling trajectory in the fluid model, $z(t) = (z_{11}(t), z_{12}(t), z_{22}(t))$, is determined by a fluid admissible control that satisfies

$$q_1(t) \geq 0, q_2(t) \geq 0, z_{11}(t) \leq s_1, z_{12}(t) + z_{22}(t) \leq s_2, z_{11}(t) \geq 0, z_{12}(t) \geq 0, z_{22}(t) \geq 0.$$

We denote the set of admissible controls at time t as $\mathcal{Z}(t)$. We impose the following assumptions on the arrival rate functions.

Assumption 1. *The arrival rates $\lambda_1(t)$ and $\lambda_2(t)$ satisfy:*

1. *For $i = 1, 2$, $\lambda_i(t) \geq s_i \mu_{ii}$ when $t < \kappa_i$ and $\lambda_i(t) < s_i \mu_{ii}$ when $t \geq \kappa_i$.*
2. *$(\lambda_i(t))_{0 \leq t \leq \kappa_i}$'s are piecewise monotone with a finite number of pieces.*
3. *$\int_{\kappa_i}^{\infty} (s_i \mu_{ii} - \lambda_i(t)) dt = \infty$.*
4. *Given $X(0) = x$, for any $t \leq \kappa_1 \vee \kappa_2$, where $\kappa_1 \vee \kappa_2 = \max\{\kappa_1, \kappa_2\}$, $W(x, t) > 0$, where*

$$\begin{aligned} W(x,t) = \min_{z} \ & q_1(t) + q_2(t) \\ & \text{s.t. } \dot{q}_1(u) = \lambda_1(u) - \mu_{11} z_{11}(u) - \mu_{12} z_{12}(u), \ \ q_1(0) = x_1 \\ & \qquad \dot{q}_2(u) = \lambda_2(u) - \mu_{22} z_{22}(u), \ \ q_2(0) = x_2 \\ & \qquad z(u) \in \mathcal{Z}(u) \text{ for all } u \in [0, t]. \end{aligned} \tag{2.2}$$

The last condition in Assumption 1 indicates that the demand surge is large enough such that the fluid queue can not be emptied by any admissible control before $\kappa_1 \vee \kappa_2$. (Because the objective function in $W(x, t)$ solely focuses on minimizing the queue length, it leads to policies that aim to

empty the queue as fast as possible.) In addition, note that Assumption 1 considers a single demand surge for each class. We relax this assumption in Section 2.3.4 to consider multiple surges.

The fluid control problem corresponding to (2.1) is formulated as

$$\begin{aligned} \min_{z} \quad & \int_0^{\sigma} h_1 q_1(t) + h_2 q_2(t) + \phi_{12} z_{12}(t) dt \\ \text{s.t. } & q(0) = x \\ & \dot{q}_1(t) = \lambda_1(t) - \mu_{11} z_{11}(t) - \mu_{12} z_{12}(t) \\ & \dot{q}_2(t) = \lambda_2(t) - \mu_{22} z_{22}(t) \\ & z(t) \in \mathcal{Z}(t) \text{ for all } t \geq 0, \end{aligned} \tag{2.3}$$

where $\sigma = \inf\{t \geq \kappa_1 \wedge \kappa_2 : q_1(t) + q_2(t) = 0\}$. Note that under Assumption 1, with a proper scheduling policy, the fluid queue will eventually hit zero and stay there. Before presenting the optimal policy for (2.3), we make some remarks on Assumption 1.

Remark 1 (Future information on arrival rates)**.** *In the baseline fluid analysis, we assume the arrival rates $\{\lambda_i(t)\}_{t\geq 0}$ are known exactly and fully. Later, we show an adaptation of the optimal policy to scenarios where (i) we only have access to estimated arrival rates that can have prediction errors (Section 2.3.2), and (ii) we only have access to a limited look-ahead time window (Section 2.3.3). When translating the fluid control policy to the stochastic system, we will show that our proposed policy is asymptotically optimal even when the arrival rates are estimated with certain errors (Section 2.4).*

For $i = 1, 2$ and $t \geq 0$, define the function $G_i^t : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ as follows. For $x > 0$,

$$G_i^t(x) := \inf\left\{\Delta \geq 0 : \int_t^{t+\Delta} (s_i \mu_{ii} - \lambda_i(s))\, ds = x\right\}, \tag{2.4}$$

and for $x = 0$,

$$G_i^t(0) := \lim_{x \downarrow 0} G_i^t(x). \tag{2.5}$$

We can interpret $G_i^t(x)$ as the time it takes to empty queue i after time t using only primary resources, given that $q_i(t) = x$. For a fixed value of t, it is continuous and strictly increasing in x. Note that under (2.5), $G_i^t(0)$ could be positive if there is an upcoming demand surge, and it is the time until the effects of the demand surge can be fully absorbed using only primary resources. The next theorem characterizes the optimal routing policy for the fluid control problem.

Theorem 1 (Optimal control policy in N-model)**.** *Under Assumption 1, the optimal control for* (2.3) *takes the following form. Pool 1 serves as many class 1 customers as possible – i.e.,*

$$z_{11}^*(t) = s_1 1\{q_1(t) > 0\} + \left(s_1 \wedge \frac{\lambda_1(t)}{\mu_{11}}\right) 1\{q_1(t) = 0\}.$$

Moreover,

I. When $h_1\mu_{12} \geq h_2\mu_{22}$, pool 2 gives priority to class 1 when queue 1 is large enough relative to queue 2. In particular,

a. If $h_1\mu_{12}G_1^t(q_1(t)) > h_2\mu_{22}G_2^t(q_2(t)) + \phi_{12}$, pool 2 gives priority to class 1 – i.e.,

$$z_{12}^*(t) = s_2 1\{q_1(t) > 0\} + \left(s_2 \wedge \frac{\lambda_1(t) - z_{11}^*(t)\mu_{11}}{\mu_{12}}\right) 1\{q_1(t) = 0\}, \text{ and}$$
$$z_{22}^*(t) = (s_2 - z_{12}^*(t)) 1\{q_2(t) > 0\} + \left((s_2 - z_{12}^*(t)) \wedge \frac{\lambda_2(t)}{\mu_{22}}\right) 1\{q_2(t) = 0\}.$$

b. Otherwise, pool 2 serves class 2 only – i.e.,

$$z_{12}^*(t) = 0 \text{ and } z_{22}^*(t) = s_2 1\{q_2(t) > 0\} + \left(s_2 \wedge \frac{\lambda_2(t)}{\mu_{22}}\right) 1\{q_2(t) = 0\}.$$

II. When $h_1\mu_{12} \leq h_2\mu_{22}$, pool 2 gives priority to class 2 and helps class 1 only when $q_2(t) = 0$ and $q_1(t)$ is large enough. In particular,

a. If $q_2(t) = 0$ and $h_1\mu_{12}G_1^t(q_1(t)) > \phi_{12}$, pool 2 provides partial help to class 1 – i.e.,

$$\begin{aligned} z_{12}^*(t) &= (s_2 - z_{22}^*(t))1\{q_1(t) > 0\} \\ &\quad + \left((s_2 - z_{22}^*(t)) \wedge \frac{\lambda_1(t) - z_{11}^*(t)\mu_{11}}{\mu_{12}}\right) 1\{q_1(t) = 0\}; \\ z_{22}^*(t) &= s_2 \wedge \frac{\lambda_2(t)}{\mu_{22}}. \end{aligned}$$

b. Otherwise, pool 2 serves class 2 only – i.e.,

$$z_{12}^*(t) = 0 \text{ and } z_{22}^*(t) = s_2 1\{q_2(t) > 0\} + \left(s_2 \wedge \frac{\lambda_2(t)}{\mu_{22}}\right) 1\{q_2(t) = 0\}.$$

The proof for Theorem 1 is in Appendix A.4. It utilizes the Pontryagin's Minimum Principle. The indices are derived based on the adjoint vector (i.e., dual function). This optimal control specified in Theorem 1 can be summarized as a two-stage index-based look-ahead policy. In the first stage, we compare the $h\mu$ index to decide whether pool 2 should fully prioritize class 1, or only partially help when having spare capacity. In particular, when $h_1\mu_{12} > h_2\mu_{22}$, pool 2 may prioritize class 1; otherwise, pool 2 prioritizes its own class and may provide partial help. Then, in the second stage, we decide how long pool 2 should help class 1 (either through full prioritization or partial help), by comparing $h_1\mu_{12}G_1^t(q_1(t)) - \phi_{12}$ with $h_2\mu_{22}G_2^t(q_2(t))$; help is provided only when the former index is larger than the latter. The $G_i^t(\cdot)$ term is the "look-ahead" component as it takes the future demand into account. In what follows, we refer to the scenario in which pool 2 prioritizes class 1 as providing full help to class 1, and the scenario in which pool 2 serves class 1 only when there is spare capacity (i.e., when queue is emptied) as providing partial help to class 1.

2.3.1 Interpretation of the Two-Stage Policy

Leveraging future arrival information

The optimal policy depends on $G_i^t(q_i(t))$, the time to empty the queue, which requires one to look ahead and take the future arrival rate into account. In particular, we note that (a) when

class 1 is not very congested but is about to experience a demand surge, pool 2 may already start to prioritize class 1 in anticipation of the upcoming demand surge; (b) when class 1 has a large queue, but the demand surge is about to dissipate, pool 2 may decide to stop serving class 1 in anticipation of the upcoming drop in demand. Figure 2.2 provides a demonstration of the role of the future arrival rate here. In this example, we set $\kappa_1 = 20$ and $\kappa_2 = 10$. For $t \leq 10$, class 1 is experiencing a moderate demand surge with $\lambda_1(t) = 1.5$; for $t \in (10, 20]$, class 1 is experiencing a more severe demand surge with $\lambda_1(t) = 2$. We observe that at time 0, even though class 2 is more congested than class 1 – i.e., $q_2(0) = 2$ while $q_1(0) = 0$ – we still choose to prioritize class 1 in pool 2 – i.e., $z_{12}(0) = 4$. This corresponds to scenario (a) as we are anticipating a demand surge for class 1. We also observe scenario (b), that is, even though the demand surge for class 1 ends at time 20, pool 2 stops prioritizing class 1 at time 15.8 – i.e., $z_{12}(t) = 0$ for $t \geq 15.8$. We discuss how to modify this look-ahead component when we only have estimated arrival rates with potential prediction errors in Section 2.3.2.

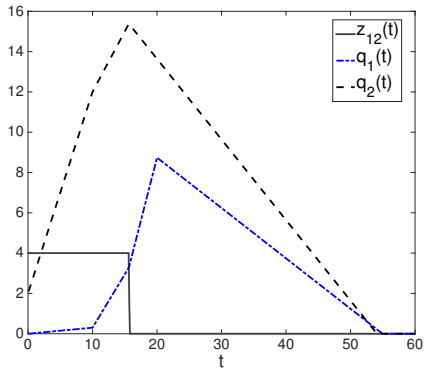


Figure 2.2: Optimal trajectory of the N-model. (Parameter setting: $s_1 = 3$, $s_2 = 4$, $\mu_{11} = \mu_{22} = 0.25$, $\mu_{12} = 0.18$, $h_1 = 1.5$, $h_2 = 1$, $\phi_{12} = 1$, $\lambda_1(t) = 1.5 \times \mathbf{1}\{0 \leq t \leq 10\} + 2 \times \mathbf{1}\{10 < t \leq 20\} + 0.5 \times \mathbf{1}\{t > 20\}$, $\lambda_2(t) = 1 \times \mathbf{1}\{0 \leq t \leq 10\} + 0.6 \times \mathbf{1}\{t > 10\}$, $q_1(0) = 0$ and $q_2(0) = 2$).

Even in the case of constant arrival rates, our policy takes the arrival rates into account by considering the difference between the arrival rate and the service capacity. In this case, $G_i^t(x_i) = \frac{x_i}{s_i\mu_{ii} - \lambda_i}$ for $x_i \geq 0$. The optimal scheduling policy is similar to the maximum pressure policy but takes the slack capacity $s_i\mu_{ii} - \lambda_i$ into account.

The effect of overflow costs

As discussed in the introduction, though it seems intuitive that one should use future arrival information when available, it is nontrivial to identify the proper form of incorporating this information, especially when there are costs associated with flexibility. We discuss in this section the importance of comparing the overflow cost and the holding cost at the right scale in routing decisions.

Our fluid-control policy shows that, when having a positive overflow cost, we should compare the per customer overflow cost to the holding cost over a time interval that is determined by the time it takes to empty the queue. To see this, we rewrite the condition for Case I in the following equivalent form:

$$h_1 G_1^t(q_1(t)) - \frac{\phi_{12}}{\mu_{12}} > h_2 G_2^t(q_2(t)) \frac{\mu_{22}}{\mu_{12}}. \tag{2.6}$$

Similarly, in Case II, we check whether

$$h_1 G_1^t(q_1(t)) > \frac{\phi_{12}}{\mu_{12}} \tag{2.7}$$

to decide whether pool 2 should provide partial help to class 1. Here, ϕ_{ij}/μ_{ij} corresponds to the expected amount of overflow cost for a class i customer completing service in pool j (with $1/\mu_{ij}$ being the average service time), while $h_i G_i^t(q_i(t))$ corresponds to the expected holding cost accumulated till the queue is depleted using primary resources only. In other words, the cost comparison needs to account for the future impact of the routing action via the accumulated overflow cost and the accumulated holding cost over a look-ahead time window. Note that for a (virtual) customer joining the queue at time t, $h_i G_i^t(q_i(t))$ also measures the queueing externality cost of this customer – i.e., the additional holding cost it imposes on the entire system; see [44] for an interpretation using the Last-in-First-out discipline.

We note that this cost comparison is in contrast to comparing both costs (overflow and holding) at the myopic cost-rate scale. For the cost rate, when using pool 2 to serve class i customers, the

holding cost decreases at rate $h_i\mu_{i2}$, while the overflow cost increases at rate $\phi_{i2}\mu_{i2}$. If we compare the instantaneous cost rate to determine which class should have priority, then we should check whether

$$h_1\mu_{12} - \phi_{12}\mu_{12} > h_2\mu_{22}$$

to decide if pool 2 should prioritize class 1. This myopic rule corresponds to the modified $c\mu$ rule that we consider in the numerical experiments in Section 2.6, which can result in significantly worse performance than our proposed policy in many settings. Our policy suggests that we should look beyond the instantaneous cost reduction rate and consider overflow versus holding cost from the system perspective, i.e., how the overflow decision impacts the future system congestion.

Simple priority switching structure

Our policy can be characterized via a time-and-state-dependent switching curve, which is defined as

$$\psi(t) = h_1\mu_{12}G_1^t(q_1(t)) - \phi_{12} - h_2\mu_{22}G_2^t(q_2(t)).$$

In Case I, when $\psi(t) > 0$, pool 2 gives priority to class 1; otherwise, pool 2 serves class 2 only. Furthermore, we can establish the following property for the switching curve.

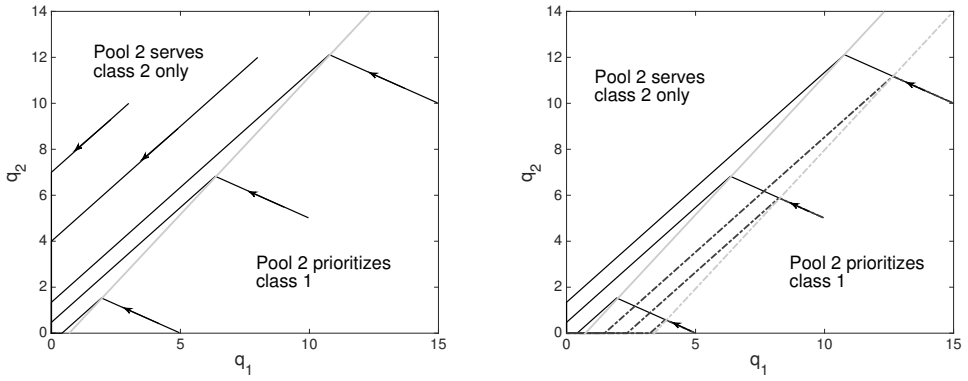
Lemma 1. *Under Assumption 1 and the control characterized by Theorem 1, let* $\tau_1 = \inf\{t \geq 0 : G_1^t(q_1(t)) = 0\}$. $\psi(t)$ *is monotonically decreasing in* t *for* $t \leq \tau_1$ *and* $\psi(t) < 0$ *for* $t > \tau_1$.

Lemma 1 indicates that in Case I, pool 2 switches priority at most once throughout the planning horizon. If it switches priority, it is from class 1 to class 2. If it does not switch priority, it serves class 2 only throughout the planning horizon. This is a highly desirable feature for policy implementation, since frequent priority switching can impose additional administrative burdens.

In the case of constant arrival rates, the switching curve reduces to a simple threshold policy. In particular, the switching curve partitions the state space of $(q_1(t), q_2(t))$ into two regions. In one region, pool 2 gives priority to class 1; in the other region, pool 2 serves class 2 only. To demonstrate this, Figure 2.3 (a) plots the optimal fluid trajectory $(q_1(t), q_2(t))$ for different initial

queue lengths. The switching curve is the grey line in the figure. When $(q_1(t), q_2(t))$ is below the curve (i.e., when $q_1(t)$ is sufficiently larger than $q_2(t)$), pool 2 prioritizes class 1; otherwise, pool 2 serves class 2 only.

The switching curve structure also allows us to conduct sensitivity analyses to visualize the impact of different system parameters. For example, Figure 2.3 (b) compares the fluid trajectories when $\phi_{12} = 1$ (solid) to the fluid trajectories when $\phi_{12} = 5$ (dashed). As the overflow cost increases, the optimal policy switches priority from class 1 to class 2 "earlier".



(a) $\phi_{12} = 1$ (b) $\phi_{12} = 1$ (solid) versus $\phi_{12} = 5$ (dashed)

Figure 2.3: Optimal trajectory of the N-model with different initial queues and overflow costs. (Parameter setting: $s_1 = s_2 = 2$, $\mu_{11} = \mu_{22} = 0.25$, $\mu_{12} = 0.2$, $\lambda_1 = \lambda_2 = 0.3$, $h_1 = 1.5$, $h_2 = 1$.)

2.3.2 Adaptivity to Estimation Errors

In this section, we consider the case where we only have access to estimated arrival rates. In particular, the estimated arrival rate for class i takes the form $\tilde{\lambda}_i(t) = \lambda_i(t) + \epsilon_i(t)$ where $\lambda_i(t)$ is the true arrival rate and $\epsilon_i(t)$ is the prediction error.

We propose to use the same two-stage policy even when the estimation is inaccurate. The estimation error affects the performance of the policy because the look-ahead function is now calculated based on the estimated arrival rate:

$$\tilde{G}_i^t(x) = \inf\left\{\Delta \geq 0 : \int_t^{t+\Delta} (s_i\mu_{ii} - \tilde{\lambda}_i(s))\, ds = x\right\}$$

for $x > 0$. In this case, pool 2 decides whether to help class 1 by checking whether $h_1\mu_{12}\tilde{G}_1^t(q_1(t)) - \phi_{12} > h_2\mu_{22}\tilde{G}_2^t(q_2(t))$.

When $\epsilon_i(t)$ is small, we expect $\tilde{G}_i^t(x)$ to be close to $G_i^t(x)$, and our policy should perform well. This intuition will be made rigorous when translating the fluid policy back to the stochastic system. In particular, we will show in Section 2.4 that under suitable conditions on the estimation error $\epsilon_i(t)$, the policy based on $\tilde{G}_i^t(x)$ is asymptotically optimal in the stochastic systems.

2.3.3 Adaptivity to Limited Look-ahead Windows

In this subsection, we consider the restriction of having only a limited look-ahead time window. Specifically, we assume that at time t, only the future arrival rate up to time $t + W$ is known. The constant $W \geq 0$ controls the amount of future information available: $W = 0$ corresponds to case with no future arrival rate information; $W = \infty$ corresponds to knowing the full future information.

With a limited look-ahead window, we adapt our policy as follows. We can calculate $\tilde{G}_i^t$'s using the nominal arrival rates outside the available time window. For example, when $W = 0$, we use $\tilde{G}_i^t(q_i(t)) = q_i(t)/(s_i\mu_{ii} - \lambda_i)$, where λ_i is the nominal arrival rate (i.e., the arrival rate before or after the demand surge).

Beyond this adaptation, we also note that the policy characterized in Theorem 1 actually does not require $W = \infty$. For example, consider the case where $h_1\mu_{12} \geq h_2\mu_{22}$, i.e., Case I in Theorem 1. In this case, pool 2 prioritizes class 1 if

$$G_1^t(q_1(t)) > \frac{h_2\mu_{22}G_2^t(q_2(t)) + \phi_{12}}{h_1\mu_{12}}. \tag{2.8}$$

Suppose that the class 2 arrival rate is known, so that $G_2^t(q_2(t))$ is known. Then, we only need to know the arrival rate of class 1 up to $W = (h_2\mu_{22}G_2^t(q_2(t)) + \phi_{12})/(h_1\mu_{12})$, which is sufficient to determine whether (2.8) is satisfied. For the case where $h_1\mu_{12} \leq h_2\mu_{22}$, i.e., Case II in Theorem 1, because pool 2 gives priority to class 2, no future arrival rate information is needed when $q_2(t) > 0$. When $q_2(t) = 0$, it is sufficient to know the arrival rate of class 1 up to $W = \phi_{12}/(h_1\mu_{12})$

to determine whether pool 2 should help class 1. Furthermore, in Section 2.6.3, we conduct simulation experiments to test the performance of our policy with varying values of W in the stochastic systems. We observe that our proposed policy achieves a good performance even with a relatively small look-ahead time window.

2.3.4 Adaptivity to Multiple Surges

Our analytical framework applies to very general arrival rates, including scenarios with multiple demand surges. In this section, we show an example in which class 1 experiences two demand surges, as characterized in Assumption 2.

Assumption 2. *The arrival rates $\lambda_1(t)$ and $\lambda_2(t)$ satisfy:*

1. *For class 1, there exist constants $0 < \kappa_a < \kappa_b < \kappa_c$ such that $\lambda_1(t) \geq s_1\mu_{11}$ for $t \in [0, \kappa_a] \bigcup [\kappa_b, \kappa_c]$ and $\lambda_1(t) < s_1\mu_{11}$ otherwise. For class 2, $\lambda_2(t) < s_2\mu_{22}$ for all $t \geq 0$.*
2. *$(\lambda_1(t))_{0\leq t\leq \kappa_c}$ is piecewise monotone with a finite number of pieces.*
3. *$\int_0^\infty (s_1\mu_{11} - \lambda_1(t))\, dt = \infty$.*
4. *Given $X(0) = x$, for any $t \in [0, \kappa_a) \bigcup (\kappa_b, \kappa_c)$, $W(x, t) > 0$, where $W(x, t)$ is defined in* (2.2).

We redefine $\sigma = \inf\{t > \kappa_c : q_1(t) + q_2(t) = 0\}$. The following theorem shows that the optimal control in this two-surge setting takes exactly the same form as before, with the look-ahead function G_i^t defined in (2.4).

Theorem 2 (Optimal control under two demand surges)**.** *Under Assumption 2, the optimal control for* (2.3) *takes the following form. Pool 1 serves as many class 1 customers as possible. Moreover:*

I. When $h_1\mu_{12} \geq h_2\mu_{22}$, pool 2 gives priority to class 1 when

$$h_1\mu_{12}G_1^t(q_1(t)) > h_2\mu_{22}G_2^t(q_2(t)) + \phi_{12};$$

otherwise, pool 2 serves class 2 only.

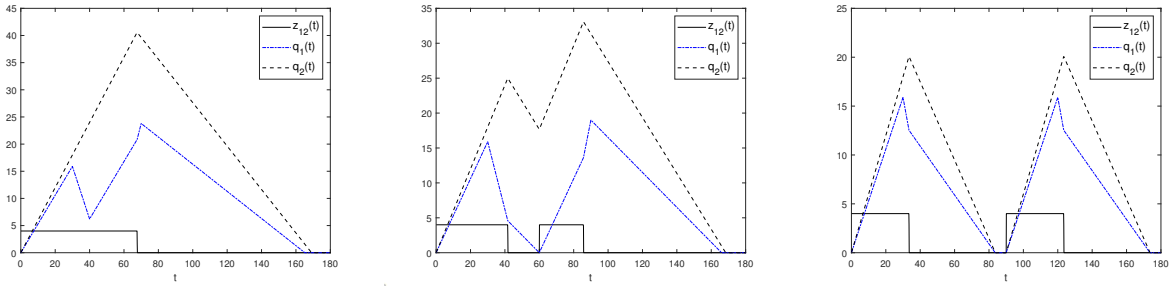
II. When $h_1\mu_{12} \le h_2\mu_{22}$, pool 2 gives priority to class 2 and will help class 1 when $q_2(t) = 0$ and $h_1\mu_{12}G_1^t(q_1(t)) > \phi_{12}$; otherwise, pool 2 serves class 2 only.

The proof of Theorem 2 is in the Appendix, and follows a similar framework as that of Theorem 1. The behavior of the system under the control characterized in Theorem 2 depends on the length of the interval between the two demand surges. We demonstrate the basic idea through a numerical example: Class 2 has a constant arrival rate $\lambda_2(t) \equiv 0.6$, while the arrival rate for class 1 takes the form

$$\lambda_1(t) = \begin{cases} 2, & 0 \le t < 30, \\ 0.5, & 30 \le t < 30 + K, \\ 2, & 30 + K \le t < 60 + K, \\ 0.5, & t \ge 60 + K. \end{cases}$$

In particular, there are two demand surges for class 1, and the length of the interval between the two surges is K, which we vary in the experiments plotted in Figure 2.4. When K is small – i.e., $K = 10$ in case (a) – the two demand surges are so close to each other that neither queue can be emptied before the beginning of the second demand surge, and we observe a single helping interval as in the single demand surge setting. When K is moderate – i.e., $K = 30$ in case (b) – the two demand surges are far enough apart for the class 1 queue to be emptied by the time the second demand surge begins, but not far enough apart for the class 2 queue to be emptied then. In this case, there are two helping intervals. Finally, when K is large – i.e., $K = 60$ in case (c) – both queues can be emptied before the start of the second demand surge. In this case, the two demand surges can be decomposed into two single-demand surge planning, and there are again two helping intervals.

We conclude this section by remarking that our optimal control policy has the same structure in both the single-surge and multi-surge settings. This is very appealing for practical implementations because one can implement the same policy but adjust the estimation of the G values as more information about the future arrival rates becomes available. We also note that even if the second



(a) $K = 10$ (b) $K = 30$ (c) $K = 60$

Figure 2.4: Optimal trajectory of the N-model when the duration between the two demand surges changes. ($h_1 = 1.5, h_2 = 1, \phi_{12} = 1, s_1 = 3, s_2 = 4, \mu_{11} = \mu_{22} = 0.25, \mu_{12} = 0.18, q_1(0) = q_2(0) = 0, \lambda_1(t) = 2\times\mathbf{1}\{0 \le t < 30\} + 0.5\times\mathbf{1}\{30 \le t < 30+K\} + 2\times\mathbf{1}\{30+K \le t < 60+K\} + 0.5\times\mathbf{1}\{t \ge 60+K\}, \lambda_2(t) = 0.6.$)

surge was not foreseen at the time of the first surge (i.e., the initial policy calculation assumes a single surge), our policy can quickly adapt to the information on the second surge when it becomes available. We substantiate this point with more numerical results in Section 2.6.3 when dealing with two demand surges and limited look-ahead windows.

2.4 Asymptotic Optimality

In this section, we translate the optimal control defined in Theorem 1 to the original stochastic system introduced in Section 2.2. We prove that the translated control is asymptotically optimal along a properly scaled sequence of stochastic systems even when there are prediction errors.

To specify the sequence of stochastic systems, we first elaborate on the planning horizon T in (2.1). Recall the planning horizon σ in the fluid control problem (2.3). The idea is to have a long enough time such that the system can get back to the normal state of operation by then. For tractability reasons, we adopt a deterministic planning horizon for the stochastic system. Given the initial state $X(0) = x$, we define the planning horizon $T(x)$ based on the fluid dynamic. Consider a fluid system in which pool 2 fully prioritizes class 1 as long as the class 1 queue is non-empty.

Specifically, we define $q^o(t) = (q_1^o(t), q_2^o(t))$ with $q^o(0) = x$ that satisfies

$$\begin{aligned}\dot{q}_1^o(t) &= (\lambda_1(t) - s_1\mu_{11} - s_2\mu_{12})1\{q_1^o(t) > 0\} + (\lambda_1(t) - s_1\mu_{11} - s_2\mu_{12})^+ 1\{q_1^o(t) = 0\},\\ \dot{q}_2^o(t) &= \lambda_2(t) - \left(s_2 - \frac{(\lambda_1(t) - s_1\mu_{11})^+}{\mu_{12}}\right)^+ \mu_{22}1\{q_1^o(t) = 0\}.\end{aligned} \tag{2.9}$$

Define

$$\tau_1^o(x) = \inf\left\{t \geq \kappa_1 : q_1^o(t) = 0\right\} \text{ and } \tau_2^o(x) = \inf\left\{t \geq \kappa_2 : q_2^o(t) = 0\right\},$$

which corresponds to the time by which the class i queue should be emptied if pool 2 prioritizes class 1 all the time. Then, we can define

$$T(x) = \max\{G_1^0(x_1), G_2^0(x_2), \tau_1^o(x), \tau_2^o(x)\} = \max\{G_1^0(x_1), \tau_2^o(x)\}.$$

Note that $T(x)$ can be interpreted as the time by which the fluid queue will be emptied under any reasonable scheduling policy. This is because the class 1 queue should be emptied by time $G_1^0(x_1)$ even if pool 2 does not serve class 1, and the class 2 queue should be emptied by time $\tau_2^o(x)$ even if pool 2 prioritizes class 1 all the time.

We now specify the setup to establish asymptotic optimality. Consider a sequence of systems indexed by n. The number of servers are fixed along the sequence. We speed up the time and scale down the space by n. Specifically, for the n-th system, the arrival rate at time t is $\lambda^n(t) := n\lambda(t)$, and the service rate is $n\mu_{ij}$. We use the superscript n to denote quantities related to the n-th system. For example, $X^n(t) = (X_1^n(t), X_2^n(t))$ denotes the number of customers in the n-th system; $Q^n(t) = (Q_1^n(t), Q_2^n(t))$ denotes the queue length; and $Z^n(t) = (Z_{11}^n(t), Z_{12}^n(t), Z_{22}^n(t))$ denotes the number of customers in service from each class in each pool at time t. For a given "base" starting state x, we assume that $X^n(0) = nx$. We also define the fluid-scaled queue length process as

$$\bar{X}^n(t) = \frac{1}{n}X^n(t).$$

A scheduling policy $\pi^n = \{\pi_t^n : t \geq 0\}$ for the n-th system maps the state of the system to

the allocation of servers – i.e., $Z^n(t) = \pi_t^n(X^n(t))$. The admissible controls are non-anticipative, but we have access to some estimated arrival rate $\Lambda^n(t)$. The server allocation policies satisfy the following conditions:

$$Z_{11}^n(t) + Z_{12}^n(t) \le X_1^n(t),\ Z_{22}^n(t) \le X_2^n(t),$$
$$Z_{11}^n(t) \le s_1,\ Z_{12}^n(t) + Z_{22}^n(t) \le s_2,\ Z^n(t) \in \mathbb{N}_0^3.$$

Let $\bar{Y}_{ij}^n(t) = \int_0^t Z_{ij}^n(s)\, ds$ be the total amount of time spent by pool j servers on class i customers up to time t. We add the scheduling policy as a superscript to the related processes when we want to emphasize the dependence of the system dynamics on the scheduling policy explicitly, e.g., $X^{n,\pi^n}(t)$ and $Z^{n,\pi^n}(t)$.

For the n-th system, the optimal scheduling problem is formulated as finding a policy that minimizes the cumulative holding and overflow costs over $[0, T(x)]$. In particular, we want to find a policy π^n that minimizes the following fluid-scaled objective:

$$\begin{aligned}\min_{\pi^n} \bar{V}^{n,\pi^n}(x) &= \mathbb{E}\left[\int_0^{T(x)} \left(\frac{h_1}{n} X_1^{n,\pi^n}(t) + \frac{h_2}{n} X_2^{n,\pi^n}(t) + \phi_{12} Z_{12}^{n,\pi^n}(t)\right) dt \middle| X^n(0) = nx\right] \\ &= \mathbb{E}\left[\int_0^{T(x)} \left(h_1 \bar{X}_1^{n,\pi^n}(t) + h_2 \bar{X}_2^{n,\pi^n}(t)\right) dt + \phi_{12} \bar{Y}_{12}^{n,\pi^n}(T(x)) \middle| X^n(0) = nx\right].\end{aligned}$$

Note that the holding costs and the overflow cost are scaled differently in $\bar{V}^{n,\pi^n}(x)$ to have a meaningful comparison. Specifically, the holding costs are scaled by n – i.e., $h_i^n = h_i/n$ – while the overflow cost is unscaled. A similar scaling is used in [51].

We next translate the optimal fluid policy to a sequence of policies for the corresponding stochastic systems. Recall that for the n-th system, the true arrival rate for class i follows $\lambda_i^n(t) = n\lambda_i(t)$. We assume the corresponding estimated arrival rate takes the form $\Lambda_i^n(t) = n\lambda_i(t) + E_i^n(t)$, where $E_i^n(\cdot)$ is the estimation error term. We impose the following assumptions on $E_i^n(\cdot)$:

Assumption 3. *$E_i^n(\cdot)$ is a stochastic process satisfying $E_i^n(\cdot)/n \to 0$ u.o.c. almost surely as*

$n \to \infty$, *i.e.,*

$$\mathbb{P}(E_i^n(\cdot)/n \to 0 \text{ } u.o.c. \text{ } as \text{ } n \to \infty) = 1.$$

In addition, for large enough n, $\Lambda_i^n(\cdot)$ *satisfies items 1 and 3 in Assumption 1.*

Under Assumption 3, the uncertainty of the arrival rate is of a smaller order than the arrival rate itself. This is a common assumption in the literature, see, for example, [52, 7].

For the n-th system, we use the look-ahead function based on the estimated arrival rate:

$$\tilde{G}_{i,n}^t(x) = \inf\left\{\Delta \geq 0 : \int_t^{t+\Delta} (s_i n\mu_{ii} - \Lambda_i^n(s))\, ds = x\right\}$$

for $x > 0$. The scheduling policy $\{\tilde{v}^n\}_{n\geq 1}$ is defined as follows. For the n-th system: pool 1 serves class 1 customers as much as possible. When $h_1\mu_{12} > h_2\mu_{22}$, at time t, if

$$h_1\mu_{12}\tilde{G}_{1,n}^t(X_1^n(t)) - \phi_{12} > h_2\mu_{22}\tilde{G}_{2,n}^t(X_2^n(t)),$$

pool 2 gives preemptive priority to class 1; otherwise, pool 2 serves class 2 only. When $h_1\mu_{12} \leq h_2\mu_{22}$, at time t, if

$$h_1\mu_{12}\tilde{G}_{1,n}^t(X_1^n(t)) - \phi_{12} > 0,$$

pool 2 serves both classes but gives preemptive priority to class 2; otherwise, pool 2 serves class 2 only. The following theorem shows that $\{\tilde{v}^n\}_{n\geq 1}$ is asymptotically optimal. Let $\bar{V}^*(x)$ denote the optimal objective value of the corresponding fluid control problem (2.3).

Theorem 3 (Asymptotic optimality)**.** *Under Assumptions 1 and 3, for any sequence of admissible controls* $\{\pi^n\}_{n\geq 1}$*,*

$$\liminf_{n\to\infty} \bar{V}^{n,\pi^n}(x) \geq \bar{V}^*(x).$$

For the sequence of systems under policy $\{\tilde{v}^n\}_{n\geq 1}$*,*

$$\lim_{n\to\infty} \bar{V}^{n,\tilde{v}^n}(x) = \bar{V}^*(x).$$

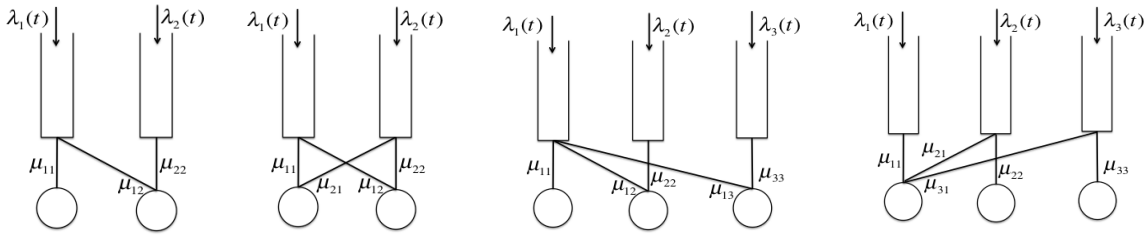
The proof of Theorem 3 is in Appendix A.5. To incorporate the prediction error in the asymptotic optimality result, we leverage the continuity properties of the look-ahead function G_i^t. This theorem suggests that when applied to the stochastic systems, the two-stage index-based look-ahead policy achieves near-optimal performance when the initial queue and/or the demand surge is large. Note that our asymptotic optimality result requires the estimation error to be of a smaller order than the arrival rate (Assumption 3). This indicates that if the estimation error is small relative to the actual arrival rate, the proposed policy achieves near-optimal performance. In Section 2.6.3, we go beyond this theoretical result and numerically investigate the performance of our policy with more general forms of prediction errors, e.g., when the estimation error is relatively large or when we only have access to a limited look-head time window. Numerical results show that even though the performance of our algorithm deteriorates as the prediction accuracy decays, it performs competitively compared to benchmark policies that are agnostic to future arrival information. We note that the index structure of our policy has some built-in resilience to perturbations. In particular, as long as the estimation errors do not reverse the order of the second-stage indices, the same policy will be implemented in the stochastic system at given time t.

2.5 Beyond the N-Model: Heuristics for General Systems

In this section, we study three more general models beyond the N-model, which helps us design a heuristic policy for general multi-class multi-pool systems. The three models are i) the X-model; ii) the many-help-one extended N-model (exN1); and iii) the one-helps-many extended N-model (exN2). See Figure 2.5 for a pictorial illustration of these models together with the N-model. Note that the exN2-model also covers the commonly-studied M-model as a special case when we set the holding cost $h_1 = 0$.

For the three models, we focus on the fluid optimal control analysis. Moreover, when presenting the results, we focus on emphasizing the key difference between these models and the N-model. In particular, by comparing the X-model with the N-model, we highlight the "unexpected" benefit from cross-training. By comparing the extended N-models with the two-class N-model, we

generate insights into how the policy changes when facing multiple classes of customers and multiple pools of servers. These insights lead us propose a heuristic policy for general multi-class multi-pool systems presented in Section 2.5.3.



(a) N-model (b) X-model (c) exN1-model (d) exN2-model

Figure 2.5: Queueing models with partial flexibility

General notation for fluid control. Consider I classes of customers and I server pools. We assume class i fluid flows into the system at rate $\lambda_i(t)$ and flows out of service at rate $\sum_j \mu_{ij} z_{ij}(t)$, where μ_{ij} is the service rate of pool j servers working on class i jobs, and $z_{ij}(t)$ is a positive real number denoting the service capacity from pool j allocated to serve class i fluid at time t. Note that if $\mu_{ij} = 0$, class i customers and pool j servers are not compatible. Let $q_i(t) \in [0, \infty)$ denote the fluid amount of class i customers in the system at time t. Then,

$$\dot{q}_i(t) = \lambda_i(t) - \sum_{j=1}^{I} \mu_{ij} z_{ij}(t).$$

A fluid scheduling policy π specifies the service capacity allocation

$$z(t) = (z_{ij}(t) : i, j = 1, \cdots, I).$$

We require the control to satisfy the following capacity constraints:

$$z_{ij}(t) \geq 0, i, j = 1, \cdots, I, \ \sum_i z_{ij}(t) \leq s_j, j = 1, \cdots, I$$

and state constraints:

$$\dot{q}_i(t) \geq 0 \text{ whenever } q_i(t) = 0, i = 1, \cdots, I.$$

Similar as the N-model, we allow the policies to utilize future arrival rates.

For arrival rates, we consider the scenario in which each class may experience an initial demand surge that lasts for a certain amount of time before the demand returns to normal; the extension to multiple surges can be done similarly as in Section 2.3.4. Let κ_i denote the demand surge period for class i and $\bar{\kappa} = \max_{1 \leq i \leq I} \kappa_i$. Based on a set of assumptions that are similar to Assumption 1 (see Assumption 5 in Appendix A.1 for a full specification of the assumptions for general multi-class multi-pool systems), we define

$$\sigma = \inf \left\{ t \geq \bar{\kappa} : \sum_i q_i(t) = 0 \right\},$$

which can be interpreted as the time to fully absorb the demand surge. Fluid waiting in the system incurs a cost of h_i per unit job per unit time. In addition, routing fluid from class i to pool j incurs an overflow cost of ϕ_{ij} per unit job per unit time, with $\phi_{ii} = 0$ by convention. Then, the fluid optimal control problem takes the form:

$$\begin{aligned}
&\min_z \int_0^\sigma \sum_{i=1}^I h_i q_i(t) + \sum_{i=1}^I \sum_{j=1}^I \phi_{ij} z_{ij}(t) dt \\
&\text{s.t. } \dot{q}_i(t) = \lambda_i(t) - \sum_{j=1}^I \mu_{ij} z_{ij}(t), i = 1, \cdots, I \\
&\quad q_i(t) \geq 0, i = 1, \cdots, I \\
&\quad z_{ij}(t) \geq 0, i, j = 1, \cdots, I, \quad \sum_{i=1}^I z_{ij}(t) \leq s_j, j = 1, \cdots, I.
\end{aligned} \tag{2.10}$$

Similar to before, we also define the function $G_i^t : \mathbb{R}_0^+ \to \mathbb{R}_0^+$ as

$$G_i^t(x) = \inf \left\{ u \geq 0 : \int_0^u (s_i \mu_{ii} - \lambda_i(s)) \, ds = x \right\}$$

for $i = 1, \ldots, I$, for $x > 0$, and $G_i^t(0) = \lim_{x \downarrow 0} G_i^t(x)$. Note that if $\lambda_i(t) \equiv \lambda_i$, $G_i^t(x) = \frac{x}{s_i \mu_{ii} - \lambda_i}$ for all $t \geq 0$. For a fixed t, the function $G_i^t(x)$ is continuous and strictly increasing in x.

Summary of the optimal policy structure. When dealing with more-general models, the optimal policy follows a similar two-stage structure as in the N-model. That is, in the first stage we decide whether a certain pool is going to fully prioritize another non-primary class or just provide partial help based on the $h\mu$ index. In the second stage we decide how long the full- or partial-help lasts. The main difference from the N-model is that, when deciding how long the "help" will last in the second stage, we compare the time it takes to empty the queues not only using their primary resources, but also taking into account the help they may receive from other pools or the help their primary pools may provide to other classes. This idea will be made more precise in the subsequent sections.

2.5.1 X-Model

The X-model has a similar network structure except that helping can happen in both ways. In particular, pool 1 can serve class 2 at rate $\mu_{21} > 0$ while pool 2 can serve class 1 at rate $\mu_{12} > 0$. We assume that $\mu_{11} > \mu_{12}$ and $\mu_{22} > \mu_{21}$, i.e., primary pool servers are preferred. Following the development of the N-model, we first compare the $h\mu$ index. Without loss of generality, we consider two possible cases:

I. $h_1\mu_{12} > h_2\mu_{22}$, which implies that $h_2\mu_{21} < h_1\mu_{11}$. In this case, pool 2 gives priority to class 1 when class 1 has a large enough backlog compared to class 2. When pool 1 empties the class 1 queue, it may provide partial help to class 2 if class 2 has a large enough backlog.

II. $h_1\mu_{12} < h_2\mu_{22}$ and $h_2\mu_{21} < h_1\mu_{11}$. In this case, when pool i, $i = 1, 2$, empties its own class, it may provide partial help to the other class if the other class has a large enough backlog.

The key difference between the X-model and the N-model comes up in Case I when deciding how long pool 2 will help class 1. In the X-model, because pool 1 can later help back class 2 – i.e., pool 1 can provide partial help to class 2 when the class 1 queue empties – the period during

which pool 2 prioritizes class 1 can be longer than the full helping period in an otherwise identical N-model.

Optimal policy.

To fully characterize the full helping period in Case I for the X-model, we define $P^t(q(t))$ as the length of the partial helping period for pool 1 to class 2:

$$P^t(q) = \inf\left\{u \geq 0 : h_2\mu_{21}G_2^{t+G_1^t(q_1)+u}(\tilde{q}_2(t + G_1^t(q_1) + u)) \leq \phi_{21}\right\},$$

where for $\tilde{q}$, its dynamic follows: $\tilde{q}(t) = q$; for $s \in (t, t + G_1^t(q_1(t)))$, pool 1 serves class 1 only; for $s \geq t + G_1^t(q_1(t))$, pool 1 provides partial help to class 2. We also define $\bar{G}^t_{X,2}(t, q(t))$ as the time it takes to empty queue 2 when taking the partial help from pool 1 into account:

$$\bar{G}^t_{X,2}(q(t)) = G_2^t(q_2(t))1\{P^t(q(t)) = 0\} + \left(G_1^t(q_1(t)) + P^t(q(t)) + \frac{\phi_{21}}{h_2\mu_{21}}\right)1\{P^t(q(t)) > 0\}.$$

The following theorem characterizes the optimal scheduling policy for the X-model.

Theorem 4 (Optimal control policy in X-model)**.** *For the X-model, under Assumption 5, the optimal control for* (2.10) *takes the following form.*

I. When $h_1\mu_{12} > h_2\mu_{22}$, pool 1 prioritizes class 1.

ia. If

$$G_1^t(q_1(t)) = 0 \text{ and } h_2\mu_{21}G_2^t(q_2(t)) > \phi_{21},$$

pool 1 provides partial help to class 2.

ib. Otherwise, pool 1 serves class 1 only.

For pool 2,

iia. If

$$h_1\mu_{12}G_1^t(q_1(t)) > h_2\mu_{22}\bar{G}_{X,2}^t(q(t)) - h_2\mu_{22}\frac{\mu_{12}\mu_{21}}{\mu_{11}\mu_{22}}P^t(q(t)) + \phi_{12}, \tag{2.11}$$

pool 2 gives priority to class 1.

iib. Otherwise, pool 2 serves class 2 only.

II. When $h_1\mu_{12} < h_2\mu_{22}$ and $h_2\mu_{21} < h_1\mu_{11}$, pool i prioritizes class i. If

$$G_i^t(q_i(t)) = 0 \text{ and } h_j\mu_{ji}G_j^t(q_j(t)) > \phi_{ji},\ j \neq i,$$

pool i provides partial help to class j; otherwise, pool i serves class i only.

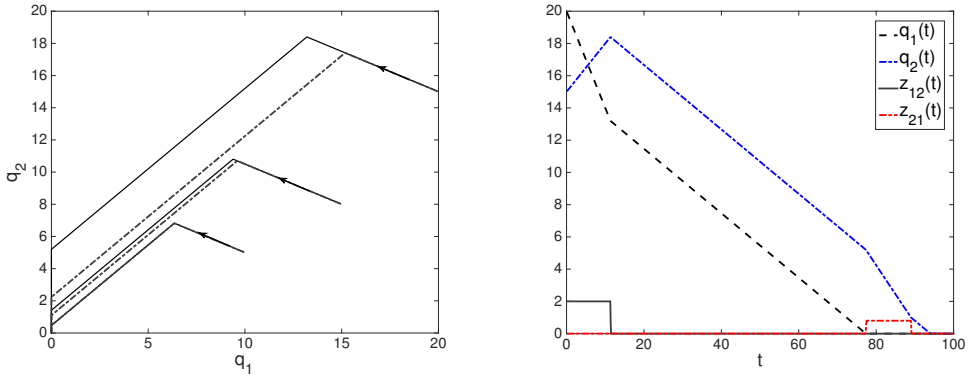
Similar to the optimal policy for the N-model, the optimal control for the X-model characterized in Theorem 4 also takes the future arrival rate information into account, and the policy has a two-stage index structure. The main difference, though, is in Case I.iia. As $\bar{G}_{X,2}^t(q(t)) \leq G_2^t(q_2(t))$,

$$h_2\mu_{22}\bar{G}_{X,2}^t(q(t)) - h_2\mu_{22}\frac{\mu_{12}\mu_{21}}{\mu_{11}\mu_{22}}P^t(q(t)) + \phi_{12} \leq h_2\mu_{22}G_2^t(q_2(t)) + \phi_{12}.$$

This implies that in the X-model, because pool 1 can help back class 2 later, pool 2 may provide more help to class 1 initially than in the N-model.

To further illustrate this point, we provide some numerical examples with time-homogeneous arrival rates in Figure 2.6. In Figure 2.6(a), we consider an N-model and an X-model with the same parameters and compare the optimal fluid trajectories $(q_1(t), q_2(t))$ with different initial values. Since we start the system with $q_1(0) \gg q_2(0)$, at the beginning, pool 2 prioritizes class 1. In particular, we note that, initially, $q_1(t)$ is decreasing and $q_2(t)$ is increasing. At some points (the wedges in the figure), pool 2 switches priority from class 1 to class 2. Comparing the optimal trajectory of the N-model (dashed) to the optimal trajectory of the X-model (solid), pool 2 in the X-model provides more help to class 1 than in the N-model, i.e., the solid line switches priority

to class 2 later than the dashed line when $q(0) = (20, 15)$ or $(15, 10)$. Figure 2.6(b) plots the optimal fluid trajectory of the X-model starting from $(20, 15)$. We note that at the beginning, pool 2 prioritizes class 1 – i.e., $z_{12}(t) = 2$ for $t \leq 11.5$, while later, after $q_1(t)$ hits zero, pool 1 provides partial help back to class 2 – i.e., $z_{21}(t) = 0.8$ for $t \in [77.4, 89.0]$.



(a) N-model (dashed) versus X-model (solid) (b) X-model with $q(0) = (20, 15)$

Figure 2.6: Optimal trajectory of the N-model versus the X-model. ($s_1 = s_2 = 2$, $\mu_{11} = \mu_{22} = 0.25$, $\mu_{12} = 0.2$, $\phi_{12} = 1$, $\lambda_1 = \lambda_2 = 0.3$, $h_1 = 1.5$, $h_2 = 1$. For the X-model, $\mu_{21} = 0.2$, $\phi_{21} = 1$.)

Managerial implication: value of cross-training.

The N-model and the X-model differ in whether pool 1 is cross-trained to help class 2 (pool 2 can help class 1 in both models). The X-model has the obvious advantage that when class 2 is overloaded, pool 1 can help class 2 to alleviate the demand surge and bring the system back to normal faster than in the N-model. Meanwhile, somewhat surprisingly, even when only class 1 is experiencing a demand surge, the extra flexibility in the X-model is beneficial. In particular, as we explained in Section 2.5.1, because pool 1 can later help back class 2 in the X-model, pool 2 can provide more help to class 1 (prioritize class 1 for a longer period of time) in the initial stage. This helps reduce the class 1 congestion faster in the X-model than in the N-model. To demonstrate the latter point, Table 2.1 compares the time to empty queue 1 and the time to empty queue 2 (which is also the time to empty the whole system) under the optimal control for the X-model versus the N-model. We vary the level of demand surge experienced by class 1, while class 2 does not experience any demand surge. Note that in all cases, not only is the X-model able to empty

the class 1 queue faster than an otherwise identical N-model, but also it empties the system (both queues) faster than the N-model does.

λ_H	2	4	6
Time to empty queue 1			
X-model	50.0	108.1	166.4
N-model	51.4	117.8	184.2
Time to empty queue 2			
X-model	59.2	138.2	217.2
N-model	59.6	140.8	222.1

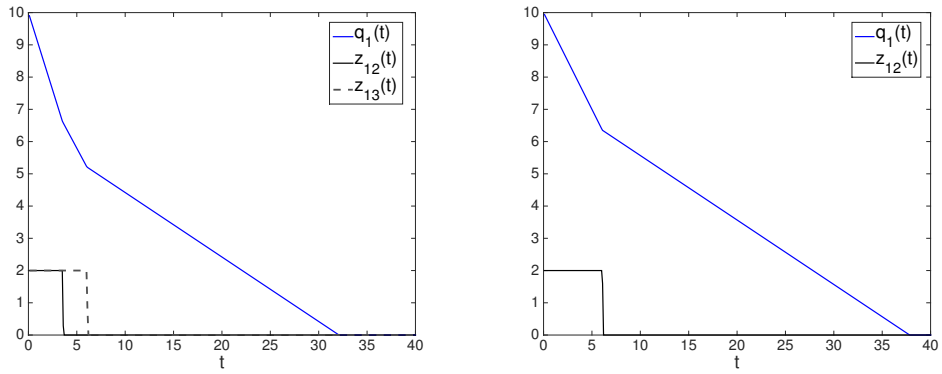
Table 2.1: Compare the N-model and the X-model under different levels of demand surge for class 1 ($s_1 = 3$, $s_2 = 4$, $\mu_{11} = \mu_{22} = 0.25$, $\mu_{12} = 0.18$, $h_1 = 2$, $h_2 = 1$, $\phi_{12} = 1$, $\lambda_1(t) = \lambda_H \times \mathbf{1}\{0 \leq t \leq 20\} + 0.5 \times \mathbf{1}\{t > 20\}$, $\lambda_2(t) = 0.6$, $q_1(0) = 10, q_2(0) = 0$. For the X-model, $\mu_{21} = 0.18$ and $\phi_{21} = 1$.)

2.5.2 Extended N-models

In this section, we discuss the main insights from the optimal policies for the two extended N-models: many-help-one (exN1) and one-helps-many (exN2) models. To keep the discussion concise, we delay the full characterization of the optimal policies to Appendix A.1.

For the exN1 model, the optimal policy still has a two-stage index-based look-ahead structure (see Theorem 9 in Appendix A.1). In the first stage, we decide which class to prioritize based on the $h\mu$ index. In the second stage, we decide how long the full or partial help will last by taking future arrival rate information into account. The main difference between the exN1 model and the N model lies in the second stage. Consider the scenario where $h_1\mu_{12} > h_2\mu_{22}$ and $h_1\mu_{13} > h_3\mu_{33}$, i.e., both pool 2 and pool 3 will give strict priority to class 1 if the class 1 queue is large enough. When pool 2 determines how long it will help class 1, it also needs to take into account the help that class 1 can receive from pool 3, in which case pool 2 may provide less help to class 1 than in a similar N-model. To demonstrate this, Figure 2.7 compares the optimal trajectory of an exN1 model (a) with the optimal trajectory of a similar N-model (b). In particular, the two models share the same parameters for the first two classes. The only difference is that the exN1 model has an extra class, class 3, and an extra server pool, pool 3 (see the caption of Figure 2.7 for more details).

For the exN1 model, we observe that both pools provide full help to class 1 at the beginning. Pool 2 stops helping class 1 at $t = 3.5$ in the exN1 model. In contrast, pool 2 stops helping class 1 at $t = 6.1$ in the N-model. This is because in the exN1 model, class 1 can also get help from pool 3, and when pool 2 decides how much to help class 1, it also takes the extra help from pool 3 into account. Lastly, we note that with the extra help from pool 3, the exN1 model is able to empty the class 1 queue faster than the N-model can.



(a) exN1 model (b) N-model

Figure 2.7: Optimal trajectory of the exN1 model versus the N-model. ($s_1 = s_2 = 2$, $\lambda_1 = \lambda_2 = 0.3$, $\mu_{11} = \mu_{22} = 0.25$, $\mu_{12} = 0.2$, $\phi_{12} = 1$, $h_1 = 1.5$, $h_2 = 1$, $q_1(0) = 10$, $q_2(0) = 5$. For the exN1 model, $s_3 = 2$, $\lambda_3 = 0.3$, $\mu_{33} = 0.25$, $\mu_{13} = 0.18$,$\phi_{13} = 1$, $h_3 = 1$, and $q_3(0) = 3$.)

For the exN2 model, the optimal policy again has a two-stage index-based look-ahead structure (see Theorem 10 in Appendix A.1). The key difference between the exN2 model and the N-model lies in the second stage. Consider the scenario where $h_2\mu_{21} > h_3\mu_{31} > h_1\mu_{11}$, i.e., pool 1 prioritizes classes 2 and 3 when there are large enough backlogs in these two classes compared to class 1. When deciding between classes 2 and 3, class 2 enjoys a higher priority. In the second stage, when pool 1 determines how long it will help class 2, it also needs to consider the help it can provide to class 3, in which case pool 1 may provide less help to class 2 than in a similar N-model.

2.5.3 A Heuristic Two-Stage Index-Based Policy for Multi-class Multi-pool Systems

Based on the results from the X-model and the two extended N-models, we observe that the structure of the optimal policy remains similar to that of the N-model. Based on this observa-

tion, we propose the following two-stage index-based look-ahead policy for more-general I-by-I networks.

First stage. Denote the set of classes that pool j can serve as $\mathcal{I}_j$, which is sorted by the $h\mu$-index. That is, for class $k(i)$ in the ith position in set $\mathcal{I}_j$, its $h\mu$-index is larger than that of class $k(i+1)$ in the $(i+1)$th position – i.e., $h_{k(i)}\mu_{k(i),j} > h_{k(i+1)}\mu_{k(i+1),j}$. The primary class j is in the set $\mathcal{I}_j$, and we denote its position as ℓ_j.

- For class $k(i) \in \mathcal{I}_j$ with $i < \ell_j$, pool j provides full help (strict priority) to class $k(i)$ if the help is initiated according to the second-stage criteria.

- For class $k(i) \in \mathcal{I}_j$ with $i > \ell_j$, pool j provides partial help (help only when there is extra capacity after serving its own class) if the help is initiated according to the second-stage criteria.

Second stage. At any time t, for each pool j, we decide which class in $\mathcal{I}_j$ it should help according to the following criteria. Set a tuning parameter $\theta > 0$.

- For classes $k(i)$'s with $i < \ell_j$, let class $k(i^*)$ be the first class for which

$$\theta h_{k(i^*)}\mu_{k(i^*),j} G^t_{k(i^*)}(q_{k(i^*)}(t)) - \phi_{k(i^*),j} > h_j \mu_{jj} G^t_j(q_j(t)). \tag{2.12}$$

  Pool j provides full help to class $k(i^*)$ if there exists such $k(i^*)$.

- If none of the full helping is initiated and $q_j(t) > 0$, pool j serves class j only;

- If none of the full helping is initiated and $q_j(t) = 0$, for classes $k(i)$'s with $i > \ell_j$, let $k(i^*)$ be the first class for which

$$\theta h_{k(i^*)}\mu_{k(i^*),j} G^t_{k(i^*)}(q_{k(i^*)}(t)) > \phi_{k(i),j}. \tag{2.13}$$

  Pool j provides partial help to class $k(i^*)$ if there exists such $k(i^*)$.

To explain the rationale of the tuning parameter θ, we note that from the analysis of the X-model and the extended N-models, depending on the system architecture, we may need to modify $G_i^t(q_i(t))$'s to take into account the help that pool i can provide to other classes or the help class i can receive from other pools. This can be partially captured by the tuning parameter θ. When $\theta > 1$, we are doing more aggressive overflow than in the N-model; when $\theta = 1$, it is equivalent to the optimal N-model policy; when $\theta < 1$, we are doing more conservative overflow than in the N-model. We show, via extensive numerical experiments in Section 2.6, that the performance of the heuristic policy is robust for θ close to 1, while a slight tuning down – i.e., setting $\theta = 0.8$ – leads to comparable or, in some cases, better performance than $\theta = 1$. Also note that the robust performance for θ near 1 suggests that our policy is resilient to prediction errors and limited look-ahead windows, which essentially add small perturbations to $G_i^t(q_i(t))$'s.

2.6 Numerical Experiments for General Stochastic Networks

The optimal control policies that we derived in prior sections are based on deterministic fluid models. In this section, we study the performance of our derived policy – namely, the two-stage index-based heuristic policy specified in Section 2.5.3, in the original stochastic systems via simulation. We refer to our derived policy as the look-ahead policy, and we compare its performance to that of several well-established benchmark policies.

Through numerical experiments, we demonstrate that, in the face of demand surges and time-nonstationary arrival rates, the performance of our look-ahead policy, even using estimated arrival rates that have prediction errors or with a limited look-ahead time window, is superior to that of the $c\mu$ rule or the maximum pressure policy or their adapted versions that account for the overflow costs. In particular, we show that without considering future arrival rates, the adapted $c\mu$ rule or maximum pressure policy can perform significantly worse than our policy with up to 50% relative cost difference. In addition, the actual format to incorporate the overflow costs is highly nontrivial as we experiment with different adapted version of the modified $c\mu$ rule or the maximum pressure policy: the cost can be arbitrarily bad when using the wrong format, which confirms the necessity

of rigorously deriving optimal routing polices in the presence of demand surge and overflow cost. Moreover, we go beyond the models studied analytically and substantiate our theoretical development with simulation results in more-general networks. In Section 2.6.2, we consider 5-by-5 networks, and the simulation results show that our policy still performs remarkably well in these more-general networks. The numerical results suggest that the insights generated from our fluid analysis of parsimonious models are robust and useful when designing routing policies in more complex systems.

To calibrate the simulation model, we consider settings motivated by hospital inpatient flow management [53]. That is, in the multi-class, multi-pool parallel processing network, each class corresponds to patients from a medical specialty, and each server pool corresponds to an inpatient ward or several wards that are dedicated to a medical specialty. Unless otherwise specified, we assume that each pool has a capacity of 20 – i.e., $s_i = 20$ for pool i, corresponding to 20 inpatient beds. The primary service rates are $\mu_{ii} = 0.25$ for each class i, corresponding to an average service time (length-of-stay) of four days. Moreover, we incorporate service slowdown – i.e., longer length-of-stay when the patient is placed in a bed in a non-primary ward [54, 55]. We assume that the overflow service rate is $\mu_{ij} = 0.2$ for $i \neq j$, corresponding to an average service time of five days. In subsequent sections, we start by presenting results for our main model – the N-model – and then results in more complicated networks.

2.6.1 Performance comparison in N-model: Value of proactive routing

The baseline arrival rate setting that we test in the N-model follows

$$\lambda_1(t) = \begin{cases} 8, & t < 40 \\ 4, & t \geq 40, \end{cases}$$

and the arrival rate for class 2 is $\lambda_2 = 3$ patients per day. That is, class 1 experiences a demand surge lasting $\kappa_1 = 40$ days, while class 2 does not experience a demand surge. The initial state is set as $\big(X_1(0), X_2(0)\big) = (60, 70)$.

Beyond this baseline setting, we also test a large combination of arrival rate settings by varying the following parameters: (i) the peak arrival rate of class 1 (max $\lambda_1(t)$) takes values of 6, 8, and 10; (ii) the arrival rate of class 2 (the constant λ_2) takes values of 3, 3.5, 4, and 4.5; (iii) the surge duration κ varies between 20 and 100; (iv) the initial state for class 1, $X_1(0)$, takes values of 40, 60, and 80; and (v) the initial state for class 2, $X_2(0)$, takes values of 70, 90, 110 and 130. These parameter combinations lead to different levels of system congestion levels. In general, the system dynamics are closer to the fluid limit when the system is more congested.

Under a given arrival setting, we fix the holding costs $h_1 = 1.5, h_2 = 1$ and vary the overflow cost – namely, (i) $\phi_{12} = 2$; (ii) $\phi_{12} = 10$; and (iii) $\phi_{12} = 25$. The holding costs correspond to Case I of Theorem 1, where pool 2 may provide full help to class 1. We choose to focus on this case since the resulting policy is less trivial than the partial helping case. In particular, if the help is not exercised properly, it can lead to both a high overflow cost and a high holding cost. We simulate 10^4 replications for each scenario (policy and system) to estimate the expected cost and the corresponding standard error. Each replication contains 250 days. A common sequence of random numbers is used when comparing different policies.

Benchmark policies

We compare five policies in the stochastic N-model: (i) our look-ahead policy (Look-ahead); (ii) the classic $c\mu$-rule (Cmu); (iii) the classic maximum pressure policy (MaxPres); (iv) the modified $c\mu$-rule which takes the overflow cost into account (ModCmu); and (v) the modified maximum pressure policy which takes the overflow cost into account (ModMaxP). We provide more details about policies (iv)-(v) next.

For the **modified $c\mu$ rule**, we prioritize different classes according to the following index (from high to low): $h_i\mu_{ij} - \phi_{ij}\mu_{ij}$. This index adjusts the original $c\mu$-index, $h_i\mu_{ij}$, by subtracting the weighted overflow cost $\phi_{ij}\mu_{ij}$. The intuition here is to maximize the instantaneous cost reduction rate under the preemptive service setting (with μ_{ij} being the chance of clearing a customer).

Similarly, for the **modified maximum pressure policy**, we prioritize different classes accord-

ing to the following index (from high to low): $h_i X_i(t)\mu_{ij} - \phi_{ij}\mu_{ij}$, where $X_i(t)$ is the number of class i customers in the system at time t. Compared to the original maximum pressure policy, we adjust the index $h_i X_i(t)\mu_{ij}$ with the weighted overflow cost $\phi_{ij}\mu_{ij}$. As discussed in Section 2.3, comparing our proposed policy with this modified maximum pressure policy, the main difference is that the modified maximum pressure policy weights $h_i\mu_{ij}$ by $X_i(t)$, while our policy weights $h_i\mu_{ij}$ by $G_i^t(X_i(t))$, which takes future arrival rate information into account.

It is worth noting that the adjustment to the $c\mu$ and maximum pressure policies is heuristic. It is possible to motivate other heuristics such as subtracting ϕ_{ij} or ϕ_{ij}/μ_{ij} instead of $\phi_{ij}\mu_{ij}$. However, if we use ϕ_{ij} or ϕ_{ij}/μ_{ij}, the index would be sensitive to the time unit we choose, because these terms will scale differently with μ_{ij} than $h_i\mu_{ij}$. In this case, by either choosing seconds or hours to be the time unit, the resulting policy could vary drastically from no overflow at all to full-sharing. This highlights the necessity of properly comparing the overflow cost with the holding cost when routing in the face of demand surges – which is the main goal of this work.

Robust performance

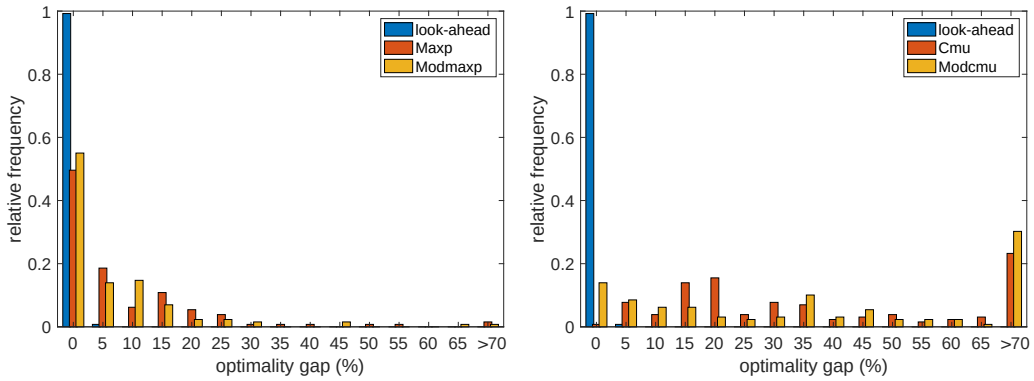
Table 2.2 shows the cost comparison among the five policies in the baseline setting. Our proposed look-ahead policy performs significantly better than the $c\mu$ and the modified $c\mu$ rules. The maximum pressure policy and its modified version perform better than the $c\mu$ rules but have a larger gap from our policy when ϕ is large, e.g., the gap is 15% when $\phi = 25$.

To have a more complete picture of the our policy's performance versus that of other benchmarks beyond just the baseline setting, Figure 2.8 plots a histogram of the optimality gap among all the tested combinations of arrival rates, initial states, and overflow costs, as described above. The optimality gap is defined as the relative cost difference between the investigated policy and the best-performed policy in the corresponding parameter setting. It is clear from the figure that our policy always performs best or near-best (the optimality gap is within 5%) among all tested parameter combinations. This demonstrates the robustness of our policy, which is an appealing feature in application. In contrast, other policies can perform well in some settings but poorly in

		Look-ahead	MaxPres	ModMaxP	Cmu	ModCmu
$\phi = 2$	Holding	1.09	1.10	1.10	1.28	2.75
	Overflow	0.14	0.13	0.13	0.17	0.00
	Total	1.23	1.23	1.23	1.45	2.75
	SE	0.003	0.003	0.003	0.004	0.008
$\phi = 10$	Holding	1.10	1.10	1.11	1.28	2.75
	Overflow	0.56	0.63	0.62	0.86	0.00
	Total	1.67	1.74	1.73	2.14	2.75
	SE	0.004	0.004	0.004	0.005	0.008
$\phi = 25$	Holding	1.28	1.10	1.12	1.28	2.75
	Overflow	1.00	1.58	1.50	2.14	0.00
	Total	2.28	2.68	2.62	3.42	2.75
	SE	0.005	0.005	0.005	0.007	0.008

Table 2.2: Expected total cost for the baseline N-model under different routing policies. The costs shown in the table are in units of 10^4. "SE" stands for the standard error for the corresponding average total cost (holding + overflow). (Parameter setting: $h = (1.5, 1)$, $s_i = 20$ and $\mu_{ii} = 0.25$, for $i = 1, 2$, $\mu_{12} = 0.2$, $\phi_{12} = \phi$, $\lambda_2(t) = 3$, $\lambda_1(t) = 8 \times \mathbf{1}\{t < 40\} + 4 \times \mathbf{1}\{t \geq 40\}$ and $X(0) = (60, 70)$.)

others. For example, the modified $c\mu$ policy tends to perform well when the surge arrival rate of class 1 is smaller (e.g., 6) and not much overflow is needed; however, it results in significantly worse performance when the surge arrival rate is large, and/or the initial queue length of class 1 is large. Similarly, the two maximum pressure policies tend to have a better performance than the two $c\mu$-based policies when the system is congested. However, their performance deteriorates when (i) the surge period is short (e.g., 20), but the initial queue length is high, (ii) the surge period is long, but the initial queue length is low, or (iii) pool 2 has less slackness in general. This is because the maximum pressure policy will help class 1 when its current queue length is large enough compared to the class 2 queue without looking into the future – this could be unnecessary (as in (i)), or too late (as in (ii)), or hurting class 2 too much (as in (iii)). Overall, the maximum pressure policies are reactive, while our policy is more proactive; that is, our policy can provide help to class 1 before the queue builds up in anticipation of the demand surge and can also end help earlier in anticipation of the drop in demand when the surge is over. Given the modified maximum pressure policy is the most competitive benchmark, we next take a further investigation into it.



(a) Our policy vs two max-pressure based policies

(b) Our policy vs two $c\mu$ based policies

Figure 2.8: Histogram of the optimality gap. Parameter setting combinations are described in the main text.

Reactive versus proactive: value of future information

The modified maximum pressure policy is the best-performing benchmark policy of Table 2.2, i.e., it performs better than other benchmark policies in most tested scenarios in Figure 2.8. Hence, it may appear that knowing the future arrival information is not as beneficial as one would expect, and that policies that do not take that information into account, such as the maximum pressure policy, may be sufficient. However, a closer investigation into different arrival settings reveals that this is not true – not considering future arrival information in a time-nonstationary setting can result in much worse performance. For illustration, consider the following arrival rate setting as an example:

$$\lambda_1(t) = \begin{cases} 8, & t < 40 \\ 1, & t \geq 40 \end{cases} \quad \text{and} \quad \lambda_2(t) = \begin{cases} 3, & t < 40 \\ 4.5, & t \geq 40. \end{cases}$$

That is, the arrival rate of class 1 drops sharply once the demand surge is over, while the arrival rate of class 2 increases slightly at the same time. All other parameters are the same as in the baseline setting. In Figure 2.9 we compare two sample paths when $\phi = 2$, one under our policy and the other under the modified maximum pressure policy.

We see that under the modified maximum pressure policy, pool 2 helps class 1 throughout the demand surge (till $t = 40$). This is because class 1 has a large queue and the policy is reacting to

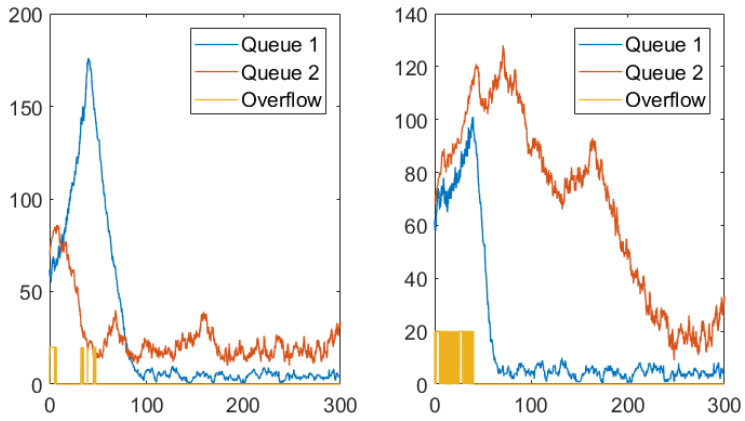


Figure 2.9: Sample path comparison between our proposed policy (left) and the modified maximum pressure policy (right). ($\lambda_1(t) = 8 \times \mathbf{1}\{t < 40\} + 1 \times \mathbf{1}\{t \geq 40\}$, $\lambda_2(t) = 3 \times \mathbf{1}\{t < 40\} + 4.5 \times \mathbf{1}\{t \geq 40\}$. Other parameters are the same as the baseline with $\phi = 2$.)

this. In contrast, under our policy, pool 2 proactively stops helping class 1 much sooner (around $t = 8$). This is because our policy anticipates that the class 1 arrival rate will soon drop to 1, so that not much help is necessary. This early-stopping also takes into account the fact that the class 2 arrival rate will soon increase to 4.5. We see from the rest of the trajectories that our policy achieves a much lower holding cost for class 2, while having only a slightly higher holding cost for class 1. On the other hand, the modified maximum pressure policy provides too much unnecessary help (from pool 2 to class 1), which results in the class 2 queue building up to a high value (around 100) at $t = 40$; from then on pool 2 has little slackness and it takes a long time to reduce the class 2 queue.

Table 2.3 reports the average costs under our policy as well as under the benchmark policies in this arrival rate setting. The costs for the two policies in Figure 2.9 are in the first panel ($\phi = 2$): 1.11×10^4 for our policy versus 1.68×10^4 for the modified maximum pressure policy – 50% higher than ours. This performance gap further enlarges when ϕ increases. The cost under maximum pressure policy is more than twice the cost under our policy when $\phi = 25$. More generally, the two maximum pressure policies can perform arbitrarily worse than our policy as the arrival rate of class 2, after the demand surge of class 1 is over, gets closer to 5 (while remaining stable). This is because the unnecessary helping in the demand surge period will result in the class 2 queue taking

an extremely long time to deplete when the slackness of pool 2 approaches 0. On the other hand, the modified $c\mu$ policy performs well in this case, since it is a no-overflow policy. However, the modified $c\mu$ policy doubles the cost of our policy in the baseline setting (Table 2.2). This indicates that, the performance of policies that do not use future arrival rate can vary a lot depending on the arrival rate patterns, which highlights the value of our look-ahead policy in time-nonstationary environment.

		Look-ahead	MaxPres	ModMaxP	Cmu	ModCmu
$\phi = 2$	Holding	1.07	1.60	1.59	2.91	1.28
	Overflow	0.03	0.09	0.09	0.14	0.00
	Total	1.11	1.69	1.68	3.05	1.28
	SE	0.002	0.006	0.006	0.010	0.002
$\phi = 10$	Holding	1.11	1.60	1.54	2.91	1.28
	Overflow	0.11	0.46	0.44	0.73	0.00
	Total	1.22	2.06	1.99	3.64	1.28
	SE	0.002	0.006	0.006	0.010	0.002
$\phi = 25$	Holding	1.28	1.60	1.46	2.91	1.28
	Overflow	0.00	1.16	1.06	1.84	0.00
	Total	1.28	2.76	2.52	4.75	1.28
	SE	0.002	0.007	0.006	0.012	0.002

Table 2.3: Expected total cost for the N-model under different routing policies. The costs shown in the table are in units of 10^4. "SE" stands for the standard error for the corresponding average total cost (holding + overflow). $(\lambda_1(t) = 8 \times \mathbf{1}\{t < 40\} + 1 \times \mathbf{1}\{t \geq 40\}$, $\lambda_2(t) = 3 \times \mathbf{1}\{t < 40\} + 4.5 \times \mathbf{1}\{t \geq 40\}$. Other parameters are the same as the baseline.)

2.6.2 Extensions to General Multi-class Multi-pool Systems

In this section, we test our proposed heuristic policy as given in Section 2.5.3 on (1) the X-model and (2) two networks with five classes of customers and five pools of servers (5-by-5). In the interest of space, we focus on the results for the 5-by-5 networks. Appendix A.2 details the results for the X-model. Because we already know the optimal fluid control for the X-model, we can compare the performance of the optimal fluid-translated control to the heuristic policy. The results for the X-model show that our heuristic policy has performance comparable to that of the fluid-optimal policy. This provides some evidence that our heuristic policy would work well

for systems beyond the N-model. For the 5-by-5 networks, we compare the performance of the heuristic look-ahead policy with that of other modified benchmark policies since the optimal policy is unknown (prohibitive to get) in this setting.

In the 5-by-5 networks, the holding costs are set to be $(1.5, 1, 1, 1.5, 1)$, and the overflow costs ϕ are the same for all overflow assignments. We consider two arrival rate settings. For the first setting, the arrival rates for class 1 and class 4 are

$$\lambda_1(t) = \begin{cases} 12, & t < 40 \\ 4.5, & t \geq 40 \end{cases} \text{ and } \lambda_4(t) = \begin{cases} 8, & t < 40 \\ 4, & t \geq 40 \end{cases},$$

and the arrival rates for other classes are constants: $\lambda_2(t) = 3$, $\lambda_3(t) = 4$, $\lambda_5(t) = 3$. Classes 1 and 4 experience demand surges while the others do not. For the second setting, the arrival rates for class 1 and class 2 are

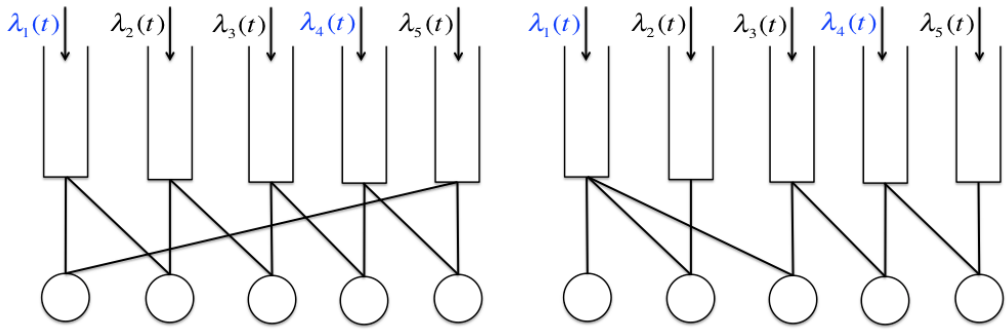
$$\lambda_1(t) = \begin{cases} 12, & t < 40 \\ 2, & t \geq 40 \end{cases} \text{ and } \lambda_2(t) = \begin{cases} 3, & t < 40 \\ 4.5, & t \geq 40 \end{cases},$$

while the arrival rates for the other classes are the same as in the first setting, including the surge for class 4. The two arrival rate settings are chosen to be consistent with those used in the N-model experiments.

We consider two network structures as depicted in Figure 2.10. The first network has a closed-chain structure, which is a commonly advocated flexibility architecture in supply chain and manufacturing applications [21, 56].

Tables 2.4 and 2.5 compare the cost under the five policies for the two network structures. The look-ahead policy is our proposed heuristic policy with $\theta = 0.8$ – a 20% tuning down on G. We find that this tuning parameter performs very well across all experiments. Thus, we recommend this policy for practical use.[1] We observe that our proposed policy again performs the best in both

[1]Tuning, in general, improves the cost from the untuned version by 0.5% to 4%; see Appendix A.2.2 for the detailed results for the X-model and 5-by-5 network. Meanwhile, even using the untuned version, our heuristic policy



(a) Network I (b) Network II

Figure 2.10: Two network structures. Classes 1 and 4 have demand surge in both arrival rate settings. In the second setting, class 2 also has a demand surge.

network structures and the two arrival rate settings. The modified $c\mu$ rule does not overflow, and thus performs the worse in the first arrival rate setting. The modified maximum pressure policy tends to perform better in the first arrival rate setting, showing a similar performance to that of our policy when $\phi = 2$; however, when $\phi = 10$ or 25, the modified maximum pressure policy results in a much higher overflow cost than our policy does. This is despite the fact that it incorporates overflow costs. The modified maximum pressure policy shows an even larger performance gap from our policy in the second arrival setting. Similar to what we explained in Section 2.6.1, this is because the maximum pressure policy does not account for the future arrival rate information, and ends up providing too much help during the demand surge, which hurts the class 2 queue.

2.6.3 Impact of Prediction Error in Arrival Rates

In Sections 2.3.2, 2.3.3, and 2.4, we have analyzed the effects of prediction errors and limited look-ahead time window in the N-model. In this section, we numerically investigate these effects in the original stochastic N-model system. We show that our look-ahead policy still performs significantly better than the benchmark policies even when the prediction errors are large (compared to the mean arrival rate) and when the look-ahead time window is small (even 0). These results show that our policy is appealing in practice since it is robust to various prediction imperfections.

performs reasonably well.

		Look-ahead	ModMaxP	ModCmu
		Arrival Rate Setting I		
$\phi = 2$	Holding	3.21	3.27	11.08
	Overflow	0.52	0.59	0.00
	Total	3.73	3.87	11.08
	SE	0.007	0.007	0.014
$\phi = 10$	Holding	3.45	3.27	11.08
	Overflow	2.05	2.75	0.00
	Total	5.50	6.03	11.08
	SE	0.008	0.009	0.014
$\phi = 25$	Holding	4.29	3.31	11.08
	Overflow	3.57	6.27	0.00
	Total	7.86	9.58	11.08
	SE	0.012	0.013	0.014
		Arrival Rate Setting II		
$\phi = 2$	Holding	3.00	3.27	11.08
	Overflow	0.48	0.42	0.00
	Total	3.48	3.69	11.08
	SE	0.005	0.007	0.014
$\phi = 10$	Holding	3.07	3.27	11.08
	Overflow	1.98	2.06	0.00
	Total	5.05	5.34	11.08
	SE	0.007	0.008	0.014
$\phi = 25$	Holding	3.57	3.27	11.08
	Overflow	3.77	5.03	0.00
	Total	7.34	8.31	11.08
	SE	0.009	0.010	0.014

Table 2.4: Simulation costs for the 5-by-5 model under Network Structure I. The costs shown in the table are in units of 10^4. "SE" stands for the standard error for the corresponding total cost. (Parameter setting: $h = (1.5, 1, 1, 1.5, 1)$, $s_i = 20$, $\mu_{ii} = 0.25$, $\mu_{ij} = 0.2$ and $\phi_{ij} = \phi$ for $i \neq j$ and $X(0) = (30, 40, 50, 60, 70)$. For arrival rate setting I, $\lambda_1(t) = 12 \times \mathbf{1}\{t < 40\} + 4.5 \times \mathbf{1}\{t \geq 40\}$, $\lambda_2(t) = 3$, $\lambda_3(t) = 4$, $\lambda_4(t) = 8 \times \mathbf{1}\{t < 40\} + 4 \times \mathbf{1}\{t \geq 40\}$, $\lambda_5(t) = 3$. For arrival rate setting II, $\lambda_1(t) = 12 \times \mathbf{1}\{t < 40\} + 2 \times \mathbf{1}\{t \geq 40\}$, $\lambda_2(t) = 3 \times \mathbf{1}\{t < 40\} + 4.5 \times \mathbf{1}\{t \geq 40\}$, $\lambda_3(t) = 4$, $\lambda_4(t) = 8 \times \mathbf{1}\{t < 40\} + 4 \times \mathbf{1}\{t \geq 40\}$, $\lambda_5(t) = 3$.)

		Look-ahead	ModMaxP	ModCmu
		Arrival Rate Setting I		
$\phi = 2$	Holding	3.00	4.26	5.94
	Overflow	0.29	0.29	0.00
	Total	3.29	4.55	5.94
	SE	0.005	0.008	0.010
$\phi = 10$	Holding	3.36	4.22	5.94
	Overflow	0.89	1.42	0.00
	Total	4.25	5.64	5.94
	SE	0.006	0.009	0.010
$\phi = 25$	Holding	4.20	4.14	5.94
	Overflow	1.10	3.46	0.00
	Total	5.31	7.61	5.94
	SE	0.007	0.010	0.010
		Arrival Rate Setting II		
$\phi = 2$	Holding	2.93	3.69	5.94
	Overflow	0.25	0.34	0.00
	Total	3.18	4.03	5.94
	SE	0.005	0.008	0.010
$\phi = 10$	Holding	3.15	3.64	5.94
	Overflow	0.93	1.68	0.00
	Total	4.09	5.33	5.94
	SE	0.006	0.009	0.010
$\phi = 25$	Holding	4.19	3.55	5.94
	Overflow	1.11	4.05	0.00
	Total	5.30	7.61	5.94
	SE	0.007	0.010	0.010

Table 2.5: Simulation costs for the 5-by-5 model under Network Structure II. The costs shown in the table are in units of 10^4. "SE" stands for the standard error for the corresponding total cost. (Parameter setting: $h = (1.5, 1, 1, 1.5, 1)$, $s_i = 20$, $\mu_{ii} = 0.25$, $\mu_{ij} = 0.2$ and $\phi_{ij} = \phi$ for $i \neq j$ and $X(0) = (30, 40, 50, 60, 70)$. For arrival rate setting I, $\lambda_1(t) = 12 \times \mathbf{1}\{t < 40\} + 4.5 \times \mathbf{1}\{t \geq 40\}$, $\lambda_2(t) = 3$, $\lambda_3(t) = 4$, $\lambda_4(t) = 8 \times \mathbf{1}\{t < 40\} + 4 \times \mathbf{1}\{t \geq 40\}$, $\lambda_5(t) = 3$. For arrival rate setting II, $\lambda_1(t) = 12 \times \mathbf{1}\{t < 40\} + 2 \times \mathbf{1}\{t \geq 40\}$, $\lambda_2(t) = 3 \times \mathbf{1}\{t < 40\} + 4.5 \times \mathbf{1}\{t \geq 40\}$, $\lambda_3(t) = 4$, $\lambda_4(t) = 8 \times \mathbf{1}\{t < 40\} + 4 \times \mathbf{1}\{t \geq 40\}$, $\lambda_5(t) = 3$.)

Error in the arrival rate estimate

For simplicity, we assume that the estimated arrival rate takes the form $\tilde{\lambda}_1(t) = \lambda_1(t) + \epsilon$, where ϵ is a random variable. We draw 25 different realizations of ϵ in the simulation experiments and report the average cost over these realizations. We use the baseline parameter setting as specified in Section 2.6.1 as well as the second arrival rate setting of Section 2.6.1.

We consider three scenarios: (i) no prediction error, (ii) small error: ϵ follows a standard normal random variable $N(0, 1)$; and iii) large error: a normal random variable of mean 0 and variance 4, $N(0, 4)$. Recall that $\lambda_1(t)$ takes the value of 8 during the surge and value of 4 or 1 after the surge, so this variance of 4 is large compared to the mean. The results are shown in Table 2.6, which compares the costs under our look-ahead policy in these three scenarios. In addition, we also show the costs under the modified maximum pressure policy and modified $c\mu$ policy. We observe that our look-ahead policy achieves consistently good performance in all three estimation error scenarios and both arrival rate settings. For example, even when $\phi = 25$ and $\epsilon \sim N(0, 4)$, the total cost incurred is 2.336×10^4 (1.297×10^4) in the first (second) arrival rate setting, which is only 2.6% (1.2%) higher than the total cost with perfect arrival rate information (i.e., $\epsilon = 0$). Moreover, our policy performs consistently good comparing to the benchmarks. It performs similarly to or better than the modified maximum pressure policy in the first setting, and similarly to or better than the modified $c\mu$ policy in the second setting, even when the prediction error is large ($\epsilon \sim N(0, 4)$). This robust performance of our policy is in contrast to the swaying performance of the modified maximum pressure or modified $c\mu$ policy: they could have good performance in one setting but perform much worse in another setting. This difference in the robust versus non-robust performance is consistent with what we observed in Figure 2.8 and discussed in Section 2.6.1.

To further investigate the effect of over- and under-estimation on our policy performance, we consider scenarios where ϵ equals a fixed value that ranges from -4 to 4. In other words, there is a persistent estimation bias. Note that a bias of 4 is large when compared to the surge arrival rate of 8. The results are summarized in Table 2.7. We observe that even large values of $|\epsilon|$ cause

		0	$N(0,1)$	$N(0,4)$	ModMaxP	ModCmu
		Arrival Rate Setting I				
$\phi = 2$	Holding	1.092	1.094	1.098	1.104	2.754
	Overflow	0.141	0.142	0.142	0.126	0.000
	Total	1.232	1.236	1.240	1.230	2.754
	SE	0.003	0.001	0.002	0.003	0.008
$\phi = 10$	Holding	1.099	1.104	1.107	1.108	2.754
	Overflow	0.562	0.571	0.579	0.619	0.000
	Total	1.661	1.675	1.686	1.727	2.754
	SE	0.004	0.002	0.005	0.004	0.008
$\phi = 25$	Holding	1.276	1.249	1.230	1.118	2.754
	Overflow	0.994	1.054	1.106	1.504	0.000
	Total	2.276	2.303	2.336	2.622	2.754
	SE	0.005	0.007	0.014	0.005	0.008
		Arrival Rate Setting II				
$\phi = 2$	Holding	1.071	1.071	1.074	1.590	1.282
	Overflow	0.035	0.035	0.037	0.092	0.000
	Total	1.106	1.107	1.111	1.682	1.282
	SE	0.002	0.003	0.003	0.006	0.002
$\phi = 10$	Holding	1.116	1.110	1.110	1.543	1.282
	Overflow	0.112	0.119	0.126	0.449	0.000
	Total	1.228	1.230	1.236	1.992	1.282
	SE	0.002	0.003	0.003	0.006	0.002
$\phi = 25$	Holding	1.282	1.277	1.256	1.466	1.282
	Overflow	0.000	0.007	0.040	1.061	0.000
	Total	1.282	1.284	1.297	2.527	1.282
	SE	0.002	0.004	0.011	0.006	0.002

Table 2.6: Simulation costs for the N-model when ϵ is random with distribution standard normal and normal with mean zero and variance 4. Costs under the modified maximum pressure policy and modified $c\mu$ policy are also shown. The costs shown in the table are in units of 10^4. "SE" stands for the standard error for the corresponding total cost. (Parameter setting: $h = (1.5, 1)$, $s_i = 20$ and $\mu_{ii} = 0.25$, for $i = 1, 2$, $\mu_{12} = 0.2$, $\phi_{12} = \phi$, and $X(0) = (60, 70)$. For the first arrival rate setting, $\lambda_1(t) = 8 \times \mathbf{1}\{t < 40\} + 4 \times \mathbf{1}\{t \geq 40\}$ and $\lambda_2(t) = 3$. For the second arrival rate setting, $\lambda_1(t) = 8 \times \mathbf{1}\{t < 40\} + 1 \times \mathbf{1}\{t \geq 40\}$ and $\lambda_2(t) = 3 \times \mathbf{1}\{t < 40\} + 4.5 \times \mathbf{1}\{t \geq 40\}$.)

		-4	-2	0	2	4
		Arrival Rate Setting I				
$\phi = 2$	Holding	1.096	1.089	1.092	1.099	1.106
	Overflow	0.139	0.140	0.141	0.143	0.145
	Total	1.235	1.228	1.232	1.242	1.251
	SE	0.003	0.003	0.003	0.003	0.003
$\phi = 10$	Holding	1.126	1.111	1.104	1.103	1.107
	Overflow	0.548	0.551	0.562	0.586	0.607
	Total	1.674	1.661	1.666	1.689	1.714
	SE	0.004	0.004	0.004	0.004	0.004
$\phi = 25$	Holding	1.354	1.329	1.276	1.194	1.147
	Overflow	0.934	0.939	0.999	1.148	1.272
	Total	2.287	2.268	2.276	2.342	2.419
	SE	0.005	0.005	0.005	0.005	0.005
		Arrival Rate Setting II				
$\phi = 2$	Holding	1.079	1.074	1.071	1.071	1.076
	Overflow	0.030	0.032	0.035	0.038	0.042
	Total	1.109	1.106	1.106	1.109	1.118
	SE	0.002	0.002	0.002	0.003	0.003
$\phi = 10$	Holding	1.183	1.145	1.116	1.096	1.084
	Overflow	0.061	0.087	0.112	0.135	0.160
	Total	1.244	1.232	1.228	1.232	1.244
	SE	0.002	0.002	0.002	0.002	0.002
$\phi = 25$	Holding	1.282	1.282	1.282	1.282	1.189
	Overflow	0.000	0.000	0.000	0.000	0.142
	Total	1.282	1.282	1.282	1.282	1.330
	SE	0.002	0.002	0.002	0.002	0.002

Table 2.7: Simulation costs for the N-model when ϵ is equal to a fixed value ranging from -4 to 4. The costs shown in the table are in units of 10^4. "SE" stands for the standard error for the corresponding total cost. (Parameter setting: $h = (1.5, 1)$, $s_i = 20$ and $\mu_{ii} = 0.25$, for $i = 1, 2$, $\mu_{12} = 0.2$, $\phi_{12} = \phi$, and $X(0) = (60, 70)$. For the first arrival rate setting, $\lambda_1(t) = 8\times\mathbf{1}\{t < 40\}+4\times\mathbf{1}\{t \geq 40\}$ and $\lambda_2(t) = 3$. For the second arrival rate setting, $\lambda_1(t) = 8 \times \mathbf{1}\{t < 40\} + 1 \times \mathbf{1}\{t \geq 40\}$ and $\lambda_2(t) = 3 \times \mathbf{1}\{t < 40\} + 4.5 \times \mathbf{1}\{t \geq 40\}$.)

relatively small changes in the performance of our policy. In the scenarios tested, over-estimation tends to lead to a higher cost than underestimation. This is because over-estimation leads to extra overflow, which results in a higher overflow cost (especially when ϕ is large) and sometimes also a higher holding cost (due to service slowdown).

Limited future arrival rate information

We next study the impact of a limited look-ahead time window. Using the same notation introduced in Section 2.3.3, we assume only the future arrival rate up to time $t + W$ is known for a given time window W. We consider two arrival rate settings that each feature two demand surges. When there are two surges, a limited look-ahead time may have a larger effect on performance, because the policy might not be able to anticipate the second demand surge when planning during the first demand surge.

The two arrival rate settings are derived from the two arrival rate settings studied earlier for the N-model, except that we break the initial demand surge into two separate surges. The first arrival rate setting is the same as the baseline, except that the class 1 arrival rate is now

$$\lambda_1(t) = \begin{cases} 8, & t < 20 \\ 4, & 20 \le t < 30 \\ 8, & 30 \le t < 50 \\ 4, & t \ge 50. \end{cases}$$

The policy implemented under the two-demand surges is specified in Section 2.3.4 and is adapted to the limited time-window. For example, if $W = 0$, we implement the look-ahead policy with $\tilde{G}_1^t(X_1(t)) = X_1(t)/(20 \times 0.25 - 4)$, because it is believed that the class 1 arrival rate is constant and equal to 4.

The second arrival rate setting is the same as that in Section 2.6.1, except that the class 1 and 2

arrival rates are

$$\lambda_1(t) = \begin{cases} 8, & t < 20 \\ 1, & 20 \le t < 30 \\ 8, & 30 \le t < 50 \\ 1, & t \ge 50 \end{cases} \text{ and } \lambda_2(t) = \begin{cases} 3, & t < 20 \\ 4.5, & 20 \le t < 30 \\ 3, & 30 \le t < 50 \\ 4.5, & t \ge 50. \end{cases}$$

For the numerical experiments, we test three different values of W: $0, 5$, and 10. Note that these time windows are smaller than or equal to the length of the interval between the two demand surges, so that the policy does not 'know' about the second demand surge during the first. Table 2.8 summarizes the simulation results under these two arrival rate settings. As W increases, more of the demand surge rate is revealed at the beginning, and, hence, the value of $\tilde{G}_1^t(q_1(t))$ increases. Consequently, more help is offered following (2.8), which explains why the overflow cost increases while the holding cost decreases. As expected, the total cost generally decreases with W, but the performance change is quite small. Our policy, even with a small value of W, generally performs much better than the two look-ahead policies. The modified maximum pressure policy performs close to our policy in the first arrival rate setting when ϕ is small (2 or 10), but otherwise performs significantly worse, even compared to the case $W = 0$. Meanwhile, the modified maximum $c\mu$ performs well in the second arrival rate setting, but performs extremely poorly in the first arrival rate setting. The numerical results indicate that our policy is robust to limited future arrival rate information regardless of the parameter setting, which is very desirable for practical implementations.

2.7 Conclusion

In this work, we study the value of future arrival rate information in designing the optimal routing policy for systems with partial flexibility when facing demand surges. Our model incorporates salient features of service systems, such as efficiency loss and inconvenience costs associated

		0	5	10	ModMaxP	ModCmu
		Arrival Rate Setting I				
$\phi = 2$	Holding	0.981	0.979	0.980	0.992	2.632
	Overflow	0.133	0.134	0.135	0.122	0.000
	Total	1.114	1.113	1.115	1.114	2.632
	SE	0.003	0.003	0.003	0.003	0.008
$\phi = 10$	Holding	1.011	0.999	0.992	0.995	2.632
	Overflow	0.537	0.543	0.549	0.598	0.000
	Total	1.549	1.543	1.542	1.593	2.632
	SE	0.004	0.004	0.004	0.004	0.008
$\phi = 25$	Holding	1.270	1.237	1.215	1.006	2.632
	Overflow	0.896	0.920	0.939	1.452	0.000
	Total	2.166	2.158	2.154	2.458	2.632
	SE	0.005	0.005	0.005	0.005	0.008
		Arrival Rate Setting II				
$\phi = 2$	Holding	1.027	0.960	0.915	1.112	1.049
	Overflow	0.004	0.021	0.027	0.075	0.000
	Total	1.032	0.981	0.943	1.187	1.049
	SE	0.002	0.002	0.002	0.004	0.002
$\phi = 10$	Holding	1.049	1.045	1.004	1.080	1.049
	Overflow	0.000	0.003	0.035	0.361	0.000
	Total	1.049	1.049	1.040	1.442	1.049
	SE	0.002	0.002	0.004	0.000	0.002
$\phi = 25$	Holding	1.049	1.049	1.049	1.030	1.049
	Overflow	0.000	0.000	0.000	0.841	0.000
	Total	1.049	1.049	1.049	1.871	1.049
	SE	0.002	0.002	0.002	0.005	0.002

Table 2.8: Simulation costs for the N-model with time window $W = 0, 5,$ and 10. Costs under the modified maximum pressure policy and modified $c\mu$ policy are also shown. The costs shown in the table are in units of 10^4. "SE" stands for the standard error for the corresponding total cost. (Parameter setting: $h = (1.5, 1)$, $s_i = 20$ and $\mu_{ii} = 0.25$, for $i = 1, 2$, $\mu_{12} = 0.2$, $\phi_{12} = \phi$, and $X(0) = (60, 70)$. For the first arrival rate setting, $\lambda_1(t) = 8 \times \mathbf{1}\{t < 20\} + 4 \times \mathbf{1}\{20 \le t < 30\} + 8 \times \mathbf{1}\{30 \le t < 50\} + 4 \times \mathbf{1}\{t \ge 50\}$ and $\lambda_2(t) = 3$. For the second arrival rate setting, $\lambda_1(t) = 8 \times \mathbf{1}\{t < 20\} + 1 \times \mathbf{1}\{20 \le t < 30\} + 8 \times \mathbf{1}\{30 \le t < 50\} + 1 \times \mathbf{1}\{t \ge 50\}$ and $\lambda_2(t) = 3 \times \mathbf{1}\{t < 20\} + 4.5 \times \mathbf{1}\{20 \le t < 30\} + 3 \times \mathbf{1}\{30 \le t < 50\} + 4.5 \times \mathbf{1}\{t \ge 50\}$.)

with overflow, general time-varying demand (arrival rates), and various service compatibility architectures (e.g., the N- and X-models). We study the fluid control problems for the N-model, the X-model, and two extensions of the N-model, and explicitly characterize the optimal control. All of these policies have a two-stage index-based structure and take future arrival rates into account. Based on the insights from the the fluid analysis, we propose a two-stage index-based look-ahead policy for general stochastic systems. Our proposed policy is interpretable, easy to implement, and able to achieve superior performance in various multi-class multi-pool networks, even when there are prediction errors or limited future arrival rate information.

Several future extensions may be considered. One is to jointly optimize the system's architectural design and the real-time routing policy (see, for example, [57]). Our analysis reveals the potential benefit of cross-training when, for example, comparing the N-model with the X-model under demand surge for class 1. In the N-model, pool 1 can only serve class 1, while in the X-model, pool 1 is cross-trained to serve class 2. The X-model is able to absorb the demand surge faster and at a lower cost than the N-model. This is because, due to cross-training, pool 2 can provide more help to class 1 since pool 1 can "pay back" the help later. Another extension is to study the effect of more general prediction errors. Our current analysis assumes the prediction error is of a smaller order than the arrival rate. In this case, our policy performs well even with prediction errors. When the prediction error and the arrival rate are of similar orders, it would be interesting to study how to adjust the routing policy to properly account for the prediction error. Lastly, it would be interesting to see how our analytical framework, which leverages optimal control theory (Pontryagin's minimum principle) and fluid-scale convergence analysis, can be applied to solve other transient queueing control problems.

Chapter 3: Optimal sizing and scheduling of flexible servers

3.1 Introduction

In multi-class service systems, servers can either be dedicated (only able to serve one class) or flexible (able to serve multiple classes). Increasing the size of the flexible server pool can help balance the workload between different classes of customers, and improve system performance. Specifically, when managing queues with multiple classes of jobs, the benefit of load-balancing and capacity flexibility have been studied and demonstrated in various settings (see, for example, [58, 24]).

However, as noted in Chapter 1, flexibility may come at a cost, such as less efficient service or more expensive staffing. Given the cost and benefit of flexible capacity, it is important to understand how to strike a balance in resource management.

When designing the service system, the service provider has to make multiple decisions. Chief among them are how many of each type of server to staff and how to match customers with servers. These problems are often referred to as the staffing and scheduling problems in the literature. In this work, we study the joint staffing and scheduling problem in multi-class queues with both dedicated and flexible servers. In particular, to highlight the key tradeoff, we consider a stylized M-model with two classes of customers and three potential pools of servers: two dedicated pools and one flexible pool that can serve both classes of customers. To capture the cost of flexibility, we assume that the flexible servers may be more costly to staff and may serve at a slower rate than dedicated servers. The objective is to find the optimal staffing and scheduling policies that minimize the sum of the staffing cost, holding cost, and abandonment cost.

We consider two demand scenarios. One has deterministic arrival rates, which is the case when we have a very accurate estimate of customer demand. In this case, the flexible pool can be used to

hedge against stochasticity, i.e., the stochastic fluctuation of interarrival times and service times. In particular, due to the stochasticity in system dynamics, one queue may incur a higher than average load while the other is at or below its normal load from time to time. In such situations, the flexible pool can be used to help the class with a heavier load, and thus balance the load between the two classes. The other scenario has random arrival rates, which is the case when there is a high degree of uncertainty in customer demand. In this case, the flexible pool is mainly used to hedge against parameter uncertainty. In particular, when the realized arrival rate of one class is higher than average while the realized arrival rate of the other class is at or below average, the flexible pool can be used to help the class with a higher realized arrival rate, and thus balance the load. The differences between the two scenarios described above give rise to different hedging mechanisms, which in turn lead to different sizes of the flexible pool in optimality. To see this, let λ denote the average arrival rate. When λ is large, the stochastic fluctuation of the system with a given arrival rate is in general of order $\sqrt{\lambda}$ [59]. The parameter uncertainty, on the other hand, can be of a different order than $\sqrt{\lambda}$ [7]. Indeed, the case we are interested in is one where the standard deviation of the random arrival rate is of a larger order than $\sqrt{\lambda}$. Lastly, the different hedging mechanisms also lead to different scheduling policies in our developments.

Because staffing and scheduling decisions interact, the joint optimization problem can be very challenging. When arrival rates are deterministic and symmetric, we use a coupling construction to derive the optimal scheduling policy for any staffing level. The scheduling policy prioritizes the dedicated servers (faster servers) when routing customers to servers, and prioritizes the class with more customers in the system when scheduling flexible servers, assuming the abandonment rate is less than the service rates. Given the optimal scheduling policy, we then optimize the staffing policy. To derive structural insights into the size of the flexible pool, we employ a heavy-traffic asymptotic approach, where we send the arrival rate to infinity and study how the size of the flexible pool scales with the arrival rate. Our result provides necessary and sufficient conditions for staffing rules to be asymptotically optimal. The key insight is that when flexibility comes at a cost, the optimal size of the flexible pool only leads to partial resource pooling. In particular, the

flexible pool helps create some load-balancing, but the effect is not large enough to equalize the two queues asymptotically.

When arrival rates are random and the magnitude of the parameter uncertainty dominates the system stochasticity, we employ a stochastic-fluid relaxation of the optimal staffing problem. In this relaxation, we ignore the stochasticity of the queueing dynamics and focus on the parameter uncertainty only. The stochastic-fluid optimization problem is a special case of the single-period multi-product inventory problem with demand substitution, for which we can characterize the optimal solution explicitly. The relaxation also motivates a simple scheduling rule that essentially decomposes the M-model into two independent inverted-V models for any realization of the arrival rates. When the average arrival rates grow to infinity, we show that the staffing and scheduling rules derived based on the stochastic-fluid relaxation are asymptotically optimal. The key insight is that when facing both parameter uncertainty and cost of flexibility, the optimal size of the flexible pool provides some hedging against the parameter uncertainty, and the cost saving, compared to the no-flexible resource case, is increasing with the magnitude of the uncertainty.

In addition to providing prescriptive solutions to managing flexibility, we also highlight the following contributions of our work.

1. When the arrival rates are symmetric and deterministic, we construct the optimal scheduling policy for any arrival rates and staffing levels. In contrast to most of the optimal scheduling literature for multi-server queues, our results do not rely on any asymptotic argument (for development on asymptotically optimal scheduling policies, see, for example, [4]). Instead, the proof uses a coupling argument that can be of interest to the analysis of other Markovian queueing systems. Our coupling technique also allows us to establish the optimality of a non-standard scheduling policy when the abandonment rate is larger than the service rates, see Theorem 14.

2. When the arrival rates are deterministic and the flexible pool is of the optimal order, we derive the diffusion limit of the M-model under heavy-traffic. The limit is a two-dimensional diffusion process. In particular, the complete resource pooling condition is not satisfied

when the flexible pool is optimally sized, i.e., the flexible pool size is not large enough to instantaneously balance the queue lengths between the two classes. Thus, we do not have state space collapse in the limit, i.e., the two-dimensional queue length process does not reduce to a one-dimensional process in the limit. This is in contrast to most of the optimal scheduling literature (see, for example, [5, 6]). On the other hand, the limiting process cannot be fully decomposed along each dimension, i.e., the drift terms of the two component diffusion processes are interconnected. Thus, we achieve partial resource pooling.

3. When the arrival rates are random and the parameter uncertainty is of a larger order than the stochasticity of the queueing dynamics, we quantify the optimality gap for policies derived based on the stochastic fluid approximation. This extends the results in [7] from a multi-server queue with a single class of customers and a single pool of servers to a multi-class queue with multiple server types. We also allow the arrival rate distributions of the two classes to be asymmetric, i.e., they can have different means and different levels of uncertainty.

3.1.1 Literature review

We first review related works on queues with deterministic arrival rates. The M-model studied in this work is a special case of parallel server systems (PSSs). Due to the interplay between staffing and scheduling decisions, the joint staffing and scheduling problem can be highly nontrivial for general PSSs. In the literature, most works only look at one of the two problems in isolation. However, there are a few exceptions. Noticeably, [60] consider the joint optimization problem for an inverted-V model where there is a single class of customers and multiple types of servers. Using a coupling argument, they establish the optimality of the fastest-server-first policy. [61] study the problem of staffing and scheduling PSSs to minimize total staffing costs subject to quality-of-service constraints. They establish that the queue-and-idleness-ratio control is asymptotically optimal in heavy-traffic. When dealing with a single class of customers and a single pool of servers, [8] study the optimal staffing problem in an $M/M/n$ queue. They find that the quality-

and-efficiency-driven (QED) regime, which is also known as the Halfin-Whitt regime [62], arises naturally when staffing is set to balance the staffing cost and the system performance. The work is then extended by [63] to allow for customer abandonment.

The work that is most related to ours is [23], which studies the sizing of flexible resources when service rates can be continuously chosen. They find that the linear staffing and holding costs often lead to an $O(\sqrt{\lambda})$ flexibility when flexible capacity is more expensive. The main difference between our work and theirs is the modeling of the service resources. They use a single-server mode of analysis and assume the service rate can be optimally chosen. This modeling approach is reasonable for computer or manufacturing systems. Motivated mostly by large-scale service systems, our work adopts a many-server mode of analysis. As [23] point out, the many-server regime that we consider introduces substantial complexity to the analysis, and they leave this extension as a potential future research direction. In addition, [23] assumes a longest-queue-first scheduling and hypothesize that it is likely to be optimal. We establish the optimality of a scheduling policy that prioritizes the class with more customers in the system.

More broadly, optimal scheduling of various PSSs has been extensively studied in the literature. For example, [64] study the optimal scheduling of the N-model. They show that a $c\mu$-type of greedy policy is asymptotically optimal in the many-server QED regime. [4] studies the optimal scheduling problem of general PSSs, i.e., with multiple classes of customers and multiple pool of servers, and customer abandonment. The work establishes the asymptotical optimality of policies derived based on the corresponding optimal diffusion control problem in the many-server QED regime. [65] study the optimal scheduling of V-model with general patience-time distributions. The main feature that distinguishes our work from the stream of works on PSSs in the QED regime is the size of our flexible server pool. In our analysis, the size of the flexible pool is asymptotically negligible in the fluid scale, whereas in the literature, it is almost universally assumed that the fluid-scaled pool sizes are non-negligible (see, for example, Assumption 1 in [4], Assumption 2.1 in [6], and equation (20) in [5]). Due to the difference in the size of our server pools, the asymptotic behavior (diffusion limit) of our system can be qualitatively different from what is observed in the

literature.

When the arrival rates are random, our work is related to works that look at staffing queues when facing parameter uncertainty. The stochastic-fluid relaxation was first proposed in [66]. Its efficacy has been studied in several subsequent works. [67] show that it leads to an asymptotically optimal staffing policy under a non-conventional asymptotic regime that features large arrival rates and short service times. The asymptotic framework is then extended in [15], who consider the case when the arrival rate distribution is unknown and has to be estimated from data. Compared to these works, the analysis in this work takes a different asymptotic approach. In particular, we increase the system demand, i.e., arrival rates, but do not scale other system parameters such as service rates and abandonment rates. The paper [7] takes a similar heavy-traffic asymptotic approach as ours and establishes the optimality gap of the staffing policy derived from the stochastic-fluid relaxation for an Erlang-A model with a random arrival rate. We extend their results to a multi-class network setting, where in addition to the staffing decision, we also have to decide on the scheduling policy.

[68] develops a different stochastic-fluid model that allows non-exponential service times and patience times, and studies the staffing problem with both random arrival rates and staffing levels (due to employee absenteeism). When facing demand uncertainty, [69] study the staffing problem with a chance constraint for the quality of service. They first use mixed integer programming to obtain a first-order staffing solution, and then refine the staffing level using simulation. [70] study the staffing and outsourcing problem when demand is random.

Our work contributes to this stream of literature in two key ways. First, we show that when dealing with random demand, it is the staffing, not the scheduling decision, that is of paramount importance. This supports why many papers tend to focus on the staffing instead of the scheduling decision in this setting (see, for example [69, 71]). Second, we quantify the benefit of flexibility. Specifically, we extend the notion that the order of flexibility should match the order of system stochasticity in [23] to the case where the order of flexibility should match the order of demand uncertainty. In general, the notion that just a small degree of flexibility is enough has been much investigated over the years in various different contexts (see, for example, [21, 72] for manufacturing

systems, [24, 73] for PSSs, etc.) Our work contributes to this literature as well.

3.1.2 Structure and Notation

In Section 3.2, we introduce the queueing model and the optimization problem. In Section 3.3, we study the optimal scheduling and staffing policy for a symmetric M-model with deterministic arrival rates. The goal is to highlight the cost and benefit of flexibility in a classical setting with no parameter uncertainty. In Section 3.4, we study the staffing and scheduling problem for systems with random arrival rates. To highlight the effect of demand uncertainty, we focus on the regime where the demand uncertainty dominates the system stochasticity. We complement our theoretical analysis with numerical experiments in Section 3.5. In particular, the numerical analysis focuses on the pre-limit performance of our proposed staffing and scheduling rules. The proofs of all the theoretical results are delayed until the Appendix.

We next introduce some notation that is used throughout the work. The set of non-negative integers is denoted by $\mathbb{N}_0$, and the set of real numbers is denoted by $\mathbb{R}$. We define $\eta(t) = 0$, $\chi(t) = t$, and $I(t) = 1$, for $t \geq 0$. Let D denote the space of functions from $[0, \infty)$ to $\mathbb{R}$ that are right-continuous with left limits and is endowed with Skorohod J_1 topology. Let e_i be a unit vector with the i-th element equal to 1. The dimension of e_i depends on the context. We write $1\{\cdot\}$ for the indicator function. A random variable A is said to be stochastically larger than a random variable B, $A \geq_{st} B$, if $\mathbb{P}(A > x) \geq \mathbb{P}(B > x)$ for any $x \in \mathbb{R}$. For real sequences $\{a_n\}$ and $\{b_n\}$, we say that $a_n = O(b_n)$ if $\limsup_{n\to\infty} |a_n|/b_n < \infty$, $a_n = o(b_n)$ if $\limsup_{n\to\infty} |a_n|/b_n = 0$, and $a_n = \Theta(b_n)$ if $\liminf_{n\to\infty} |a_n|/b_n > 0$. For $a \in \mathbb{R}$, write $a^+ = \max(a, 0)$ and $a^- = \max(-a, 0)$.

3.2 The Model

We consider a classical M-model with possible demand uncertainty as depicted in Figure 3.1. In particular, the model has two customer classes, Class 1 and Class 2, and three pools of servers: two dedicated pools for the two customer classes and one flexible pool that can serve both classes. We allow the arrival rate for Class i, Λ_i, $i = 1, 2$, to be a random variable. For a given realization of

Λ_i, i.e., $\Lambda_i = \lambda_i$, Class i arrivals follow a Poisson process with rate λ_i. Each server pool can have multiple servers. We write n_i for the number of servers in the dedicated pool for Class i, and n_F for the number of servers in the flexible pool. If a customer is served by the dedicated server, its service time follows an exponential distribution with rate μ. If a customer is served by the flexible server, its service time follows an exponential distribution with rate μ_F. We assume $\mu_F \leq \mu$ to account for the potential efficiency loss of flexible servers. Each customer has a patience time that follows an exponential distribution with rate θ. Once a customer's waiting time (in the queue) exceeds its patience time, it abandons the system.

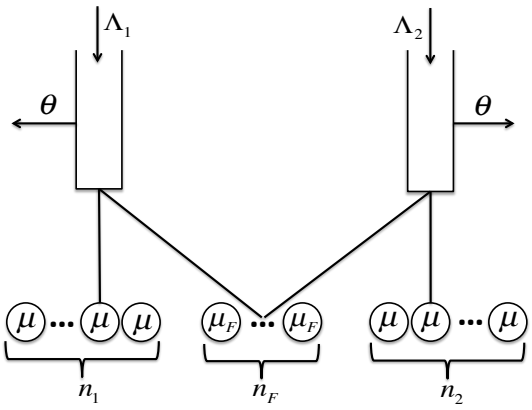


Figure 3.1: The M-model

For $i = 1, 2$, let $X_i(t)$ denote the number of Class i customers in the system at time t. We denote $Z_i(t)$ and $Z_{Fi}(t)$ as the number of dedicated servers and flexible servers serving Class i customers at time t respectively. Note that

$$Z_i(t) \leq n_i, Z_F(t) := Z_{F1}(t) + Z_{F2}(t) \leq n_F, \text{ and } Z_i(t) + Z_{Fi}(t) \leq X_i(t). \tag{3.1}$$

Let $Q_i(t)$ denote the number of Class i customers waiting in the queue at time t. Then $Q_i(t) = X_i(t) - Z_i(t) - Z_{Fi}(t)$. Let $X(t) := (X_1(t), X_2(t))$, $Z(t) := (Z_1(t), Z_2(t), Z_{F1}(t), Z_{F2}(t))$, and $Q(t) := (Q_1(t), Q_2(t))$. We also define the total number of customers in the system and the total queue length processes as $X_\Sigma(t) = X_1(t) + X_2(t)$ and $Q_\Sigma(t) = Q_1(t) + Q_2(t)$ respectively. Let A_i, S_i, S_{Fi}, G_i be independent unit-rate Poisson processes, which will be used to represent the

arrival, departure, and abandonment events respectively. At the beginning of the planning horizon Λ_i is realized. Given $\Lambda_i = \lambda_i$, $X_i(t)$ satisfies the following dynamics:

$$X_i(t) = X_i(0) + A_i\left(\lambda_i t\right) - G_i\left(\theta \int_0^t Q_i(s)\, ds\right) - S_i\left(\mu \int_0^t Z_i(s)\, ds\right) - S_{Fi}\left(\mu_F \int_0^t Z_{Fi}(s)\, ds\right).$$

To fully describe dynamics of the system, we need to specify the scheduling policy – how to allocate the servers. We restrict ourselves to preemptive deterministic Markovian policies, where the allocation of servers $Z(t)$ can be viewed as a function of the current state of the system $X(t)$ [74]. Let ν denote such a policy (mapping), i.e., $Z(t) = \nu(X^\lambda(t))$ and it satisfies the feasibility conditions listed in (3.1).

Let $Q_\Sigma(\infty; n_1, n_2, n_F; \nu)$ be the steady-state total queue length given staffing level (n_1, n_2, n_F) and scheduling policy ν. If the system is not stable under a certain staffing and scheduling rule, we define $Q_\Sigma(\infty; n_1, n_2, n_F; \nu) \equiv \infty$. Our goal is to jointly choose the staffing levels for each pool and the scheduling policy to minimize the sum of the staffing costs and the steady-state average holding and abandonment costs:

$$\min_{n_1, n_2, n_F, \nu} \Pi(n_1, n_2, n_F; \nu) := c(n_1 + n_2) + c_F n_F + (h + a\theta)\mathbb{E}[Q_\Sigma(\infty; n_1, n_2, n_F; \nu)], \tag{3.2}$$

where $c > 0$ is the per server per unit time staffing cost for the dedicated pools, $c_F > 0$ is the per server per unit time staffing cost for the flexible pool, $h > 0$ is the per customer per unit time holding cost, and $a > 0$ is the per customer abandonment cost. Note that the abandonment cost is $a\theta\mathbb{E}[Q_\Sigma(\infty; n_1, n_2, n_F; \nu)]$ because $\theta\mathbb{E}[Q_\Sigma(\infty; n_1, n_2, n_F; \nu)]$ is the rate at which customers abandon in stationarity. We also note that

$$\mathbb{E}[Q_\Sigma(\infty; n_1, n_2, n_F; \nu)] = \mathbb{E}[\mathbb{E}[Q_\Sigma(\infty; n_1, n_2, n_F; \nu)|\Lambda]],$$

where $\mathbb{E}[Q_\Sigma(\infty; n_1, n_2, n_F; \nu)|\Lambda = (\lambda_1, \lambda_2)]$ is the steady-state average queue length of an M-model with arrival rates (λ_1, λ_2), and the outer expectation is taken with respect to the random

arrival rates Λ, i.e., the demand uncertainty.

In order to avoid trivial situations, we impose the following condition on the rates and cost parameters:

$$c/\mu < c_F/\mu_F < h/\theta + a. \tag{3.3}$$

The first inequality ensures that flexible servers have some disadvantage over dedicated servers. Otherwise, we would never staff dedicated servers. The second inequality ensures that the cost of serving a customer using a flexible server is less than the cost of letting the customer wait and abandon. Otherwise, we would never staff flexible servers.

We highlight two challenges in solving (3.2). First, even for a given staffing level, characterizing the optimal scheduling policy can be highly nontrivial. Second, even after pinning down the optimal scheduling policy, it remains difficult to solve for the optimal staffing level due to the lack of an analytical characterization of $\mathbb{E}[Q_\Sigma(\infty; n_1, n_2, n_F; \nu)]$. We will address these challenges in subsequent sections. In particular, a lot of our developments rely on a heavy-traffic asymptotic mode of analysis, in which our goal is to characterize how the optimal decisions scale with the arrival rate (average arrival rate) λ as $\lambda \to \infty$. To explicitly mark the dependence of the policies and system dynamics on λ, we use the superscript λ. For example, ν^λ and $(n_1^\lambda, n_2^\lambda, n_F^\lambda)$ are the scheduling policy and staffing levels for the system with the arrival rate parameter λ (i.e., the λ-th system). Similarly, X^λ, Z^λ, and Q^λ are the number-in-system, number-in-service, and number-in-queue processes of the λ-th system.

3.3 The Case with Deterministic Arrival Rate

In this section, we study a special case of the system where the arrival rate is deterministic. In particular, we assume $\Lambda_1 = \Lambda_2 = \lambda$ with probability 1. In this case, we have a symmetric M-model. The goal is to highlight how to strike a balance between the cost and benefit of flexibility.

We start by providing an overview of how we address the two challenges listed in Section 3.2 to derive the optimal staffing and scheduling rules jointly. In Section 3.3.1, we use a coupling argu-

ment to derive the optimal scheduling policy for any given staffing level. This optimal scheduling rule turns out to have a very neat and intuitive structure. In particular, the policy prioritizes the faster servers (the dedicated servers) when routing customers to servers, and the flexible servers prioritize the class with more customers in the system (the larger $X_i(t)$). This is similar in structure to the queue-idleness ratio policy [6], the fastest server first policy [60], and the max-pressure policy [34]. However, we emphasize that using the coupling argument, we are able to establish exact optimality instead of asymptotic optimality. Moreover, in the staffing regime we are interested in, there is no state-space collapse. Thus, the asymptotic optimality framework leveraged in the literature can no longer be applied.

Next, in Section 3.3.2, we take a heavy-traffic asymptotic approach to derive a necessary and sufficient characterization of the optimal staffing rules. In particular, we gradually send the arrival rate λ to infinity and study how the optimal staffing level scales with λ. Our analysis shows that the optimal staffing rule leads the system to operate in the QED regime, and the optimal size of the flexible pool is $O(\sqrt{\lambda})$. This extends the insights developed in [23] to the many-server setting.

Due to the symmetry of the system, we assume, without loss of optimality, that $n_1^\lambda = n_2^\lambda = n^\lambda$. Thus, our decision variables for the staffing rule reduce to n^λ and n_F^λ. For the model analyzed in this section, we allow $\theta = 0$, i.e., no abandonment. When $\theta = 0$, we need to put more restrictions on the staffing levels to ensure system stability. In particular, we define

$$\Omega^\lambda(0) := \left\{(n^\lambda, n_F^\lambda) \in \mathbb{N}_0^2 : 2\lambda < 2n^\lambda\mu + n_F^\lambda\mu_F\right\}.$$

The following lemma show that when $\theta = 0$, having $(n^\lambda, n_F^\lambda) \in \Omega^\lambda(0)$ ensures that the system is stable under the optimal scheduling rule.

Lemma 2. *If $\theta = 0$, for any $(n^\lambda, n_F^\lambda) \in \Omega^\lambda(0)$ there exists a scheduling policy ν^λ, under which the stochastic process X^λ has a unique stationary distribution.*

To ensure consistent notation, for $\theta > 0$ we define

$$\Omega^\lambda(\theta) = \{(n^\lambda, n_F^\lambda) \in \mathbb{N}_0^2\}.$$

3.3.1 Optimal Scheduling Rule

Intuitively, a good scheduling policy should reduce the queues as fast as possible and balance the queues of the two classes. This motivates the following scheduling rule. For the dedicated pool of servers,

$$Z_i^\lambda(t) = \min\{n^\lambda, X_i^\lambda(t)\} \text{ for } i = 1, 2; \tag{3.4}$$

and for the flexible pool of servers, if $X_1^\lambda(t) \geq X_2^\lambda(t)$,

$$Z_{F1}^\lambda(t) = \min\{n_F^\lambda, (X_1^\lambda(t) - n^\lambda)^+\}, \quad Z_{F2}^\lambda(t) = \min\{n_F^\lambda - Z_{F1}^\lambda(t), (X_2^\lambda(t) - n^\lambda)^+\}; \tag{3.5}$$

otherwise,

$$Z_{F1}^\lambda(t) = \min\{n_F^\lambda - Z_{F2}^\lambda(t), (X_1^\lambda(t) - n^\lambda)^+\}, \quad Z_{F2}^\lambda(t) = \min\{n_F^\lambda, (X_2^\lambda(t) - n^\lambda)^+\}. \tag{3.6}$$

Note that under this policy, we first try to assign as many customers to the dedicated pools as possible, i.e., (3.4). Then, for the flexible pool, we give priority to the class with more customers in the system, i.e., (3.5) and (3.6). We comment that for our scheduling policy, ties can be broken in an arbitrary way. For simplicity of exposition, we assume that when $X_1^\lambda(t) = X_2^\lambda(t)$, the flexible pool gives priority to Class 1. We denote the policy defined in (3.4) - (3.6) as $\nu^{\lambda,*}$.

The next theorem shows that when $\theta \leq \mu_F$, for any fixed staffing level (n^λ, n_F^λ), $\nu^{\lambda,*}$ is optimal.

Theorem 5. *Suppose $\theta \leq \mu_F$. For any Markovian scheduling policy ν^λ,*

$$\mathbb{E}[Q_\Sigma^\lambda(\infty; n^\lambda, n_F^\lambda; \nu^\lambda)] \geq \mathbb{E}[Q_\Sigma^\lambda(\infty; n^\lambda, n_F^\lambda; \nu^{\lambda,*})],$$

which implies that $\Pi^\lambda(n^\lambda, n_F^\lambda; \nu^\lambda) \geq \Pi^\lambda(n^\lambda, n_F^\lambda; \nu^{\lambda,*})$.

Note that for $\theta \leq \mu_F$, the policy $\nu^{\lambda,*}$ tries to equalize X_1^λ and X_2^λ at the maximum rate. Due to the symmetry in the system structure, we expect this policy to perform well. We prove the theorem by developing a coupling construction based on the transition rates of the underlying Markov processes (see Appendix B.2.2 for more details). We also comment that the condition $\theta \leq \mu_F$ is necessary for $\nu^{\lambda,*}$ to be optimal. If $\theta > \mu_F$, $\nu^{\lambda,*}$ no longer equalizes X_1^λ and X_2^λ at the maximum rate, because a larger rate can be attained by keeping customers waiting in the queue instead of sending them to the flexible severs. Indeed, when $\theta \geq \mu_F = \mu$, we can show that a scheduling rule that prioritizes the class with fewer customers in the system is optimal (see Theorem 14 in Appendix B.2.3).

3.3.2 Asymptotically Optimal Staffing Rule

Based on the analysis in Section 3.3.1, the scheduling policy $\nu^{\lambda,*}$ is optimal for any λ and (n^λ, n_F^λ) when $\theta \leq \mu_F$. In subsequent analysis, we assume without loss of optimality that the policy $\nu^{\lambda,*}$ is employed. When there is no confusion, we will omit the scheduling policy from the notation of the corresponding stochastic processes. Now, the problem of jointly optimizing staffing and scheduling rules, i.e., (3.2), reduces to optimizing the staffing levels only:

$$\min_{(n^\lambda, n_F^\lambda) \in \Omega^\lambda(\theta)} \Pi^\lambda(n^\lambda, n_F^\lambda) := 2cn^\lambda + c_F n_F^\lambda + (h + a\theta)\mathbb{E}[Q_\Sigma^\lambda(\infty; n^\lambda, n_F^\lambda)]. \tag{3.7}$$

Solving (3.7) analytically is still challenging due to the lack of a closed-form expression for $\mathbb{E}[Q_\Sigma^\lambda(\infty; n^\lambda, n_F^\lambda)]$. In this section, we study the structure of the optimal staffing levels under heavy traffic. In particular, we send $\lambda \to \infty$ while keeping the service rates and abandonment rates fixed. Our analysis reveals how the optimal sizes of the dedicated pool and flexible pool scale with the arrival rate λ.

Let

$$\Pi^{\lambda,*} := \min_{(n^\lambda, n_F^\lambda) \in \Omega^\lambda(\theta)} \Pi^\lambda(n^\lambda, n_F^\lambda) \text{ and } (n^{\lambda,*}, n_F^{\lambda,*}) \in \operatorname*{arg\,min}_{(n^\lambda, n_F^\lambda) \in \Omega^\lambda(\theta)} \Pi^\lambda(n^\lambda, n_F^\lambda).$$

Define $R^\lambda := \lambda/\mu$, which is the offered load of Class i, $i = 1, 2$.

Lemma 3. *Suppose* (3.3) *holds. Then,* $\Pi^{\lambda,*} = 2cR^\lambda + O(\sqrt{\lambda})$. *Moreover, for* $(n^{\lambda,*}, n_F^{\lambda,*})$,

$$-\infty < \liminf_{\lambda\to\infty} \frac{n^{\lambda,*} - R^\lambda}{\sqrt{\lambda}} \le \limsup_{\lambda\to\infty} \frac{n^{\lambda,*} - R^\lambda}{\sqrt{\lambda}} < \infty$$

and

$$\limsup_{\lambda\to\infty} \frac{n_F^{\lambda,*}}{\sqrt{\lambda}} < \infty.$$

Motivated by Lemma 3, our goal in subsequent analysis is to close the $O(\sqrt{\lambda})$ optimality gap. In particular, we employ the following notion of asymptotic optimality.

Definition 1. *A sequence of staffing levels* (n^λ, n_F^λ) *(indexed by* λ*) is asymptotically optimal if*

$$\Pi^\lambda(n^\lambda, n_F^\lambda) = \Pi^{\lambda,*} + o(\sqrt{\lambda}).$$

The key question we would like to address is how much flexibility is optimal. We first note that if there is no 'cost' of flexibility, then we would want as much flexibility as possible. This is because flexible servers create resource pooling in the system. To be more precise, we have the following result:

Lemma 4. *When* $\mu = \mu_F \ge \theta$, *we have*

$$\mathbb{E}[Q_\Sigma^\lambda(\infty; n^\lambda, n_F^\lambda)] \ge \mathbb{E}[Q_\Sigma^\lambda(\infty; 0, 2n^\lambda + n_F^\lambda)].$$

In practice, flexibility often comes at a cost. Here, we consider two forms of cost: a higher staffing cost, i.e., $c_F \ge c$, and an efficiency cost, i.e., $\mu_F \le \mu$. In this case, Lemma 3 indicates that, overall, it is optimal to follow the square-root staffing rule, i.e., $2n^{\lambda,*} + n_F^{\lambda,*} = 2R^\lambda + O(\sqrt{\lambda})$. More importantly, $n_F^{\lambda,*}$ cannot be too large, i.e., $n_F^{\lambda,*} = O(\sqrt{\lambda})$.

To derive an asymptotically optimal staffing rule, we need to have a good approximation of $\mathbb{E}[Q_\Sigma^\lambda(\infty; n^\lambda, n_F^\lambda)]$ in (3.7). In the many-server heavy-traffic analysis, there are two commonly used approximations: the fluid approximation and the diffusion approximation. To achieve $o(\sqrt{\lambda})$

optimality, we need to use the finer-scale diffusion approximation. We next define some diffusion-scaled processes. Let

$$\hat{X}_i^\lambda(t) = \frac{X_i^\lambda(t) - n^\lambda}{\sqrt{\lambda}},\ \hat{Q}_i^\lambda(t) = \frac{Q_i^\lambda(t)}{\sqrt{\lambda}},\ \text{ and }\ \hat{Z}_i^\lambda(t) = \frac{Z_i^\lambda(t) - n^\lambda}{\sqrt{\lambda}} \text{ for } i = 1, 2.$$

We also write $\hat{X}^\lambda = (\hat{X}_1^\lambda, \hat{X}_2^\lambda)$, and define

$$\hat{X}_\Sigma^\lambda(t) = \frac{X_\Sigma^\lambda(t) - 2n^\lambda - n_F^\lambda}{\sqrt{\lambda}} \text{ and } \hat{Q}_\Sigma^\lambda(t) = \frac{Q_\Sigma^\lambda(t)}{\sqrt{\lambda}}.$$

In our subsequent development, for any stochastic process $Y(t)$, we write $Y(\infty)$ as its stationary distribution.

Recall that $n_F^{\lambda,*} = O(\sqrt{\lambda})$. The following theorem characterizes the diffusion limit of the number-in-system processes in this case.

Theorem 6. *For $(n^\lambda, n_F^\lambda) \in \Omega^\lambda(\theta)$, suppose $n^\lambda = R^\lambda + \beta\sqrt{R^\lambda} + o(\sqrt{R^\lambda})$ and $n_F^\lambda = \beta_F\sqrt{R^\lambda} + o(\sqrt{R^\lambda})$, where $\beta \in \mathbb{R}, \beta_F \geq 0$, and if $\theta = 0$, $2\beta\mu + \beta_F\mu_F > 0$. Then, if $\hat{X}^\lambda(0) \Rightarrow \hat{X}(0)$ as $\lambda \to \infty$,*

$$\hat{X}^\lambda \Rightarrow \hat{X} \text{ in } D^2 \text{ as } \lambda \to \infty,$$

where $\hat{X}$ is a two-dimensional diffusion process with

$$d\hat{X}_i(t) = \left(-\beta\sqrt{\mu} + \mu\hat{X}_i(t)^- - (\mu_F - \theta) f_i\left(\hat{X}_1(t), \hat{X}_2(t)\right) - \theta\hat{X}_i(t)^+\right) dt + \sqrt{2}\, dB_i(t),$$

for $i = 1, 2$, B_1 and B_2 are independent standard Brownian motions, and

$$f_1(x_1, x_2) = \begin{cases} x_1^+ \wedge \frac{\beta_F}{\sqrt{\mu}} & \text{if } x_1 \geq x_2, \\ x_1^+ \wedge \left(\frac{\beta_F}{\sqrt{\mu}} - x_2^+\right)^+ & \text{if } x_1 < x_2; \end{cases} \quad f_2(x_1, x_2) = \begin{cases} x_2^+ \wedge \left(\frac{\beta_F}{\sqrt{\mu}} - x_1^+\right)^+ & \text{if } x_1 \geq x_2, \\ x_2^+ \wedge \frac{\beta_F}{\sqrt{\mu}} & \text{if } x_1 < x_2. \end{cases}$$

Moreover,

$$\mathbb{E}[\hat{Q}_\Sigma^\lambda(\infty)] \to \mathbb{E}[(\hat{X}_1(\infty)^+ + \hat{X}_2(\infty)^+ - \beta_F/\sqrt{\mu})^+] \text{ as } \lambda \to \infty.$$

We make two important observations from Theorem 6. First, to characterize $\hat{X}^\lambda_\Sigma$, we need to keep track of a two-dimensional diffusion process $\hat{X}$ in the limit. In this sense, we do not achieve complete resource pooling. On the other hand, the drift terms of $\hat{X}$ cannot be fully decomposed along each dimension, i.e., $f_i(x_1, x_2)$ depends on both x_1 and x_2. Thus, we achieve partial resource pooling. Second, $\mathbb{E}[(\hat{X}_1(\infty)^+ + \hat{X}_2(\infty)^+ - \beta_F/\sqrt{\mu})^+]$ serves as a good approximation for $\mathbb{E}[\hat{Q}^\lambda_\Sigma(\infty)]$, which suggests approximating $\min_{(n^\lambda, n^\lambda_F)\in\Omega^\lambda(\theta)}(\Pi^\lambda(n^\lambda, n^\lambda_F) - 2cR^\lambda)/\sqrt{\lambda}$ by the following optimization problem:

$$\begin{aligned}\min_{(\beta,\beta_F)\in\hat{\Omega}(\theta)} \hat{V}_p(\beta, \beta_F) :=& 2c\beta/\sqrt{\mu} + c_F\beta_F/\sqrt{\mu} \\ &+ (h + a\theta)\mathbb{E}\left[\left(\hat{X}_1(\infty; \beta, \beta_F)^+ + \hat{X}_2(\infty; \beta, \beta_F)^+ - \beta_F/\sqrt{\mu}\right)^+\right],\end{aligned} \tag{3.8}$$

where, if $\theta = 0$,

$$\hat{\Omega}(0) := \{(\beta, \beta_F) : \beta \in \mathbb{R}, \beta_F \geq 0, 2\beta\mu + \beta_F\mu_F > 0\},$$

and, if $\theta > 0$,

$$\hat{\Omega}(\theta) := \{(\beta, \beta_F) : \beta \in \mathbb{R}, \beta_F \geq 0\}.$$

We do not have a closed-form expression for $\mathbb{E}\left[\left(\hat{X}_1(\infty; \beta, \beta_F)^+ + \hat{X}_2(\infty; \beta, \beta_F)^+ - \beta_F/\sqrt{\mu}\right)^+\right]$. Thus, (3.8) can only be solved numerically. In Figure 3.2, we plot $\hat{V}_p(\beta, \beta_F)$ for different values of β and β_F. We observe that $\hat{V}_p(\beta, \beta_F)$ is convex and is minimized at $(0.5, 0.5)$ in this example.

We next characterize the optimal staffing rule by rigorously drawing the connection between the solution of the optimal staffing problem (3.7) and the diffusion optimization problem (3.8). Due to the lack of an analytical solution for $\hat{V}_p(\beta, \beta_F)$, we impose the following technical assumption.

Assumption 4. *The set* $\arg\min_{(\beta,\beta_F)\in\hat{\Omega}(\theta)} \hat{V}_p(\beta, \beta_F)$ *is non-empty and finite.*

Theorem 7. *For $\theta \leq \mu_F \leq \mu$, under Assumption 4, a sequence of staffing policies (n^λ, n^λ_F) is asymptotically optimal if and only if the following two conditions hold:*

1. $n^\lambda = R^\lambda + \beta^\lambda\sqrt{R^\lambda} + o(\sqrt{R^\lambda})$
2. $n^\lambda_F = \beta^\lambda_F\sqrt{R^\lambda} + o(\sqrt{R^\lambda})$

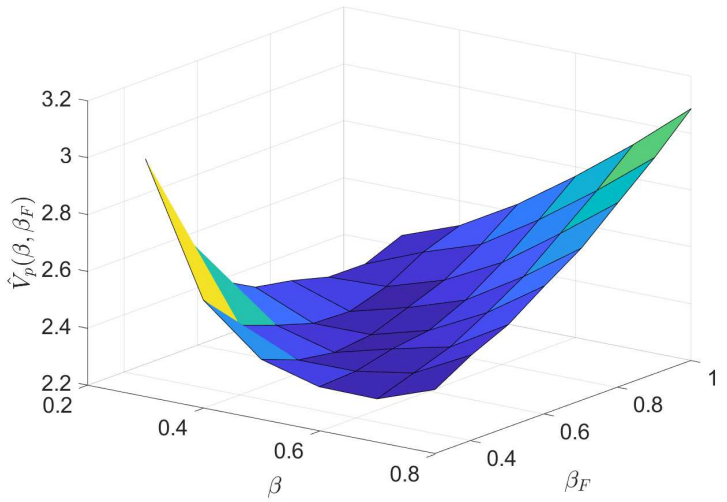


Figure 3.2: $\hat{V}_p(\beta, \beta_F)$ as a function of β and β_F. ($\mu = 1, \mu_F = 0.85, \theta = 0, c = 1, c_F = 1.4, h = 1$)

where $(\beta^\lambda, \beta_F^\lambda) \in \arg\min_{(\beta,\beta_F)\in\hat{\Omega}(\theta)} \hat{V}_p(\beta, \beta_F)$.

Remark 2. *If $\hat{V}_p(\beta, \beta_F)$ has a unique minimizer (β^*, β_F^*), then the asymptotically optimal staffing levels satisfy $n^\lambda = R^\lambda + \beta^*\sqrt{R^\lambda} + o(\sqrt{R^\lambda})$ and $n_F^\lambda = \beta_F^*\sqrt{R^\lambda} + o(\sqrt{R^\lambda})$.*

To illustrate why $n_F^\lambda = O(\sqrt{\lambda})$ is necessary for asymptotic optimality, we plot $\mathbb{E}[Q_\Sigma^\lambda(\infty; (60 - n_F^\lambda)/2, n_F^\lambda)]$ as a function of n_F^λ in Figure 3.3. We set $\lambda = 25$ and test two different scenarios for the service rate when $\theta = 0$. In the left plot, $\mu = \mu_F = 1$. In the right plot, $\mu = 1$ while $\mu_F = 0.85$. The stationary queue lengths are estimated through simulation. The simulation errors (estimated using the batch means method) are less than 0.01 and hence omitted. When $\mu = \mu_F$ (left plot in Figure 3.3), we observe that increasing n_F^λ beyond $2\sqrt{\lambda} = 10$ has almost no effect on the stationary total queue length. In this case, if $c < c_F$, the staffing cost increases linearly with n_F^λ while the holding cost does not decrease much as n_F^λ increases beyond 10. Thus, the optimal n_F^λ cannot be too large. When $\mu > \mu_F$ (right plot in Figure 3.3), the stationary total queue length is not monotone in n_F^λ. The minimum is achieved at a relatively small value of n_F^λ, i.e., $n_F^\lambda = 6$. Therefore, the optimal n_F^λ cannot be too large in this case as well.

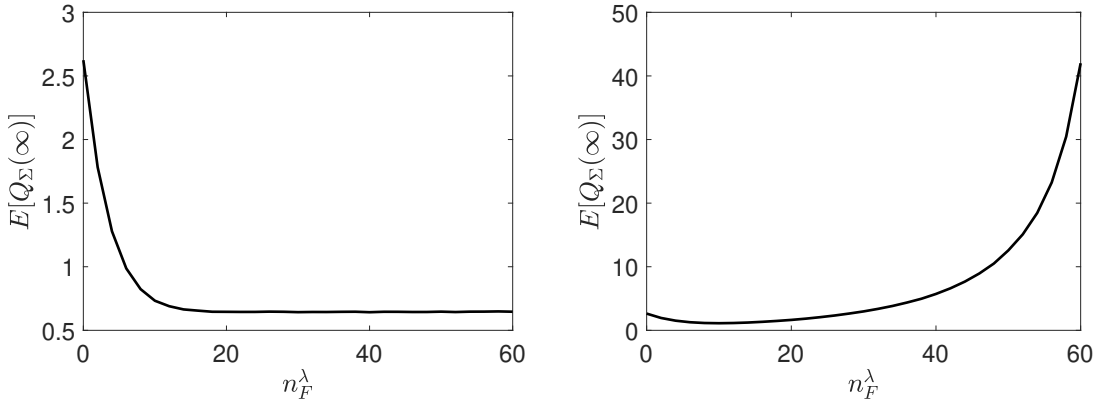


Figure 3.3: $\mathbb{E}[Q_\Sigma(\infty; (60 - n_F^\lambda)/2, n_F^\lambda)]$ as a function of n_F^λ. Left: $\mu = \mu_F = 1$; Right $\mu = 1$, $\mu_F = 0.85$. ($\lambda = 25, \theta = 0$)

We conclude this section with some sensitivity analysis on β^* and β_F^*. Let $h = c = 1$, $\mu = 1$, and $\theta = 0$. Note that setting $c = 1$ and $\mu = 1$ is without loss of generality as it is equivalent to choosing units for cost and time. We first test how (β^*, β_F^*) varies with c_F, when $\mu_F = 0.85$. Table 3.1 shows one such experiment. We observe that β^* is increasing in c_F while β_F^* is decreasing in c_F. When c_F is large, i.e., $c_F \geq 1.6$, $\beta_F^* = 0$, suggesting it becomes too expensive to use the flexible servers then.

c_F	1	1.2	1.4	1.6	1.8
β^*	−0.2	0.2	0.5	0.9	0.9
β_F^*	1.9	1.1	0.5	0	0

Table 3.1: Sensitivity of (β^*, β_F^*) with respect to c_F

We next test how (β^*, β_F^*) varies with μ_F, when $c_F = 1.4$. Table 3.2 shows one such experiment. We observe that β^* is decreasing in μ_F while β_F^* is increasing in μ_F. For very small values of μ_F, i.e., $\mu_F \leq 0.55$, flexible servers are too inefficient to be staffed.

μ_F	0.55	0.65	0.75	0.85	0.95
β^*	0.8	0.8	0.7	0.5	0.4
β_F^*	0	0.1	0.2	0.5	0.6

Table 3.2: How optimal (β, β_F) varies with μ_F

3.4 The Case with Demand Uncertainty

In this section, we study the joint staffing and scheduling optimization problem (3.2) with random arrival rates. We assume $\Lambda_i = p_i\lambda + \lambda^{\alpha_i} Y_i$, where $p_i > 0$, $1/2 < \alpha_i \leq 1$, and Y_i is a random variable with $\mathbb{E}[Y_i] = 0$ and $\text{Var}(Y_i) = \sigma_i^2 < \infty$. As Λ_i is an arrival rate, we assume $\Lambda_i \geq 0$ with probability 1. For example, when $\alpha_i = 1$, we assume $Y_i \geq -p_i$ with probability 1. For ease of exposition, we also assume Y_i's are continuous random variables with strictly increasing marginal cdf on their domains of definition. We allow Y_1 and Y_2 to be dependent and denote by g their joint density. Without loss of generality, we assume $\alpha_1 \geq \alpha_2$. For the analysis in this section, we also require $\theta > 0$ to ensure system stability regardless of the realized arrival rates.

We next make some comments about our modeling assumptions for this section. First, we allow quite some asymmetry between the two classes. In particular, p_i's, α_i's, and the marginal distribution of Y_i's can be different for the two classes. This implies that the optimal n_1 and n_2 might be different. Second, the mean of Λ_i is of order λ while the standard deviation of Λ_i is of order λ^{α_i}. For queues with deterministic arrival rate λ, our analysis in Section 3.3 reveals that the stochastic fluctuation of the system is of order $\lambda^{1/2}$. In this section, we are interested in the case where $\alpha_i > 1/2$, so that the demand uncertainty is of a larger order of magnitude than the stochastic fluctuation of the system.

We start by providing an overview of how we address the two challenges listed Section 3.2 to derive the optimal scheduling and staffing rules jointly. When facing demand uncertainty, solving (3.2) analytically is more challenging than the case with deterministic arrival rates. This is because we now face two sources of randomness: One is the parameter uncertainty, the other is the stochas-

ticity of the queue, i.e., random interarrival, service, and patience times. In this section, we again take a heavy-traffic asymptotic approach where we send the arrival rate parameter λ to infinity and quantify how the optimal staffing rule scales with λ. Under the assumption that $\alpha_i > 1/2$, we employ a stochastic-fluid approximation where we suppress the stochastic fluctuation of the queues and focus on parameter uncertainty only [66]. In our setting, the stochastic-fluid optimal staffing problem is a special case of the single-period multi-product inventory problem with demand substitution [75]. Based on the stochastic-fluid staffing solution, it also becomes easier to develop good scheduling policies. We then show that the staffing and scheduling rules derived from the stochastic fluid problem achieve an $O(\lambda^{1-\alpha_2})$ optimality gap.

The key intuition behind our development is that when parameter uncertainty dominates system stochasticity, optimally hedging against parameter uncertainty is more important. Indeed, with high probability, the system with realized arrival rate is no longer in the QED regime. In these cases, any "fluid-optimal" scheduling policy is "good enough". We will make these intuitions more precise in the subsequent development.

3.4.1 Stochastic-Fluid Optimization Problem

For our model, the rate of customer abandonment can be expressed as

$$\theta \mathbb{E}[Q_\Sigma(\infty; n_1, n_2, n_F)].$$

By rate conservation, the rate of customer abandonment can also be approximated by

$$\mathbb{E}\left[((\Lambda_1 - n_1\mu)^+ + (\Lambda_2 - n_2\mu)^+ - n_F\mu_F)^+\right].$$

Thus, we can approximate the steady-state queue length by

$$\frac{1}{\theta}\mathbb{E}\left[((\Lambda_1 - n_1\mu)^+ + (\Lambda_2 - n_2\mu)^+ - n_F\mu_F)^+\right].$$

This allows us to approximate (3.2) by the following stochastic-fluid optimization problem

$$\begin{aligned}\min_{\tilde{n}_1\geq 0,\tilde{n}_2\geq 0,\tilde{n}_F\geq 0} \tilde{\Pi}(\tilde{n}_1,\tilde{n}_2,\tilde{n}_F) :=& c(\tilde{n}_1+\tilde{n}_2)+c_F\tilde{n}_F \\ &+(h/\theta+a)\mathbb{E}\left[((\Lambda_1-\tilde{n}_1\mu)^+ +(\Lambda_2-\tilde{n}_2\mu)^+ -\tilde{n}_F\mu_F)^+\right].\end{aligned} \tag{3.9}$$

For (3.9), we relax the integer requirement on n_1, n_2, n_F and only require them to be non-negative. We denote its optimal solution as $(\tilde{n}_1^*, \tilde{n}_2^*, \tilde{n}_F^*)$ and the optimal value as $\tilde{\Pi}^*$.

The optimization problem (3.9) can be viewed as a special case of the single-period multi-product inventory management problem with demand substitution, i.e., demand Λ_i is best met by dedicated resources, but may also be met by flexible resources if there is a shortfall of dedicated resources. This class of inventory management problems has been studied in the literature in much more general forms [76, 77, 78, 75, 79]. Restricting it to our special setting allows us to derive more analytical insights.

To simplify the notation, we define $c_P := h/\theta + a$, i.e., the performance cost. Let q_i denote the solution of the following equation

$$\mathbb{P}(Y_i > q_i) = \frac{c}{c_P\mu}.$$

We first study the case where $\alpha_1 = \alpha_2 = \alpha$. If $\mathbb{P}(Y_1 > q_1 \text{ or } Y_2 > q_2) > \frac{c_F}{c_P\mu_F}$, let $r_1, r_2 \in \mathbb{R}$, and $r_F > 0$ denote the solution to the following system of equations:

$$\begin{aligned}\mathbb{P}(Y_1 > r_1, Y_1 - r_1 + (Y_2 - r_2)^+ > r_F) &= \frac{c}{c_P\mu},\\ \mathbb{P}((Y_1 - r_1)^+ + (Y_2 - r_2)^+ > r_F) &= \frac{c_F}{c_P\mu_F}.\end{aligned}$$

The next lemma characterizes the optimal solution to (3.9) when $\alpha_1 = \alpha_2$.

Lemma 5. *Suppose $\alpha_1 = \alpha_2 = \alpha$.*

If $\mathbb{P}(Y_1 > q_1 \text{ or } Y_2 > q_2) \leq \frac{c_F}{c_P\mu_F}$, $\tilde{n}_i^* = (p_i\lambda + q_i\lambda^\alpha)/\mu$ for $i = 1, 2$, and $\tilde{n}_F^* = 0$.

If $\mathbb{P}(Y_1 > q_1 \text{ or } Y_2 > q_2) > \frac{c_F}{c_P\mu_F}$, $\tilde{n}_i^* = (p_i\lambda + r_i\lambda^\alpha)/\mu$ for $i = 1, 2$, and $\tilde{n}_F^* = r_F\lambda^\alpha/\mu_F$.

Lemma 5 reveals that the optimal solution to (3.9) has a very neat structure. The optimal

number of dedicated servers involves a baseline level to meet the mean demand, $p_i\lambda/\mu$, and an uncertainty hedging of order λ^α. We also note that the size of the flexible pool is $O(\lambda^\alpha)$, indicating that the flexible pool is mostly used for uncertainty hedging.

We next consider the case where $\alpha_1 > \alpha_2$. In this case, we do not have explicit expressions for $\tilde{n}_i^*$'s and $\tilde{n}_F^*$ as in Lemma 5. However, (3.9) can still be solved numerically very efficiently, as it is a convex optimization problem. In addition, we can derive structural insights into the optimal staffing levels. When $\mathbb{P}(Y_1 > q_1 \text{ or } Y_2 > q_2) > \frac{c_F}{c_P\mu_F}$, define $l \in \mathbb{R}, l_F > 0$ to be the solution to the following system of equations:

$$\mathbb{P}(Y_2 > l + l_F \text{ or } \{Y_1 > q_1, Y_2 > l\}) = \frac{c}{c_P\mu},$$
$$\mathbb{P}(Y_1 > q_1 \text{ or } Y_2 > l + l_F) = \frac{c_F}{c_P\mu_F}.$$

The next lemma characterizes the optimal solution to (3.9) when $\alpha_1 > \alpha_2$.

Lemma 6. *Suppose $\alpha_1 > \alpha_2$.*
If $\mathbb{P}(Y_1 > q_1 \text{ or } Y_2 > q_2) \le \frac{c_F}{c_P\mu_F}$, $\tilde{n}_i^* = (p_i\lambda + q_i\lambda^{\alpha_i})/\mu + o(\lambda^{\alpha_i})$ for $i = 1, 2$ and $\tilde{n}_F^* = o(\lambda^{\alpha_2})$.
If $\mathbb{P}(Y_1 > q_1 \text{ or } Y_2 > q_2) > \frac{c_F}{c_P\mu_F}$, $\tilde{n}_1^* = (p_1\lambda + q_1\lambda^{\alpha_1})/\mu + o(\lambda^{\alpha_1})$, $\tilde{n}_2^* = (p_2\lambda + l\lambda^{\alpha_2})/\mu + o(\lambda^{\alpha_2})$ and $\tilde{n}_F^* = l_F\lambda^{\alpha_2}/\mu_F + o(\lambda^{\alpha_2})$.

We note from Lemma 6 that the optimal size of the dedicated pool again contains a baseline level to meet the mean demand and an uncertainty hedging. The size of the flexible pool is $O(\lambda^{\alpha_2})$. As $\alpha_2 < \alpha_1$, the hedging functionality of the flexible pool is targeted for the less uncertain class.

We conduct numerical sensitivity analysis. For illustration, consider $p_1 = p_2 = 1$, $\alpha_1 = \alpha_2 = \alpha$, $Y_1 = Z_1$, and $Y_2 = \rho Z_1 + \sqrt{1-\rho^2}Z_2$, where Z_1 and Z_2 are independent standard Normal random variables. In this case, $\text{Cor}(Y_1, Y_2) = \rho$. Due to the symmetry of the two classes, we have $q_1^* = q_2^* := q^*$, $\tilde{n}_i^* = \lambda/\mu + q^*\lambda^\alpha/\mu$, and $\tilde{n}_F^* = q_F^*\lambda^\alpha/\mu_F$. Figures 3.4 and 3.5 show how q^* and q_F^* vary with ρ for different values of c_F or μ_F. We note that as ρ increases, q_F^* decreases while q^* increases. This is because when the demand of the two classes are highly positively correlated, there is not much room for load-balancing. We also observe in Figure 3.4 that for a fixed value of

ρ, the higher the cost of flexible servers, the smaller the value of q_F^*. Similarly, the less efficient the flexible servers, the smaller the value of q_F^* (see Figure 3.5).

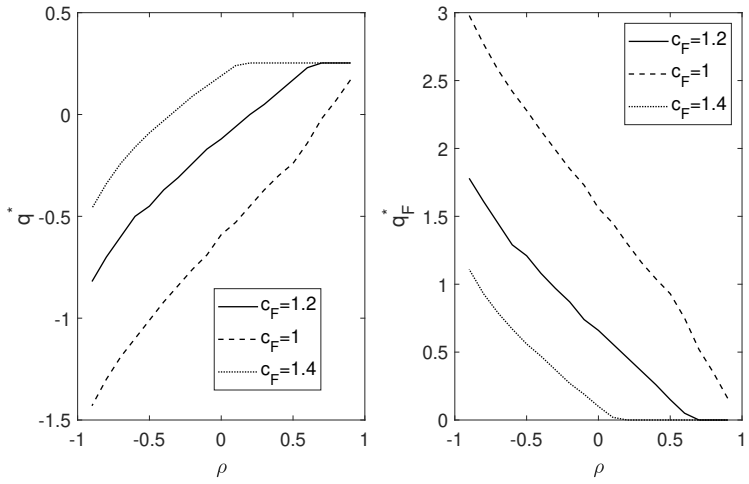


Figure 3.4: How q^* and q_F^* vary with ρ when $\mu_F = 0.9$ and $c_F \in \{1, 1.2, 1.4\}$

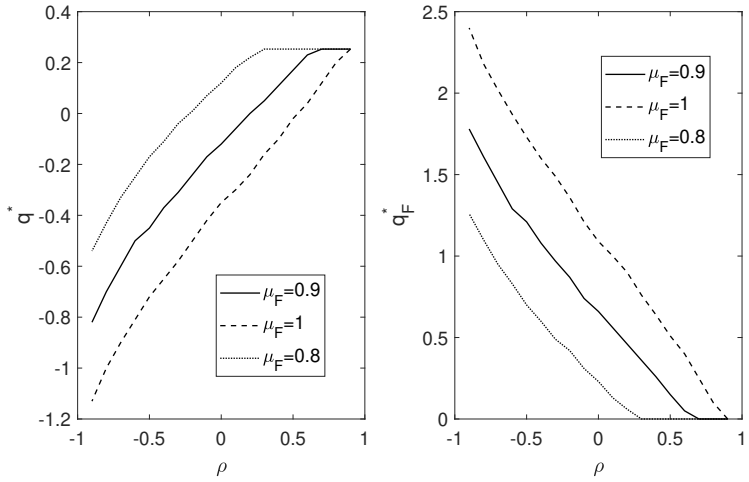


Figure 3.5: How q^* and q_F^* vary with ρ when $c_F = 1.2$ and $\mu_F \in \{0.8, 0.9, 1\}$

3.4.2 Asymptotically Optimal Staffing and Scheduling Rules

In this section, we quantify the quality of the staffing rule derived from the stochastic-fluid approximation (3.9). We also develop a corresponding scheduling rule.

Consider a sequence of systems indexed by λ. The superscript λ is used to denote the quantities

related to the λ-th system. For example, $\tilde{n}^{\lambda,*}, \tilde{n}_F^{\lambda,*}$ is the stochastic-fluid optimal solution when $\Lambda_i = p_i\lambda + \lambda^{\alpha_i} Y_i$ for Class i, $i = 1, 2$.

Our proposed staffing rule for the λ-th system is $(\lceil \tilde{n}_1^{\lambda,*} \rceil, \lceil \tilde{n}_2^{\lambda,*} \rceil, \lfloor \tilde{n}_F^{\lambda,*} \rfloor)$. We next introduce a scheduling policy. Given a realization of the arrival rate $\Lambda = \gamma := (\gamma_1, \gamma_2)$, let $\delta(\gamma) \in [0, 1]$ be a solution of

$$((\gamma_1 - n_1^\lambda \mu)^+ + (\gamma_2 - n_2^\lambda \mu)^+ - n_F^\lambda \mu_F)^+ = (\gamma_1 - n_1^\lambda \mu - \delta n_F^\lambda \mu_F)^+ + (\gamma_2 - n_2^\lambda \mu - (1-\delta) n_F^\lambda \mu_F)^+. \quad (3.10)$$

Note that the solution to (3.10) may not be unique. When (3.10) has multiple optimal solutions, we can set $\delta(\gamma)$ to be any one of them. For a fixed $\delta(\gamma)$, the scheduling policy $\tilde{\nu}^\lambda$ allocates $\lfloor \delta(\gamma) n_F^\lambda \rfloor$ flexible servers to Class 1 and the remaining $\lceil (1 - \delta(\gamma)) n_F^\lambda \rceil$ flexible servers to Class 2. When assigning customers to servers, the dedicated servers are prioritized over the flexible servers. That is, upon each realization of the arrival rates $\Lambda = \gamma$, the policy $\tilde{\nu}^\lambda$ turns the M-model into two independent inverted-V models. For each inverted-V model, we follow the fastest-server-first policy.

To quantify the optimality gap of the stochastic-fluid based policies, we first quantify the difference between Π^λ defined in (3.2) and $\tilde{\Pi}^\lambda$ defined in (3.9).

Lemma 7. *For $\theta > 0$, $\alpha_1 \geq \alpha_2 > 1/2$, and $n_i^\lambda + n_F^\lambda = \Theta(\lambda)$, for any scheduling policy ν^λ,*

$$\tilde{\Pi}^\lambda(n_1^\lambda, n_2^\lambda, n_F^\lambda) \leq \Pi^\lambda(n_1^\lambda, n_2^\lambda, n_F^\lambda; \nu^\lambda).$$

For policy $\tilde{\nu}^\lambda$, we also have

$$\Pi^\lambda(n_1^\lambda, n_2^\lambda, n_F^\lambda; \tilde{\nu}^\lambda) \leq \tilde{\Pi}^\lambda(n_1^\lambda, n_2^\lambda, n_F^\lambda) + O(\lambda^{1-\alpha_2}).$$

Based on Lemma 7, we have the following optimality gap quantification.

Theorem 8. *For $\alpha_1 \geq \alpha_2 > 1/2$ and $\theta > 0$,*

$$\Pi^{\lambda}(\lceil \tilde{n}_1^{\lambda,*} \rceil, \lceil \tilde{n}_2^{\lambda,*} \rceil, \lfloor \tilde{n}_F^{\lambda,*} \rfloor; \tilde{\nu}^{\lambda}) = \Pi^{\lambda,*} + O(\lambda^{1-\alpha_2}).$$

Theorem 8 indicates that the staffing rule based on the stochastic-fluid approximation together with the scheduling policy $\tilde{\nu}^{\lambda}$ is asymptotically optimal, i.e., it achieves an $o(\sqrt{\lambda})$ optimality gap. In addition, we note from Theorem 8 that the optimality gap of our proposed staffing and scheduling rule is determined by the smaller α_i. This is expected as the size of flexible pool, $\tilde{n}_F^{\lambda,*}$, is determined by the smaller α_i (Lemma 6).

When $\tilde{n}_F^{\lambda,*} > 0$, comparing to the case where no flexible server is available, i.e., $n_F^{\lambda} \equiv 0$, we have

$$\min_{\tilde{n}_1^{\lambda} \geq 0, \tilde{n}_2^{\lambda} \geq 0} \tilde{\Pi}^{\lambda}(\tilde{n}_1^{\lambda}, \tilde{n}_2^{\lambda}, 0) = \tilde{\Pi}^{\lambda,*} + \Theta(\lambda^{\alpha_2}).$$

Then, Theorem 8 indicates that in this case, having access to flexible servers can lead to an $\Theta(\lambda^{\alpha_2})$ cost-saving. This is different from the case without demand uncertainty (Section 3.3), where flexible servers only lead to an $\Theta(\sqrt{\lambda})$ cost-saving.

We conclude this section with some remarks about good scheduling policies when facing a high level of uncertainty in demand. Our proposed scheduling policy $\tilde{\nu}^{\lambda}$ is quite simple but is sufficient for achieving a good performance. This is because for most realized arrival rates, the system is no longer in the critically loaded regime. Thus, any fluid-optimal scheduling policy will achieve a similar optimality gap. To reinforce this point, consider another scheduling policy $\tilde{\nu}_R^{\lambda}$ defined as follows. Similar to $\tilde{\nu}^{\lambda}$, for a realized arrival rate $\Lambda = \gamma$, we allocate $\lfloor \delta(\gamma) n_F^{\lambda} \rfloor$ flexible servers to Class 1 and the remaining $\lceil (1 - \delta(\gamma)) n_F^{\lambda} \rceil$ flexible servers to Class 2. However, unlike $\tilde{\nu}^{\lambda}$, when assigning customers to servers, $\tilde{\nu}_R^{\lambda}$ prioritizes the slower flexible servers over the faster dedicated servers, i.e., the policy turns the M-model into two independent slowest-server-first inverted-V models. Following similar lines of argument as the proof of Theorem 8, one can show that this policy also achieves an $O(\lambda^{1-\alpha_2})$ optimality gap.

Although the scheduling policy $\tilde{\nu}^{\lambda}$ is asymptotically optimal, it can be improved further. We

next introduce a simple improved version of $\tilde{\nu}^\lambda$, which we denote as $\tilde{\nu}_I^\lambda$. For a realized arrival rate, $\Lambda = \gamma$, we again follow the same server allocation rule as $\tilde{\nu}^\lambda$, and when assigning customers to servers, we prioritize the dedicated servers. However, under $\tilde{\nu}_I^\lambda$, the $\lfloor \delta(\gamma) n_F^\lambda \rfloor$ flexible servers 'assigned' to Class 1 only give priority to Class 1 customers, and the remaining $\lceil (1 - \delta(\gamma)) n_F^\lambda \rceil$ flexible servers 'assigned' to Class 2 only give priority to Class 2. For example, when one of the $\lfloor \delta(\gamma) n_F^\lambda \rfloor$ flexible servers assigned to Class 1 becomes available and there is no Class 1 customer waiting, the flexible server can then serve a Class 2 customer waiting in queue. It is easy to see that $\tilde{\nu}_I^\lambda$ is also asymptotically optimal. Indeed, following similar coupling arguments as those in Appendix B.2, one can show that $\tilde{\nu}_I^\lambda$ leads to a smaller steady-state average queue length than $\tilde{\nu}^\lambda$ for any given arrival rate realization.

In Section 3.5.2, we conduct some numerical experiments demonstrating the pre-limit performance of $\tilde{\nu}^\lambda$, $\tilde{\nu}_R^\lambda$, and $\tilde{\nu}_I^\lambda$ introduced above (see Table 3.6).

3.5 Numerical Experiments

In this section, we demonstrate the pre-limit performance of our proposed staffing and scheduling rules using simulation experiments.

3.5.1 Deterministic Arrival Rates

Based on the result in Theorem 7, we set the staffing levels

$$(\hat{n}^\lambda, \hat{n}_F^\lambda) = (\lceil R^\lambda + \beta^* \sqrt{R^\lambda} \rceil, \lfloor \beta_F^* \sqrt{R^\lambda} \rfloor). \tag{3.11}$$

In the first numerical experiment, we consider the case with no abandonment, i.e., $\theta = 0$. We set $h = c = 1$, $\mu = 1, \mu_F = 0.85$, and vary the values of λ and c_F. In Table 3.3, we compare the staffing rule (3.11) to the optimal staffing levels $(n^{\lambda,*}, n_F^{\lambda,*})$ (solved by exhaustive search using simulation). We observe that the staffing levels suggested by the diffusion optimization problem is almost identical to the optimal staffing levels. In most cases, the difference between the two is less

than or equal to 1, and the largest difference is 3. Table 3.3 also reports $\Pi^{\lambda,*}$ and the optimality gaps, i.e., $\Pi^{\lambda}(\hat{n}^{\lambda}, \hat{n}_F^{\lambda}) - \Pi^{\lambda,*}$. As expected, the optimality gaps are extremely small, even for systems as small as $\lambda = 25$.

c_F	$(\hat{n}^{\lambda}, \hat{n}_F^{\lambda})$	$(n^{\lambda,*}, n_F^{\lambda,*})$	$\Pi^{\lambda,*}$	Gap
	$\lambda = 25$			
1	(27,10)	(26,11)	65.91	0.17
1.2	(28,7)	(28,7)	67.76	0
1.4	(29,5)	(30,4)	69.12	0.05
	$\lambda = 100$			
1	(103,20)	(102,22)	230.94	0.08
1.2	(106,15)	(106,15)	234.79	0
1.4	(108,11)	(108,10)	237.27	0.19
	$\lambda = 400$			
1	(406,40)	(405,42)	861.42	0.16
1.2	(412,30)	(413,27)	868.71	0.25
1.4	(416,22)	(416,21)	873.85	0.01

Table 3.3: Performance of $(\hat{n}^{\lambda}, \hat{n}_F^{\lambda})$ for systems with different scales, λ's. ($\mu = 1, \mu_F = 0.85, \theta = 0, h = 8, c = 1$)

Table 3.4 reports the results of a similar experiment when there is abandonment. In this example, we set $h = a = 8$, $c = 1$, $\mu = 1$, and $\mu_F = \theta = 0.85$. We observe again that the prescription (3.11) works very well for all system sizes. Specifically, the optimality gap across all cases are less than 0.1.

3.5.2 Random Arrival Rates

In this section, we study the pre-limit performance of the stochastic-fluid based staffing and scheduling rules when the arrival rates are random. For simplicity of illustration, we consider a symmetric system where $p_1 = p_2 = 1$ and $\alpha_1 = \alpha_2 = \alpha$. In this case, $\tilde{n}_1^{\lambda,*} = \tilde{n}_2^{\lambda,*} := \tilde{n}^{\lambda,*}$. Based on the result in Theorem 8, we set the staffing level

$$(\hat{n}^{\lambda}, \hat{n}_F^{\lambda}) = (\lceil \tilde{n}^{\lambda,*} \rceil, \lfloor \tilde{n}_F^{\lambda,*} \rfloor), \tag{3.12}$$

where $\hat{n}_1^{\lambda} = \hat{n}_2^{\lambda} := \hat{n}^{\lambda}$.

c_F	$(\hat{n}^\lambda, \hat{n}_F^\lambda)$	$(n^{\lambda,*}, n_F^{\lambda,*})$	$\Pi^{\lambda,*}$	Gap
	$\lambda = 25$			
1	(26,11)	(25,13)	65.95	0.02
1.2	(28,7)	(28,7)	67.94	0
1.4	(29,5)	(29,5)	69.26	0
	$\lambda = 100$			
1	(101,23)	(101,23)	231.29	0
1.2	(105,15)	(105,15)	235.12	0
1.4	(107,11)	(108,10)	237.72	0.03
	$\lambda = 400$			
1	(402,46)	(402,46)	862.01	0
1.2	(410,30)	(410,30)	869.50	0
1.4	(414,22)	(415,21)	874.60	0.05

Table 3.4: Performance of $(\hat{n}^\lambda, \hat{n}_F^\lambda)$ for systems with different scales, λ's. ($\mu = 1, \mu_F = 0.85 = \theta, c = 1, h = a = 8$)

Let $c = 1, c_F = 1.2, h = a = 8, \mu = 1, \mu_F = 0.9$, and $\theta = 0.5$. In addition, let $Y_1 = Z_1, Y_2 = \rho Z_1 + \sqrt{1-\rho^2} Z_2$, where Z_1 and Z_2 are independent standard Normal random variables, and $\rho = 0.5$. In this case, $r_1 = r_2 = 1.22$ and $r_F = 0.70$ in Lemma 5. Thus,

$$\hat{n}^\lambda = \lceil \lambda + 1.22\lambda^\alpha \rceil \text{ and } \hat{n}_F^\lambda = \lfloor 0.70\lambda^\alpha / 0.9 \rfloor.$$

Next, to define $\tilde{\nu}^\lambda$, we need to specify $\delta(\gamma)$ (the results of Section 3.4 are valid for any choice that satisfies (3.10)). In our experiments we choose $\delta(\gamma)$ to strike a balance between the capacity of the two classes, i.e. $\gamma_1 - \hat{n}_1^\lambda \mu - \delta \hat{n}_F^\lambda \mu_F$ versus $\gamma_2 - \hat{n}_2^\lambda \mu - (1-\delta)\hat{n}_F^\lambda \mu_F$. For example, if $\hat{n}^\lambda = 34$, $\hat{n}_F^\lambda = 5$ and $\gamma = (33, 35)$, $\tilde{\nu}^\lambda$ allocates 1 flexible server to Class 1 and 4 to Class 2.

Table 3.5 reports the performance of our proposed staffing and scheduling rules for different values of α and λ. The optimality gap in Theorem 8 cannot be computed numerically, because the optimal scheduling policy for (3.2) is unknown. However, because by Lemma 7, the optimality gap satisfies

$$\Pi^\lambda(\hat{n}^\lambda, \hat{n}_F^\lambda; \tilde{\nu}^\lambda) - \Pi^{\lambda,*} \leq \Pi^\lambda(\hat{n}^\lambda, \hat{n}_F^\lambda; \tilde{\nu}^\lambda) - \tilde{\Pi}^{\lambda,*},$$

where $\tilde{\Pi}^{\lambda,*}$ is the optimal value of (3.9), we use $\Pi^\lambda(\hat{n}^\lambda, \hat{n}_F^\lambda; \tilde{\nu}^\lambda) - \tilde{\Pi}^{\lambda,*}$ as an approximation of the optimality gap. Note that this approximation is larger than the actual optimality gap. We refer to it

as “AG" in Table 3.5.

We observe that for a fixed value of λ, the gap decreases as α increases. For example, when $\lambda = 25$, as α increases from 0.6 to 1, the gap decreases from 9.6 to 4.4. This agrees with the results in Theorem 8, i.e., the optimality gap is $O(\lambda^{1-\alpha})$, which decreases as α increases. For a fixed value of α, the ratio between the gap and $\tilde{\Pi}^{\lambda,*}$ decreases as λ increase. For example, when $\alpha = 0.8$, as λ increases from 25 to 100, the gap decreases from 5.5% of $\tilde{\Pi}^{\lambda,*}$ to 2% of $\tilde{\Pi}^{\lambda,*}$.

	$\lambda = 25$		$\lambda = 50$		$\lambda = 100$		$\lambda = 200$	
α	$\tilde{\Pi}^{\lambda,*}$	AG	$\tilde{\Pi}^{\lambda,*}$	AG	$\tilde{\Pi}^{\lambda,*}$	AG	$\tilde{\Pi}^{\lambda,*}$	AG
0.6	78.4	9.6	143.0	12.2	265.2	14.7	498.9	18.6
0.8	104.1	5.7	194.1	6.7	363.9	7.2	685.3	7.9
1	152.9	4.4	305.8	4.6	611.7	4.5	1223.3	4.2

Table 3.5: Performance of $(\hat{n}^\lambda, \hat{n}_F^\lambda; \tilde{\nu}^\lambda)$ for systems with different values of λ and α. ‘AG’ stands for approximate gap. ($c = 1, c_F = 1.2, h = a = 8, \mu = 1, \mu_F = 0.9, \theta = 0.5, \rho = 0.5$)

We next compare the pre-limit performance of three asymptotically optimal scheduling policies: $\tilde{\nu}^\lambda$, $\tilde{\nu}_R^\lambda$, and $\tilde{\nu}_I^\lambda$ introduced in Section 3.4.2. We observe in Table 3.6 that

$$\Pi^\lambda(\hat{n}^\lambda, \hat{n}_F^\lambda; \tilde{\nu}_I^\lambda) < \Pi^\lambda(\hat{n}^\lambda, \hat{n}_F^\lambda; \tilde{\nu}^\lambda) < \Pi^\lambda(\hat{n}^\lambda, \hat{n}_F^\lambda; \tilde{\nu}_R^\lambda).$$

The performance gaps between $\tilde{\nu}_R^\lambda$ and $\tilde{\nu}_I^\lambda$ are small in all cases. This demonstrates that using as crude a policy as $\tilde{\nu}_R^\lambda$ still leads to good performances.

	$\lambda = 25$			$\lambda = 50$		
α	$\tilde{\nu}_I^\lambda$	$\tilde{\nu}^\lambda$	$\tilde{\nu}_R^\lambda$	$\tilde{\nu}_I^\lambda$	$\tilde{\nu}^\lambda$	$\tilde{\nu}_R^\lambda$
0.6	86.3	88.0	88.3	153.0	155.2	155.5
0.8	108.3	109.8	110.0	199.5	200.8	201.0
1	156.2	157.3	157.5	309.6	310.4	310.6
	$\lambda = 100$			$\lambda = 200$		
α	$\tilde{\nu}_I^\lambda$	$\tilde{\nu}^\lambda$	$\tilde{\nu}_R^\lambda$	$\tilde{\nu}_I^\lambda$	$\tilde{\nu}^\lambda$	$\tilde{\nu}_R^\lambda$
0.6	276.8	279.9	280.3	513.6	517.5	518.1
0.8	369.2	371.1	371.4	691.2	693.2	693.5
1	614.3	616.2	616.4	1226.6	1227.5	1227.7

Table 3.6: The cost under other scheduling policies $\nu \in \{\tilde{\nu}_I^\lambda, \tilde{\nu}^\lambda, \tilde{\nu}_R^\lambda\}$ for different values of λ and α. ($c = 1, c_F = 1.2, h = a = 8, \mu = 1, \mu_F = 0.9, \theta = 0.5, \rho = 0.5$)

3.6 Concluding Remarks

In this work, we study the joint optimal staffing and scheduling problem for a two-class queue with both dedicated and flexible servers. We quantify how the cost of flexibility affects the optimal size of the flexible pool. We conclude the chapter with some remarks for future research.

Non-preemption For the deterministic arrival rate setting, our scheduling policy $\nu^{\lambda,*}$ is preemptive. For example, we allow a customer in service with the flexible pool to be transferred to the dedicated pool if a dedicated server becomes available. If we restrict ourselves to non-preemptive policies, one may be tempted to define a non-preemptive version of the policy, and prove that it performs asymptotically as well as the preemptive version in the many-server heavy-traffic regime. Unfortunately, this asymptotic result is unlikely to hold in our case. This is because the size of the flexible pool, $O(\sqrt{\lambda})$, is not large enough to cause instantaneous changes in X^λ in the limit, which indicates that the non-preemptive version of the policy may not be able to closely 'track' the preemptive policy (see [4] for a similar argument).

For the random arrival rate setting, our scheduling policy $\tilde{\nu}^\lambda$ is preemptive, but this is not needed to achieve the optimality gap in Theorem 8. Indeed, a simple coupling argument can show that a non-preemptive fastest-server-first scheduling policy outperforms the preemptive slowest-server-first scheduling policy $\tilde{\nu}^\lambda_R$, and the latter is still asymptotically optimal.

Multiple customer classes When there are k customer classes, servers can potentially have 2^k-1 different skill sets, i.e., each of the non-empty subsets of $\{1,\cdots,k\}$. In this case, we need to specify the optimal size of each potential server pool as well as the corresponding scheduling policy. As k increases, the number of possible system configurations can become very large, posing substantial analytical challenges.

When facing demand uncertainty, we can still approximate the optimal staffing problem with a multi-product inventory management problem with demand substitution. [79] study such inventory networks when the 'staffing' costs are affine (or convex) in the degree of flexibility. Let

$S_1, S_2 \subseteq \{1, \cdots, k\}$, and let n_{S_i} denote the number of servers with skill set S_i. We also write $|S_i|$ for the cardinality of set S_i. [79] finds that if it is optimal to set $n_{S_1}, n_{S_2} > 0$ with $S_1 \subsetneq S_2$, then $|S_1| = |S_2| - 1$. This implies that if the optimal sizes of the dedicated pools are all positive, i.e., $n_{\{i\}} > 0$ for $i = 1, 2, \ldots, k$, then the only other server pools we need to consider are those with skill set $\{i, j\}, i \neq j, i, j = 1, \ldots, k$. This can help reduce the number of possible system configurations that one needs to consider.

References

[1] N. Gans, G. Koole, and A. Mandelbaum, "Telephone call centers: Tutorial, review, and research prospects," *Manufacturing & Service Operations Management*, vol. 5, no. 2, pp. 79–141, 2003.

[2] J. Chen, J. Dong, and P. Shi, "A survey on skill-based routing with applications to service operations management," *Queueing Systems*, vol. 96, pp. 53–82, Oct. 2020.

[3] H. Song, A. L. Tucker, R. Graue, S. Moravick, and J. J. Yang, "Capacity Pooling in Hospitals: The Hidden Consequences of Off-Service Placement," *Management Science*, vol. 66, no. 9, pp. 3825–3842, Aug. 2019.

[4] R. Atar, "Scheduling control for queueing systems with many servers: Asymptotic optimality in heavy traffic," *The Annals of Applied Probability*, vol. 15, no. 4, pp. 2606–2650, Nov. 2005.

[5] J. G. Dai and T. Tezcan, "State Space Collapse in Many-Server Diffusion Limits of Parallel Server Systems," *Mathematics of Operations Research*, vol. 36, no. 2, pp. 271–320, May 2011.

[6] I. Gurvich and W. Whitt, "Queue-and-Idleness-Ratio Controls in Many-Server Service Systems," *Mathematics of Operations Research*, vol. 34, no. 2, pp. 363–396, Apr. 2009.

[7] A. Bassamboo, R. S. Randhawa, and A. Zeevi, "Capacity Sizing Under Parameter Uncertainty: Safety Staffing Principles Revisited," *Management Science*, vol. 56, no. 10, pp. 1668–1686, Oct. 2010.

[8] S. Borst, A. Mandelbaum, and M. I. Reiman, "Dimensioning Large Call Centers," *Operations Research*, vol. 52, no. 1, pp. 17–34, Feb. 2004.

[9] T. J. Best, B. Sandıkçı, D. D. Eisenstein, and D. O. Meltzer, "Managing hospital inpatient bed capacity through partitioning care into focused wings," *Manufacturing & Service Operations Management*, vol. 17, no. 2, pp. 157–176, 2015.

[10] M. Armony and A. R. Ward, "Fair Dynamic Routing in Large-Scale Heterogeneous-Server Systems," *Operations Research*, vol. 58, no. 3, pp. 624–637, Feb. 2010.

[11] Z. Aksin, M. Armony, and V. Mehrotra, "The modern call center: A multi-disciplinary perspective on operations management research," *Production and Operations Management*, vol. 16, no. 6, pp. 665–688, 2007.

[12] J. Shu, M. C. Chou, Q. Liu, C.-P. Teo, and I.-L. Wang, "Models for effective deployment and redistribution of bicycles within public bicycle-sharing systems," *Operations Research*, vol. 61, no. 6, pp. 1346–1359, 2013.

[13] H. Song, A. Tucker, and K. Murrell, "The diseconomies of queue pooling: An empirical investigation of emergency department length of stay," *Management Science*, vol. 61, no. 12, pp. 3032–3053, 2015.

[14] Y. Liu and W. Whitt, "A network of time-varying many-server fluid queues with customer abandonment," *Operations research*, vol. 59, no. 4, pp. 835–846, 2011.

[15] A. Bassamboo and A. Zeevi, "On a data-driven method for staffing large call centers," *Operations Research*, vol. 57, no. 3, pp. 714–726, 2009.

[16] R. Ibrahim and P. L'Ecuyer, "Forecasting call center arrivals: Fixed-effects, mixed-effects, and bivariate models," *Manufacturing & Service Operations Management*, vol. 15, no. 1, pp. 72–85, 2013.

[17] S. Baas *et al.*, "Real-time forecasting of covid-19 bed occupancy in wards and intensive care units," *Health Care Management Science*, vol. 24, no. 2, pp. 402–419, 2021.

[18] D. Smith and W. Whitt, "Resource sharing for efficiency in traffic systems," *Bell Systems Technical Journal*, vol. 60, no. 1, pp. 39–55, 1981.

[19] J. A. van Mieghem, "Investment strategies for flexible resources," *Management Science*, vol. 44, no. 8, pp. 1071–1078, 1998.

[20] S. Graves and B. Tomlin, "Process flexibility in supply chains," *Management Science*, vol. 49, no. 7, pp. 907–919, 2003.

[21] D. Simchi-Levi and Y. Wei, "Understanding the Performance of the Long Chain and Sparse Designs in Process Flexibility," *Operations Research*, vol. 60, no. 5, pp. 1125–1141, Sep. 2012.

[22] O. Z. Akşin and F. Karaesmen, "Characterizing the performance of process flexibility structures," *Operations Research Letters*, vol. 35, no. 4, pp. 477–484, 2007.

[23] A. Bassamboo, R. S. Randhawa, and J. A. V. Mieghem, "A little flexibility is all you need: On the asymptotic value of flexible capacity in parallel queuing systems," *Operations Research*, vol. 60, no. 6, pp. 1423–1435, 2012.

[24] J. Tsitsiklis and K Xu, "On the power of (even a little) resource pooling," *Stochastic Systems*, vol. 2, no. 1, pp. 1–66, 2012.

[25] A. Mandelbaum and M. Reiman, "On pooling in queueing networks," *Management Science*, vol. 44, no. 7, pp. 971–981, 1998.

[26] B. Ata and J. Van Mieghem, "The value of partial resource pooling: Should a service network be integrated or product-focused?" *Management Science*, vol. 55, no. 1, pp. 115–131, 2009.

[27] E. J. Pinker and R. A. Shumsky, "The efficiency-quality trade-off of cross-trained workers," *Manufacturing & Service Operations Management*, vol. 2, no. 1, pp. 32–48, 2000.

[28] O. Garnett and A. Mandelbaum, "An introduction to skills-based routing and its operational complexities," Teaching notes, 2000.

[29] J. A. van Mieghem, "Dynamic Scheduling with Convex Delay Costs: The Generalized $c\mu$ Rule," *The Annals of Applied Probability*, vol. 5, no. 3, pp. 809–833, Aug. 1995.

[30] S. L. Bell and R. J. Williams, "Dynamic scheduling of a system with two parallel servers in heavy traffic with resource pooling: Asymptotic optimality of a threshold policy," *The Annals of Applied Probability*, vol. 11, no. 3, pp. 608–649, Aug. 2001.

[31] A. Mandelbaum and A. L. Stolyar, "Scheduling Flexible Servers with Convex Delay Costs: Heavy-Traffic Optimality of the Generalized cμ-Rule," *Operations Research*, vol. 52, no. 6, pp. 836–855, Dec. 2004.

[32] J. G. Dai and W. Lin, "Maximum Pressure Policies in Stochastic Processing Networks," *Operations Research*, vol. 53, no. 2, pp. 197–218, Apr. 2005.

[33] M. Bramson, B. D'Auria, and N. Walton, "Stability and instability of the maxweight policy," *Mathematics of Operations Research*, vol. 46, no. 4, pp. 1611–1638, 2021.

[34] J. G. Dai and W. Lin, "Asymptotic optimality of maximum pressure policies in stochastic processing networks," *The Annals of Applied Probability*, vol. 18, no. 6, pp. 2239–2299, Dec. 2008.

[35] A. L. Stolyar, "MaxWeight scheduling in a generalized switch: State space collapse and workload minimization in heavy traffic," *The Annals of Applied Probability*, vol. 14, no. 1, pp. 1–53, Feb. 2004.

[36] A. Mandelbaum, W. A. Massey, and M. I. Reiman, "Strong approximations for markovian service networks," *Queueing Systems*, vol. 30, no. 1, pp. 149–201, 1998.

[37] C. Maglaras, "Discrete-review policies for scheduling stochastic networks: Trajectory tracking and fluid-scale asymptotic optimality," *Annals of Applied Probability*, vol. 10, no. 3, pp. 897–929, 2000.

[38] N. Bäuerle, "Optimal control of queueing networks: An approach via fluid models," *Advances in Applied Probability*, pp. 313–328, 2002.

[39] S. P. Sethi and G. L. Thompson, *Optimal Control Theory*. Springer, 2000.

[40] D. Grass, J. Caulkins, G. Feichtinger, G. Tragler, and D. Behrens, *Optimal Control of Nonlinear Processes: With Applications in Drugs, Corruption, and Terror*. Springer, 2008, ISBN: 978-3-540-77646-8.

[41] R. F. Hartl, S. P. Sethi, and R. G. Vickson, "A Survey of the Maximum Principles for Optimal Control Problems with State Constraints," *SIAM Review*, vol. 37, no. 2, pp. 181–218, Jun. 1995, Publisher: Society for Industrial and Applied Mathematics.

[42] R. C. Hampshire and W. A. Massey, "Dynamic optimization with applications to dynamic rate queues," in *Risk and Optimization in an Uncertain World*, INFORMS, 2010, pp. 208–247.

[43] Y. Hu, C. W. Chan, and J. Dong, "Optimal scheduling of proactive service with customer deterioration and improvement," working paper, 2019.

[44] B. Ata and X. Peng, "An optimal callback policy for general arrival processes: A pathwise analysis," *Operations Research*, vol. 68, no. 2, pp. 1–21, 2020.

[45] J. Chang, H. Ayhan, J. Dai, and C. H. Xia, "Dynamic scheduling of a multiclass fluid model with transient overload," *Queueing Systems*, vol. 48, no. 3, pp. 263–307, 2004.

[46] K. Xu and C. W. Chan, "Using future information to reduce waiting times in the emergency department via diversion," *Manufacturing & Service Operations Management*, vol. 18, no. 3, pp. 314–331, 2016.

[47] K. Delana, N. Savva, and T. Tezcan, "Proactive customer service: Operational benefits and economic frictions," *Manufacturing & Service Operations Management*, vol. 23, no. 1, pp. 70–87, 2021.

[48] I. Gurvich, J. Luedtke, and T. Tezcan, "Staffing call centers with uncertain demand forecasts: A chance-constrained optimization approach," *Management Science*, vol. 56, no. 7, pp. 1093–1115, 2010.

[49] Y. Hu, C. W. Chan, and J. Dong, "Prediction-driven surge planning with application in the emergency department," working paper, 2021.

[50] Institute for Health Metrics and Evaluation, *Covid-19 projections*, `https://covid19.healthdata.org/united-states-of-america`, Accessed: 2021-05-27, 2021.

[51] N. Bäuerle, "Asymptotic optimality of tracking policies in stochastic networks," *Annals of Applied Probability*, vol. 10, no. 4, pp. 1065–1083, Nov. 2000.

[52] S. Maman, A. Mandelbaum, and S Zeltyn, "Uncertainty in the demand for service: The case of call centers and emergency departments," Ph.D. dissertation, Technion-Israel Institute of Technology, Faculty of Industrial and Management, 2009.

[53] P. Shi, M. C. Chou, J. G. Dai, D. Ding, and J. Sim, "Models and insights for hospital inpatient operations: Time-dependent ed boarding time," *Management Science*, vol. 62, no. 1, pp. 1–28, 2016.

[54] H. Song, A. Tucker, R. Graue, S. Moravick, and J. Yang, "Capacity pooling in hospitals: The hidden consequences of off-service placement," *Management Science*, vol. 66, no. 9, pp. 3825–3842, 2019.

[55] J. Dong, P. Shi, F. Zheng, and X. Jin, "Off-service placement in inpatient ward network: Resource pooling versus service slowdown," Working paper, 2019.

[56] E. Tekin, W. J. Hopp, and M. P. Van Oyen, "Benefits of skill chaining in production lines with cross-trained workers," *Manufacturing & Service Operations Management*, vol. 4, no. 1, pp. 17–20, 2002.

[57] S. Netessine, G. Dobson, and R. A. Shumsky, "Flexible service capacity: Optimal investment and the impact of demand correlation," *Operations Research*, vol. 50, no. 2, pp. 375–388, 2002.

[58] S. Andradóttir, H. Ayhan, and D. G. Down, "Dynamic server allocation for queueing networks with flexible servers," *Operations Research*, vol. 51, no. 6, pp. 952–968, 2003.

[59] O. Garnett, A. Mandelbaum, and M. Reiman, "Designing a Call Center with Impatient Customers," *Manufacturing & Service Operations Management*, vol. 4, no. 3, pp. 208–227, Jul. 2002.

[60] M. Armony and A. Mandelbaum, "Routing and Staffing in Large-Scale Service Systems: The Case of Homogeneous Impatient Customers and Heterogeneous Servers," *Operations Research*, vol. 59, no. 1, pp. 50–65, Feb. 2011.

[61] I. Gurvich and W. Whitt, "Service-Level Differentiation in Many-Server Service Systems via Queue-Ratio Routing," *Operations Research*, vol. 58, no. 2, pp. 316–328, Oct. 2009.

[62] S. Halfin and W. Whitt, "Heavy-Traffic Limits for Queues with Many Exponential Servers," *Operations Research*, vol. 29, no. 3, pp. 567–588, Jun. 1981.

[63] A. Mandelbaum and S. Zeltyn, "Staffing Many-Server Queues with Impatient Customers: Constraint Satisfaction in Call Centers," *Operations Research*, vol. 57, no. 5, pp. 1189–1205, Jun. 2009.

[64] T. Tezcan and J. Dai, "Dynamic control of n-systems with many servers: Asymptotic optimality of a static priority policy in heavy traffic," *Operations Research*, vol. 58, no. 1, pp. 94–110, 2010.

[65] J. Kim, R. S. Randhawa, and A. R. Ward, "Dynamic scheduling in a many-server, multiclass system: The role of customer impatience in large systems," *Manufacturing & Service Operations Management*, vol. 20, no. 2, pp. 285–301, 2018.

[66] J. M. Harrison and A. Zeevi, "A Method for Staffing Large Call Centers Based on Stochastic Fluid Models," *Manufacturing & Service Operations Management*, vol. 7, no. 1, pp. 20–36, Jan. 2005.

[67] A. Bassamboo, J. M. Harrison, and A. Zeevi, "Design and Control of a Large Call Center: Asymptotic Analysis of an LP-Based Method," *Operations Research*, vol. 54, no. 3, pp. 419–435, Jun. 2006.

[68] W. Whitt, "Staffing a Call Center with Uncertain Arrival Rate and Absenteeism," *Production and Operations Management*, vol. 15, no. 1, pp. 88–102, 2006.

[69] I. Gurvich, J. Luedtke, and T. Tezcan, "Staffing Call Centers with Uncertain Demand Forecasts: A Chance-Constrained Optimization Approach," *Management Science*, vol. 56, no. 7, pp. 1093–1115, May 2010.

[70] Y. L. Koçağa, M. Armony, and A. R. Ward, "Staffing Call Centers with Uncertain Arrival Rates and Co-sourcing," *Production and Operations Management*, vol. 24, no. 7, pp. 1101–1117, 2015.

[71] D. Bertsimas and X. V. Doan, "Robust and data-driven approaches to call centers," *European Journal of Operational Research*, vol. 207, no. 2, pp. 1072–1085, Dec. 2010.

[72] C. Shi, Y. Wei, and Y. Zhong, "Process flexibility for multiperiod production systems," *Operations Research*, vol. 67, no. 5, pp. 1300–1320, 2019.

[73] R. B. Wallace and W. Whitt, "A Staffing Algorithm for Call Centers with Skill-Based Routing," *Manufacturing & Service Operations Management*, vol. 7, no. 4, pp. 276–294, Oct. 2005.

[74] J. M. Harrison and A. Zeevi, "Dynamic Scheduling of a Multiclass Queue in the Halfin-Whitt Heavy Traffic Regime," *Operations Research*, vol. 52, no. 2, pp. 243–257, Apr. 2004.

[75] S. Netessine and N. Rudi, "Centralized and Competitive Inventory Models with Demand Substitution," *Operations Research*, vol. 51, no. 2, pp. 329–335, Apr. 2003.

[76] R. Ernst and P. Kouvelis, "The Effects of Selling Packaged Goods on Inventory Decisions," *Management Science*, vol. 45, no. 8, pp. 1142–1155, Aug. 1999.

[77] K. Rajaram and C. S. Tang, "The impact of product substitution on retail merchandising," *European Journal of Operational Research*, vol. 135, no. 3, pp. 582–601, Dec. 2001.

[78] J. A. Van Mieghem and N. Rudi, "Newsvendor Networks: Inventory Management and Capacity Investment with Discretionary Activities," *Manufacturing & Service Operations Management*, vol. 4, no. 4, pp. 313–335, Oct. 2002.

[79] A. Bassamboo, R. S. Randhawa, and J. A. Van Mieghem, "Optimal Flexibility Configurations in Newsvendor Networks: Going Beyond Chaining and Pairing," *Management Science*, vol. 56, no. 8, pp. 1285–1303, Jun. 2010.

[80] J. G. Dai and J. M. Harrison, *Processing Networks: Fluid Models and Stability*. Cambridge University Press, 2020.

[81] J. Dong, P. Feldman, and G. B. Yom-Tov, "Service Systems with Slowdowns: Potential Failures and Proposed Solutions," *Operations Research*, vol. 63, no. 2, pp. 305–324, Feb. 2015.

[82] D. Gamarnik and A. Zeevi, "Validity of heavy traffic steady-state approximations in generalized Jackson networks," *The Annals of Applied Probability*, vol. 16, no. 1, pp. 56–90, Feb. 2006.

[83] W. Whitt, *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Their Application to Queues* (Springer Series in Operations Research and Financial Engineering). New York: Springer-Verlag, 2002, ISBN: 978-0-387-95358-8.

[84] H. Chen and H. Zhang, "Diffusion Approximations for Some Multiclass Queueing Networks with FIFO Service Disciplines," *Mathematics of Operations Research*, vol. 25, no. 4, pp. 679–707, Nov. 2000.

[85] S. R. E. Turner, "A join the shorter queue model in heavy traffic," *Journal of Applied Probability*, vol. 37, no. 1, pp. 212–223, Mar. 2000.

[86] I. Karatzas and S. Shreve, *Brownian Motion and Stochastic Calculus* (Graduate Texts in Mathematics), 2nd ed. New York: Springer-Verlag, 1998, ISBN: 978-0-387-97655-6.

Appendix A: Additions to and Proofs of Results in Chapter 2

A.1 Full Characterization of Optimal Policies for Extended N-Models

In this section, we provide the full characterization of the optimal scheduling polices for the two extended N-models discussed in Section 2.5. We make the following assumptions about the arrival rate functions.

Assumption 5. *The arrival rate functions $\lambda_i(t), i = 1, \cdots, I$ satisfy:*

1. $\lambda_i(t) \geq s_i\mu_{ii}$ *when* $t < \kappa_i$ *and* $\lambda_i(t) < s_i\mu_{ii}$ *when* $t \geq \kappa_i$.

2. $(\lambda_i(t))_{0\leq t\leq \kappa_i}$ *is piecewise monotone with a finite number of pieces.*

3. $\int_{\kappa_i}^{\infty}(s_i\mu_{ii} - \lambda_i(t))dt = \infty$.

4. *Given* $X(0) = x$, *for any* $0 \leq t \leq \bar{\kappa}$, $W(x,t) > 0$, *where*

$$
\begin{aligned}
W(x,t) = \min_{z} \quad & \sum_{i=1}^{I} q_i(t) \\
& s.t.\ \dot{q}_i(s) = \lambda_i(s) - \sum_{j=1}^{I} \mu_{ij} z_{ij}(s), \quad q_i(0) = x_i, \quad i = 1, \cdots, I \\
& \sum_{i=1}^{I} z_{ij}(s) \leq s_j, \quad j = 1, \cdots, I \\
& q_i(s) \geq 0, \ i = 1, \cdots, I \\
& z_{ij}(s) \geq 0, \quad i, j = 1, \cdots, I.
\end{aligned}
$$

A.1.1 Many-Help-One Extended N-Model

In this section, we consider a 3-by-3 model in which pools 2 and 3 can help class 1, while pool 1 can serve only class 1 (see Figure 2.5(c) for a pictorial illustration). We refer to this model as the exN1-model.

Following the development of the N-model, we first compare the $h\mu$ index. Without loss of generality, we consider three possible cases:

I. $h_1\mu_{12} > h_2\mu_{22}$ and $h_1\mu_{13} > h_3\mu_{33}$. In this case, pool j, $j = 2, 3$ gives priority to class 1 when class 1 has a large enough backlog compared to class j.

II. $h_1\mu_{12} < h_2\mu_{22}$ and $h_1\mu_{13} > h_3\mu_{33}$. In this case, pool 3 gives priority to class 1 when class 1 has a large enough backlog compared to class 3. Pool 2 provides partial help to class 1 after the class 2 queue empties and when class 1 has a large enough backlog.

III. $h_1\mu_{12} < h_2\mu_{22}$ and $h_1\mu_{13} < h_3\mu_{33}$. In this case, pool j, $j = 2, 3$, provides partial help to class 1 after class j queue empties and when class 1 has a large enough backlog.

The key difference between the exN1-model and the N-model is that when pool j, $j = 2, 3$ is determining how long it will help class 1, it also needs to take into account the help class 1 can receive from pool k, $k = 2, 3$, $k \neq j$. To make this notion more precise, we introduce the following notation. Define $\bar{G}^t_{\text{exN1},1,j}(q(t))$ as the time it takes pool 1 to empty queue 1 while taking into account the help it can receive from pool j. We first consider the second server pool – i.e., $j = 2$.

For Case I, let $F_2^t(q(t))$ denote the full helping period for pool 2 to class 1:

$$F_2^t(q) = \inf\left\{u \geq 0 : h_1\mu_{12}G_1^{t+u}(\tilde{q}_1(t+u)) \leq h_2\mu_{22}G_2^{t+u}(\tilde{q}_2(t+u)) + \phi_{12}\right\},$$

where, for $\tilde{q}$: $\tilde{q}(t) = q$; for $s \geq t$, pool 1 serves class 1 only; pool 3 serves class 3 only; and pool 2 prioritizes class 1. Then,

$$\bar{G}^t_{\text{exN1},1,2}(q) = F_2^t(q) + G_1^{t+F_2^t(q)}(\tilde{q}_1(t + F_2^t(q))).$$

In Cases II and III, let $P_2^t(q(t))$ denote the partial help period for pool 2 to class 1:

$$P_2^t(q) = \inf\left\{u \geq 0 : h_1\mu_{12}G_1^{t+G_2^t(q_2)+u}(\tilde{q}_1(t+G_2^t(q_2)+u)) \leq \phi_{12}\right\},$$

where, for $\tilde{q}$ with $\tilde{q}(t) = q$, pool 1 serves class 1 only and pool 3 serves class 3 only for all times $s \geq t$, while pool 2 serves queue 2 only for times between t and $t + G_2^t(q_2)$, and provides partial help to class 1 for times $s \geq t + G_2^t(q_2)$. Then,

$$\bar{G}^t_{\text{exN1,1,2}}(q) = G_2^t(q_2) + P_2^t(q) + G_1^{t+G_2^t(q)+P_2^t(q)}(\tilde{q}_1(t+G_2^t(q)+P_2^t(q))).$$

Note that $G_1^{t+G_2^t(q)+P_2^t(q)}(\tilde{q}_1(t+G_2^t(q)+P_2^t(q))) = \frac{\phi_{12}}{h_1\mu_{12}}$ if $P_2^t(q) > 0$.

We next consider the third server pool – i.e., $j = 3$. In Cases I and II, let $F_3^t(q(t))$ denote the full helping period for pool 3 to class 1:

$$F_3^t(q) = \inf\left\{u \geq 0 : h_1\mu_{13}G_1^{t+u}(\tilde{q}_1(t+u)) \leq h_3\mu_{33}G_3^{t+u}(\tilde{q}_3(t+u)) + \phi_{13}\right\},$$

where, for $\tilde{q}$, $\tilde{q}(t) = q$, for $s \geq t$, pool 1 serves class 1 only; pool 2 serves class 2 only; and pool 3 prioritizes class 1. Then,

$$\bar{G}^t_{\text{exN1,1,3}}(q) = F_3^t(q) + G_1^{t+F_3^t(q)}\left(\tilde{q}_1\left(t+F_3^t(q)\right)\right).$$

In Case III, let $P_3^t(q(t))$ denote the partial helping period for pool 3 to class 1:

$$P_3^t(q) = \inf\left\{u \geq 0 : h_1\mu_{13}G_1^{t+G_3^t(q_3)+u}(\tilde{q}_1(t+G_3^t(q_3)+u)) \leq \phi_{13}\right\},$$

where, for $\tilde{q}$: $\tilde{q}(t) = q$; for $s \geq t$, pool 1 serves class 1 only, and pool 2 serves class 2 only; between t and $t + G_3^t(q_3)$, pool 3 serves class 3 only; for $s \geq t + G_3^t(q_3)$, pool 3 provides partial help to

class 1. Then,

$$\bar{G}^t_{\text{exN1,1,3}}(q) = G^t_3(q_3) + P^t_3(q) + G_1^{t+G^t_3(q)+P^t_3(q)}(\tilde{q}_1(t + G^t_3(q) + P^t_3(q))).$$

Similar to before, $G_1^{t+G^t_3(q)+P^t_3(q)}(\tilde{q}_1(t + G^t_3(q) + P^t_3(q))) = \frac{\phi_{31}}{h_3\mu_{31}}$ if $P^t_3(q) > 0$.

Note that when pool j prioritizes class 1, it is possible that $q_1(t) = 0$, in which case, it may no longer be feasible to have $z_{1j}(t) = s_j$. To simplify the analysis, we will make the following assumption, which ensures that $q_1(t) > 0$ when pool 2 or 3 prioritizes class 1.

Assumption 6. *For $t < \kappa_1$, $\lambda_1(t) > s_1\mu_{11} + s_2\mu_{12} + s_3\mu_{13}$.*

The following theorem characterizes the optimal scheduling policy for the exN1-model.

Theorem 9. *For the exN1-model, under Assumptions 5 and 6, the optimal control for* (2.10) *takes the following form. Pool 1 serves as many class 1 customers as possible.*

I. When $h_1\mu_{12} > h_2\mu_{22}$ and $h_1\mu_{13} > h_3\mu_{33}$, for pool 2,

iia. If $\frac{h_2\mu_{22}G^t_2(q_2(t))+\phi_{12}}{h_1\mu_{12}} \geq \frac{h_3\mu_{33}G^t_3(q_3(t))+\phi_{13}}{h_1\mu_{13}}$ and

$$h_1\mu_{12}\bar{G}^t_{exN1,1,3}(q(t)) > h_2\mu_{22}G^t_2(q_2(t)) + \phi_{12}, \tag{A.1}$$

pool 2 gives priority to class 1.

iib. Otherwise, if $\frac{h_2\mu_{22}G^t_2(q_2(t))+\phi_{12}}{h_1\mu_{12}} < \frac{h_3\mu_{33}G^t_3(q_3(t))+\phi_{13}}{h_1\mu_{13}}$ and

$$h_1\mu_{12}G^t_1(q_1(t)) > h_2\mu_{22}G^t_2(q_2(t)) + \phi_{12}, \tag{A.2}$$

pool 2 gives priority to class 1.

iic. Otherwise, pool 2 serves class 2 only.

For pool 3,

iiia. If $\frac{h_3\mu_{33}G_3^t(q_3(t)+\phi_{13}}{h_1\mu_{13}} \geq \frac{h_2\mu_{22}G_2^t(q_2(t)+\phi_{12}}{h_1\mu_{12}}$ *and*

$$h_1\mu_{13}\bar{G}^t_{exN1,1,2}(q(t)) > h_3\mu_{33}G_3^t(q_3(t)) + \phi_{13}, \tag{A.3}$$

pool 3 gives priority to class 1.

iiib. Otherwise, if $\frac{h_3\mu_{33}G_3^t(q_3(t)+\phi_{13}}{h_1\mu_{13}} < \frac{h_2\mu_{22}G_2^t(q_2(t)+\phi_{12}}{h_1\mu_{12}}$ *and*

$$h_1\mu_{13}G_1^t(q_1(t)) > h_3\mu_{33}G_3^t(q_3(t)) + \phi_{13}, \tag{A.4}$$

pool 3 gives priority to class 1.

iiic. Otherwise, pool 3 serves class 3 only.

II. When $h_1\mu_{12} < h_2\mu_{22}$ *and* $h_1\mu_{13} > h_3\mu_{33}$*, pool 2 prioritizes class 2.*

iia. If

$$G_2^t(q_2(t)) = 0 \text{ and } h_1\mu_{12}\bar{G}^t_{exN1,1,3}(q(t)) > \phi_{12}, \tag{A.5}$$

pool 2 provides partial help to class 1.

iib. Otherwise, pool 2 serves class 2 only.

For pool 3,

iiia. If

$$h_1\mu_{13}\bar{G}^t_{exN1,1,2}(q(t)) > h_3\mu_{33}G_3^t(q_3(t)) + \phi_{13}, \tag{A.6}$$

pool 3 prioritizes class 1.

iiib. Otherwise, pool 3 serves class 3 only.

III. When $h_1\mu_{12} < h_2\mu_{22}$ *and* $h_1\mu_{13} < h_3\mu_{33}$*, both pool 2 and pool 3 prioritize their primary classes, respectively. For pool 2,*

iia. If

$$G_2^t(q_2(t)) = 0, \text{ and } h_1\mu_{12}\bar{G}^t_{exN1,1,3}(q(t)) > \phi_{12}, \tag{A.7}$$

pool 2 provides partial help to class 1.

iib. Otherwise, pool 2 serves class 2 only.

For pool 3,

iiia. If

$$G_3^t(q_3(t)) = 0 \text{ and } h_1\mu_{13}\bar{G}^t_{exN1,1,2}(q(t)) > \phi_{13}, \tag{A.8}$$

pool 3 provides partial help to class 1.

iiic. Otherwise, pool 3 serves class 3 only.

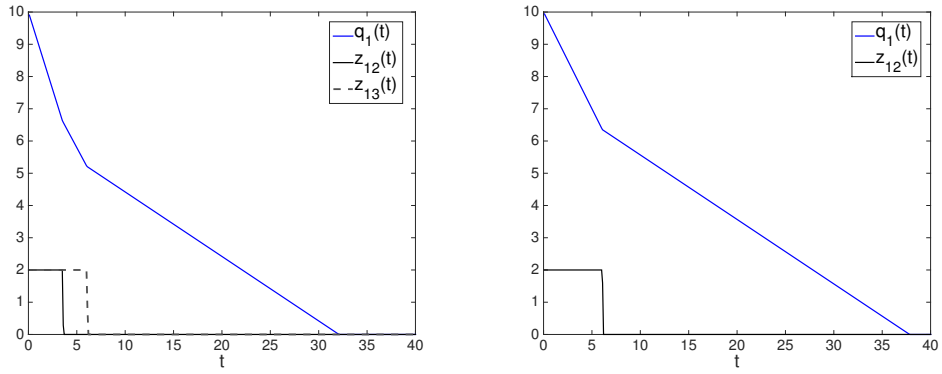
To provide more intuition behind Theorem 9, let us consider the case where the conditions of I.iia. hold. The inequalities of I. involve the $h\mu$ index, under which we show that pool j, $j = 2, 3$, should provide full help to class 1 if help is initiated. This is consistent with the first stage of the optimal policy for the N-model. The first inequality in iia. says that the "tolerance" level of overflow to pool 3 is higher than that of pool 2, noting that the index mimics condition (2.6) if we consider pool 1 and pool 2 (or pool 1 and pool 3) as a sub-system. This second-stage condition indicates that pool 3 will help class 1 longer than pool 2.

It is worth repeating that the key difference between the exN1-model and the N-model lies in the second stage; we need to compare only the $h\mu$ index in the first stage, even when there are more than two classes. Under iia., when pool 2 determines how long it will help class 1, it also needs to take into account the help that class 1 can receive from pool 3, as formalized by (A.1).

Comparing the exN1-model to the N-model, we note that $\bar{G}^t_{\text{exN1},1,3}(q(t)) \leq G_1^t(q_1(t))$. This implies that in the exN1-model, because pool 3 can also help class 1, pool 2 may provide less help to class 1 than in the N-model. Similar observations hold for the other cases, as well. To demonstrate this, Figure A.1 compares the optimal trajectory of an exN1-model (a) with the optimal trajectory of a similar N-model (b). In particular, the two models share the same parameters

for the first two classes. The only difference is that the exN1-model has an extra class, class 3, and an extra server pool, pool 3 (see the caption of Figure A.1 for more details).

For the exN1-model in our example, $h_1\mu_{12} > h_2\mu_{22}$ and $h_1\mu_{13} > h_3\mu_{33}$ – i.e., both pool 2 and pool 3 will give strict priority to class 1 if the class 1 queue is large enough. Indeed, we observe that both pools provide full help to class 1 at the beginning. Pool 2 stops helping class 1 at $t = 3.5$ in the exN1-model. In contrast, pool 2 stops helping class 1 at $t = 6.1$ in the N-model. This is because in the exN1-model, class 1 can also get help from pool 3, and when pool 2 decides how much to help class 1, it also takes the extra help from pool 3 into account. Lastly, we note that with the extra help from pool 3, the exN1-model is able to empty the class 1 queue faster than the N-model can.



(a) exN1-model (b) N-model

Figure A.1: Optimal trajectory of the exN1-model versus the N-model. ($s_1 = s_2 = 2$, $\lambda_1 = \lambda_2 = 0.3$, $\mu_{11} = \mu_{22} = 0.25$, $\mu_{12} = 0.2$, $\phi_{12} = 1$, $h_1 = 1.5$, $h_2 = 1$, $q_1(0) = 10$, $q_2(0) = 5$. For the exN1-model, $s_3 = 2$, $\lambda_3 = 0.3$, $\mu_{33} = 0.25$, $\mu_{13} = 0.18$, $\phi_{13} = 1$, $h_3 = 1$, and $q_3(0) = 3$.)

A.1.2 One-Helps-Many Extended N-Model

In this section, we consider a 3×3 model in which pool 1 can serve classes 1, 2 and 3, while pools 2 and 3 can serve only their corresponding primary class (see Figure 2.5(d) for a pictorial illustration). We refer to this model as the exN2-model.

Following the development of the N-model, we first compare the $h\mu$ index. Without loss of generality, we consider three possible cases:

I. $h_2\mu_{21} > h_3\mu_{31} > h_1\mu_{11}$. In this case, pool 1 prioritizes classes 2 and 3 when there are large enough backlogs in these two classes compared to class 1. When deciding between classes 2 and 3, class 2 enjoys higher priority over class 3.

II. $h_2\mu_{21} > h_1\mu_{11} > h_3\mu_{31}$. In this case, pool 1 prioritizes class 2 when there is a large enough backlog in class 2. Pool 1 can provide partial help to class 3 after the class 1 queue empties and when class 3 has a large enough backlog.

III. $h_1\mu_{11} > h_2\mu_{21} > h_3\mu_{31}$. In this case, pool 1 provides only partial help to classes 2 and 3 after the class 1 queue empties and when there are large enough backlogs in the two classes. When deciding between classes 2 and 3, class 2 enjoys higher priority over class 3.

The key difference between the exN2-model and the N-model is that when pool 1 is determining how long it will help class i, $i = 2, 3$, it also needs to take into account the help it can provide to class k, $k = 2, 3$, $k \neq i$. To make this notion more precise, we introduce the following notation.

In Case I, let $F^t(q(t))$ denote the length of the full helping period for pool 1 to class 3:

$$F^t(q) = \inf\left\{u \geq 0 : h_3\mu_{31}G_3^{t+u}(\tilde{q}_3(t+u)) \leq h_1\mu_{11}G_1^{t+u}(\tilde{q}_1(t+u)) + \phi_{31}\right\},$$

where for $\tilde{q}$: $\tilde{q}(t) = q$; for $s \geq t$, pool 1 prioritizes class 3. Let $\bar{G}^t_{\text{exN2},1}(q(t))$ denote the time it takes to empty queue 1 given that it may provide some help to class 3:

$$\bar{G}^t_{\text{exN2},1}(q) = F^t(q(t)) + G_1^{t+F^t(q)}(\tilde{q}_1(t+F^t(q))).$$

In Cases II and III, let $P^t(q(t))$ denote the length of pool 1's partial helping period to class 3:

$$P^t(q) = \inf\left\{u \geq 0 : h_3\mu_{31}G_3^{t+G_1^t(q_1)+u}(\tilde{q}_3(t+G_1^t(q_1)+u)) \leq \phi_{31}\right\},$$

where, for $\tilde{q}$: $\tilde{q}(t) = q$, between t and $G_1^t(q_1)$, pool 1 serves class 1 only; and for $s > t + G_1^t(q_1)$, pool 1 provides partial help to class 3.

Note that when pool 1 gives priority to class i, it is possible that $q_2(t) = 0$, in which case, it may no longer be feasible to have $z_{21}(t) = s_1$. To simplify the analysis, we will make the following assumption, which ensures that $q_i(t) > 0$ when pool 1 gives priority to class i, $i = 2, 3$.

Assumption 7. *For $i = 1, 2$ and $t < \kappa_i$, $\lambda_i(t) > s_1\mu_{i1} + s_i\mu_{ii}$.*

The following theorem characterizes the optimal scheduling policy for the exN2-model.

Theorem 10. *For the exN2-model, under Assumptions 5 and 7, the optimal control for* (2.10) *takes the following form. Pools 2 and 3 serve their primary classes as much as possible.*

I. When $h_2\mu_{21} > h_3\mu_{31} > h_1\mu_{11}$,

a. If

$$h_2\mu_{21}G_2^t(q_2(t)) > h_1\mu_{11}\bar{G}_{exN2,1}^t(q(t)) + (h_3\mu_{31} - h_1\mu_{11})F^t(q(t)) + \phi_{21}, \tag{A.9}$$

pool 1 gives priority to class 2.

b. Otherwise, if

$$h_3\mu_{31}G_3^t(q_3(t)) > h_1\mu_{11}G_1^t(q_1(t)) + \phi_{31}, \tag{A.10}$$

pool 1 gives priority to class 3.

c. Otherwise, pool 1 serves class 1 only.

II. When $h_2\mu_{21} > h_1\mu_{11} > h_3\mu_{31}$,

a. If

$$h_2\mu_{21}G_2^t(q_2(t)) > h_1\mu_{11}G_1^t(q_1(t)) + h_3\mu_{31}P^t(q(t)) + \phi_{21}, \tag{A.11}$$

pool 1 gives priority to class 2.

b. Otherwise, if $G_1^t(q_1(t)) = 0$ and $h_3\mu_{31}G_3^t(q_3(t)) > \phi_{31}$, pool 1 provides partial help to class 3.

c. Otherwise, pool 1 serves class 1 only.

III. When $h_1\mu_{11} > h_2\mu_{21} > h_3\mu_{31}$,

a. If

$$G_1^t(q_1(t)) = 0 \text{ and } h_2\mu_{21}G_2^t(q_2(t)) > h_3\mu_{31}P^t(q(t)) + \phi_{21}, \tag{A.12}$$

pool 1 provides partial help to class 2.

b. Otherwise, if $G_1^t(q_1(t)) = 0$ *and* $h_3\mu_{31}G_3^t(q_3(t)) > \phi_{31}$, *pool 1 provides partial help to class 3.*

c. Otherwise, pool 1 serves class 1 only.

To provide more intuition behind Theorem 10, let us consider Case I. In the first stage, pool 1 prioritizes classes 2 and 3 when there are large enough backlogs in these two classes compared to class 1. When deciding between classes 2 and 3, class 2 enjoys a higher priority than class 3. The key difference between the exN2-model and the N-model again lies in the second stage. In particular, when pool 1 is determining how long it will help class 2, it also needs to consider the help it can provide to class 3, as formalized by (A.9).

Comparing the exN2-model to the N-model, we note that $\bar{G}^t_{\text{exN2},1}(q(t)) \geq G_1^t(q_1(t))$, and

$$h_1\mu_{11}\bar{G}^t_{\text{exN2},1}(q(t)) + (h_3\mu_{31} - h_1\mu_{11})F^t(q(t)) + \phi_{21} \geq h_1\mu_{11}G_1^t(q_1(t)) + \phi_{21}.$$

Because pool 1 can also help class 3 in the exN2-model, it provides less help to class 2 than in an otherwise similar N-model. See also Figure A.2 in Appendix A.2 for a numerical illustration.

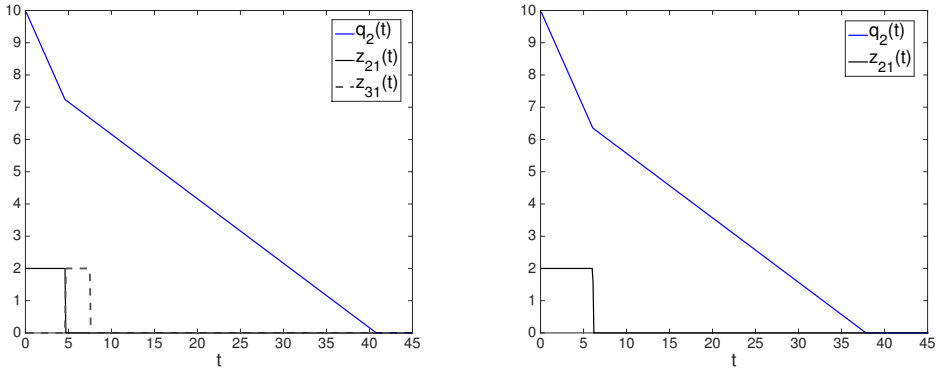
A.2 Additional Numerical Experiments

A.2.1 Fluid trajectory for exN2-model

Figure A.2 compares the optimal trajectory of an exN2-model (a) with the optimal trajectory of a similar N-model (b). For the N-model, we assume that pool 1 can serve both classes 1 and 2,

while pool 2 can serve only class 2. The two systems share the same parameters for the first two classes (see the caption of Figure A.2 for more details).

For the exN2-model in our example, we have $h_2\mu_{21} > h_3\mu_{31} > h_1\mu_{11}$. Thus, we observe that pool 1 first provides full help to class 2 and then switches priority to help class 3. Pool 1 stops helping class 2 at $t = 4.6$. In contrast, in the N-model, pool 1 stops helping class 2 at $t = 6.1$. This is because in the exN2-model, pool 1 can also help class 3 and, thus, may provide less help to class 2 in order to help class 3. In the exN2-model, pool 1 provides full help to class 3 from $t = 4.6$ to $t = 7.6$. Lastly, we note that because class 2 gets more help from pool 1 in the N-model, its queue empties faster in the N-model than in the exN2-model.



(a) exN2-model (b) N-model

Figure A.2: Optimal trajectory of the exN2-model versus the N-model. ($s_1 = s_2 = 2$, $\lambda_1 = \lambda_2 = 0.3$, $\mu_{11} = \mu_{22} = 0.25$, $\mu_{21} = 0.2$, $\phi_{21} = 1$, $h_1 = 1$, $h_2 = 1.5$, $q_1(0) = 5$, $q_2(0) = 10$. For the exN2-model, $s_3 = 2$, $\lambda_3 = 0.3$, $\mu_{33} = 0.25$, $\mu_{31} = 0.18$,$\phi_{13} = 1$, $h_3 = 1$, and $q_3(0) = 10$.)

A.2.2 Performance of different policies in the stochastic systems

Tuned versus untuned policies for the N-model

Table A.1 shows the cost comparison between the tuned and untuned policies for the N-model. We set the tuning parameter $\theta = 0.8$ in the tuned policy, where θ is used for $\theta G^t_{k(i^*)}(q_{k(i^*)}(t))$ on the left-hand side of (2.12) and (2.13) in the heuristic policy. Note that the untuned policy (with $\theta = 1$) is the fluid optimal control policy for the N-model. From this table, we observe that the tuned policy can achieve a slightly better performance than the untuned policy can in the stochastic

system. The relative cost difference between the tuned and untuned policies is 1.1% to 2.1%.

Different policies for the X-model

The X-model setting is the same as the baseline setting specified in Section 2.6.1 for the N-model, except that pool 1 can serve class 2 customers if necessary. Table A.2 reports the cost comparison among different policies. The "Look-ahead (opt)" policy corresponds to the optimal fluid control derived in Section 2.5.1, while the "Look-ahead (heu)" policy is the heuristic policy – i.e., (2.12) and (2.13) – with tuning parameter $\theta = 0.8$. In all cases tested, our heuristic policy achieves performance that is comparable to (slightly better than) that of the fluid-optimal policy.

Table A.3 shows the cost comparison between the tuned and untuned policies for the X-model. In the table, the "N-policy Untuned" and "X-policy Untuned" stand for the directly translated optimal fluid control policies derived for the N- and X-models, and for the "N-policy Tuned", we use the tuning parameter $\theta = 0.8$. From this table, we observe that the relative cost difference between the tuned and untuned policies is 0.6% to 2.1%.

Tuned versus untuned heuristic policies for the 5-by-5 networks

Table A.4 shows the cost comparison between the tuned and the untuned policies for the 5×5 model. We set the tuning parameter $\theta = 0.8$ in the tuned policy. We observe that the relative cost difference between the tuned and untuned policies is 1.2% to 4.1%.

A.3 Proof of Lemma 1

Proof. Proof. Take $G_i^t(q_i(t))$ as a function of t. Note that when class i ($i = 1, 2$) is only served by pool i and $G_i^t(q_i(t)) > 0$, $G_i^t(q_i(t))$ decreases at rate 1 until it hits zero. When pool 2 provides help to class 1, $G_1^t(q_1(t))$ decreases at rate at least 1, while $G_2^t(q_2(t))$ decreases at rate at most 1. Since $h_1\mu_{12} > h_2\mu_{22}$, $\psi(t)$ keeps decreasing in t until $G_1^t(q_1(t))$ hits zero. After $G_1^t(q_1(t))$ hits zero, say at time τ_1, it stays at zero for $t \geq \tau_1$. Since $G_2^t(q_2(t)) \geq 0$, $\psi(t) < 0$ for $t \geq \tau_1$. □

		Tuned ($\theta = 0.8$)	Untuned ($\theta = 1$)
		First Arrival Setting	
$\phi = 2$	Holding	1.08	1.09
	Overflow	0.13	0.14
	Total	1.21	1.23
	SE	0.003	0.003
$\phi = 10$	Holding	1.13	1.10
	Overflow	0.51	0.56
	Total	1.64	1.67
	SE	0.004	0.004
$\phi = 25$	Holding	1.47	1.28
	Overflow	0.78	1.00
	Total	2.25	2.28
	SE	0.005	0.005
		Tuned ($\theta = 0.8$)	Untuned ($\theta = 1$)
		Second Arrival Setting	
$\phi = 2$	Holding	2.72	2.78
	Overflow	0.28	0.29
	Total	3.00	3.07
	SE	0.007	0.007
$\phi = 10$	Holding	2.68	2.70
	Overflow	1.31	1.37
	Total	3.99	4.08
	SE	0.008	0.008
$\phi = 25$	Holding	2.78	2.72
	Overflow	2.81	2.98
	Total	5.59	5.70
	SE	0.010	0.010

Table A.1: Simulation costs for the N-model over 10000 replications. The holding cost $h = (1.5, 1)$. The costs shown in the table are in units of 10^4. "SE" stands for the standard error for the corresponding total cost. The tuning parameter θ is for $\theta G^t_{k(i^*)}(q_{k(i^*)}(t))$ on the left-hand side of (2.12) and (2.13) in the heuristic policy. ($h = (1.5, 1)$, $s_i = 20$ and $\mu_{ii} = 0.25$, for $i = 1, 2$, $\mu_{12} = 0.2$, $\phi_{12} = \phi$, $\lambda_2(t) = 3$. First arrival setting: $\lambda_1(t) = 8 \times \mathbf{1}\{t < 40\} + 4 \times \mathbf{1}\{t \geq 40\}$ and $X(0) = (60, 70)$. Second arrival setting: $\lambda_1(t) = 12 \times \mathbf{1}\{t < 40\} + 4.5 \times \mathbf{1}\{t \geq 40\}$ and $X(0) = (30, 40)$.)

		LA (opt)	LA (heu)	MaxP	ModMaxP	Cmu	ModCmu
$\phi = 2$	Holding	1.09	1.08	1.10	1.10	1.08	2.75
	Overflow	0.14	0.13	0.13	0.13	0.22	0.00
	Total	1.23	1.21	1.23	1.23	1.29	2.75
	SE	0.003	0.003	0.003	0.003	0.003	0.008
$\phi = 10$	Holding	1.10	1.13	1.10	1.11	1.08	2.75
	Overflow	0.56	0.51	0.65	0.62	1.08	0.00
	Total	1.67	1.64	1.75	1.73	2.16	2.75
	SE	0.004	0.004	0.004	0.004	0.005	0.008
$\phi = 25$	Holding	1.28	1.47	1.10	1.13	1.08	2.75
	Overflow	1.00	0.78	1.63	1.41	2.70	0.00
	Total	2.28	2.25	2.73	2.54	3.78	2.75
	SE	0.005	0.005	0.005	0.005	0.007	0.008

Table A.2: Expected total cost for the X-model under different routing policies. The costs shown in the table are in units of 10^4. "SE" stands for the standard error for the corresponding average total cost (holding + overflow). (Parameter setting: $h = (1.5, 1)$, $s_i = 20$, $\mu_{ii} = 0.25$, $\mu_{ij} = 0.2$ and $\phi_{ij} = \phi$ for $i \neq j$; $\lambda_2(t) = 3$, $\lambda_1(t) = 8 \times \mathbf{1}\{t < 40\} + 4 \times \mathbf{1}\{t \geq 40\}$, and $X(0) = (60, 70)$.)

A.4 Proof of Optimal Fluid Control Results

The proof of Theorem 1 and subsequent fluid optimal control results (Theorems 2, 4, 9 and 10) utilize Pontryagin's Minimum Principle. In its most standard version, Pontryagin's Minimum Principle provides a list of necessary conditions satisfied by any optimal solution to the optimal control problem. In this section, we first introduce a special sufficient version of Pontryagin's Minimum Principle. We then demonstrate how it can be applied to prove Theorem 1. The proofs of the other results follow similar lines of analysis and are provided later in this section.

A.4.1 Pontryagin's Minimum Principle

To state the result in a general form that can be applied to all our subsequent analysis, we first introduce some notation.

Consider a system with I classes of customers, i.e., $q = (q_1, \ldots, q_I)$ and $z = (z_{ij}, i, j = 1, \ldots, I)$. Let $F(q, z) = \sum_{i=1}^{I} h_i q_i + \sum_{j \neq i} \phi_{ij} z_{ij}$ denote the instantaneous cost function. Let $\dot{q}_i(t) = f_i(q, z, t)$ and $f(q, z, t) = (f_1(q, z, t), \ldots, f_I(q, z, t))$. We also define $g_i(q) = -q_i$ and $g(q) =$

		N-policy Tuned ($\theta = 0.8$)	N-policy Untuned	X-policy Untuned
		First Arrival Setting		
$\phi = 2$	Holding	1.08	1.09	1.05
	Overflow	0.13	0.14	0.16
	Total	1.21	1.23	1.21
	SE	0.003	0.003	0.003
$\phi = 10$	Holding	1.13	1.10	1.10
	Overflow	0.51	0.56	0.56
	Total	1.64	1.67	1.67
	SE	0.004	0.004	0.004
$\phi = 25$	Holding	1.47	1.28	1.28
	Overflow	0.78	1.00	1.00
	Total	2.25	2.28	2.28
	SE	0.005	0.005	0.005
		N-policy Tuned ($\theta = 0.8$)	N-policy Untuned	X-policy Untuned
		Second Arrival Setting		
$\phi = 2$	Holding	2.63	2.64	2.64
	Overflow	0.30	0.32	0.32
	Total	2.94	2.96	2.96
	SE	0.007	0.007	0.007
$\phi = 10$	Holding	2.67	2.67	2.67
	Overflow	1.31	1.40	1.40
	Total	3.99	4.07	4.07
	SE	0.008	0.008	0.008
$\phi = 25$	Holding	2.78	2.72	2.72
	Overflow	2.81	2.98	2.98
	Total	5.59	5.70	5.70
	SE	0.010	0.010	0.010

Table A.3: Simulation costs for the X-model over 10000 replications. The costs shown in the table are in units of 10^4. "SE" stands for the standard error for the corresponding total cost. In the table, the "N-policy Untuned" and "X-policy Untuned" stand for the optimal fluid control policies derived for the N- and X-models, and for the "N-policy Tuned", we used the tuning parameter $\theta = 0.8$ for $\theta G^t_{k(i^*)}(q_{k(i^*)}(t))$ on the left-hand side of (2.12) and (2.13). ($h = (1.5, 1)$, $s_i = 20$ and $\mu_{ii} = 0.25$, for $i = 1, 2$, $\mu_{12} = 0.2$, $\phi_{12} = \phi$, $\lambda_2(t) = 3$. First arrival setting: $\lambda_1(t) = 8 \times \mathbf{1}\{t < 40\} + 4 \times \mathbf{1}\{t \geq 40\}$ and $X(0) = (60, 70)$. Second arrival setting: $\lambda_1(t) = 12 \times \mathbf{1}\{t < 40\} + 4.5 \times \mathbf{1}\{t \geq 40\}$ and $X(0) = (30, 40)$. For the X-model, $\mu_{21} = 0.2$ and $\phi_{21} = \phi$.)

		Tuned ($\theta = 0.8$)	Untuned ($\theta = 1$)
		Network Structure 1	
$\phi = 2$	Holding	3.21	3.24
	Overflow	0.52	0.53
	Total	3.73	3.77
	SE	0.007	0.007
$\phi = 10$	Holding	3.45	3.33
	Overflow	2.05	2.24
	Total	5.50	5.58
	SE	0.008	0.008
$\phi = 25$	Holding	4.29	4.04
	Overflow	3.57	3.95
	Total	7.86	7.99
	SE	0.012	0.011
		Network Structure 2	
$\phi = 2$	Holding	3.00	3.08
	Overflow	0.48	0.50
	Total	3.48	3.58
	SE	0.005	0.006
$\phi = 10$	Holding	3.07	3.12
	Overflow	1.98	2.15
	Total	5.05	5.26
	SE	0.007	0.008
$\phi = 25$	Holding	3.57	3.42
	Overflow	3.77	4.19
	Total	7.34	7.61
	SE	0.009	0.010

Table A.4: Simulation costs for the 5×5 model over 10000 replications. The holding cost $h = (1.5, 1)$. The costs shown in the table are in units of 10^4. "SE" stands for the standard error for the corresponding total cost. The tuning parameter θ is for $\theta G^t_{k(i^*)}(q_{k(i^*)}(t))$ on the left-hand side of (2.12) and (2.13) in the heuristic policy. (Parameter setting: $h = (1.5, 1, 1, 1.5, 1)$, $s_i = 20$, $\mu_{ii} = 0.25$, $\mu_{ij} = 0.2$ and $\phi_{ij} = \phi$ for $i \neq j$; $\lambda_1(t) = 12 \times \mathbf{1}\{t < 40\} + 4.5 \times \mathbf{1}\{t \geq 40\}$, $\lambda_2(t) = 3$, $\lambda_3(t) = 4$, $\lambda_4(t) = 8 \times \mathbf{1}\{t < 40\} + 4 \times \mathbf{1}\{t \geq 40\}$, $\lambda_5(t) = 3$, and $X(0) = (30, 40, 50, 60, 70)$.)

$(g_1(q), \ldots, g_I(q))$. Lastly, let $l_{ij}(z) = -z_{ij}$, $\tilde{l}_j(z) = \sum_{i=1}^{I} z_{ij} - s_j$, and $l(z) = (l_{ij}(z), \tilde{l}_j(z), i, j = 1, \ldots, I)$. Consider a general optimal control problem

$$\begin{aligned} \min_{z} \quad & \int_0^\infty F\left(q(t), z(t)\right) dt \\ \text{s.t.} \quad & \dot{q}(t) = f\left(q(t), z(t), t\right), \quad q(0) = q_0 \\ & g(q(t)) \leq 0 \\ & l(z(t)) \leq 0 \end{aligned} \tag{A.13}$$

Note that under the assumption that $\int_{\kappa_i}^{\infty} (s_i \mu_{ii} - \lambda_i(t)) dt = \infty$ for $i = 1, \ldots, I$. The queue will eventually hit zero and stay there. After this hitting time, $F\left(q(t), z(t)\right) = 0$. Thus, even though we define (A.13) as an infinite horizon problem, it is the same as a finite horizon problem where the planning horizon is long enough (possibly depending on the initial condition) such that the queue reaches zero by the end of the planning horizon.

Let $p(t) = (p_1(t), \ldots, p_I(t)) \in \mathbb{R}^I$ denote the adjoint vector. Let $\eta(t) = (\eta_1(t), \ldots, \eta_I(t)) \in \mathbb{R}^I$ and $\xi(t) = (\xi_{ij}(t), \tilde{\xi}_j(t), i, j = 1, \ldots, I) \in \mathbb{R}^{I^2+I}$ denote the Lagrangian multipliers for the state and control constraints respectively. Define the Hamiltonian H as

$$H(q(t), z(t), p(t), t) = F(q(t), z(t)) + p(t)^T f\left(q(t), z(t), t\right)$$

and the augmented Hamiltonian L as

$$L(q(t), z(t), p(t), \eta(t), \gamma(t), \xi(t), t) = H(q(t), z(t), p(t), t) + \eta(t)^T g(q(t)) + \xi(t)^T l(z(t))$$

The following sufficient conditions are adapted from Theorems 8.2 and 8.4 in [41] for (A.13).

Theorem 11 (Arrow-type sufficient condition)**.** *Let (q^*, z^*) be a feasible pair for the optimal control problem* (A.13)*. Assume that there exists a piecewise continuously differentiable function $p^*(t) : [0, \infty) \to \mathbb{R}^I$ and piecewise continuous functions $\eta^* : [0, \infty) \to \mathbb{R}^I$ and $\xi^* : [0, \infty) \to \mathbb{R}^{I^2+I}$, such that the following conditions hold almost everywhere:*

1. *Ordinary Differential Equation condition:*

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

2. *Adjoint Vector condition:*

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

3. *Minimization condition:*

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

4. *Transversality condition:*

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

5. *Complementarity condition:*

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

6. *Jump condition: At every point β of discontinuity of p^*, there exists an $\omega^*(\beta) \in \mathbb{R}^I$ such that*

$$\begin{aligned} &p^*(\beta-) = p^*(\beta+) + \omega^*(\beta)^T \nabla_q g(q^*(\beta)) \\ &\omega^*(\beta) \geq 0, \quad \omega^*(\beta)^T g\left(q^*(\beta)\right) = 0. \end{aligned} \tag{J}$$

7. *Hamiltonian condition (H): If the minimized Hamiltonian $H(q^*(t), z^*(t), p^*(t), t)$ 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 $l(z(t))$ is quasiconvex in $z(t)$.*

Then, (q^*, z^*) *is an optimal pair.*

A.4.2 Optimal control for the N-Model under single demand surge

In this section, we provide the proof of Theorem 1. The basic strategy is to construct a feasible pair (q^*, z^*) and verify that the assumptions in Theorem 11 hold.

Proof. Proof of Theorem 1. In this case,

$$\begin{aligned} H(q(t), z(t), p(t), t) =& h_1 q_1(t) + h_2 q_2(t) + \phi_{12} z_{12}(t) \\ &+ p_1(t)\left(\lambda_1(t) - \mu_{11} z_{11}(t) - \mu_{12} z_{12}(t)\right) + p_2(t)\left(\lambda_2(t) - \mu_{22} z_{22}(t)\right) \end{aligned}$$

and

$$\begin{aligned} L(q(t), z(t), p(t), \eta(t), \xi(t), \gamma(t), t) =& H(q(t), z(t), p(t)) - \eta_1(t) q_1(t) - \eta_2(t) q_2(t) \\ &- \xi_{11}(t) z_{11}(t) - \xi_{12}(t) z_{12}(t) - \xi_{22}(t) z_{22}(t) \\ &+ \gamma_1(t)(z_{11}(t) - s_1) + \gamma_2(t)(z_{12}(t) + z_{22}(t) - s_2). \end{aligned}$$

We next verify the sufficient conditions listed in Theorem 11.

Case I: $h_1\mu_{12} \geq h_2\mu_{22}$. In this case, the policy is that pool 2 gives priority to class 1 for a time

$$\tau^* = \inf\{t \geq 0 : h_1\mu_{12} G_1^t(q_1(t)) - \phi_{12} \leq h_2\mu_{22} G_2^t(q_2(t))\} \tag{A.14}$$

assuming the inequality in case (Ia) holds initially, and $\tau^* = 0$ otherwise. After this τ^* units of time, pool 2 stops helping class 1. To see this, note that since $h_1\mu_{12} \geq h_2\mu_{22}$, from Lemma 1, if the inequality in case (Ia) does not hold at some t', it also does not hold at all subsequent $t \geq t'$.

Under the policy characterized in Case I, the times to deplete the two queues are

$$\tau_1^* = \tau^* + G_1^{\tau^*}(q_1^*(\tau^*)), \quad \tau_2^* = \tau^* + G_2^{\tau^*}(q_2^*(\tau^*)). \tag{A.15}$$

Then, we consider the following queue length trajectory:

$$q_1^*(t) = \begin{cases} q_1 + \int_0^t (\lambda_1(s) - s_1\mu_{11} - z_{12}^*(s)\mu_{12})\, ds, & t \in [0, \tau^*), \\ q_1^*(\tau^*) + \int_{\tau_*}^t (\lambda_1(s) - s_1\mu_{11})\, ds, & t \in [\tau^*, \tau_1^*), \\ 0, & t \in [\tau_1^*, \infty), \end{cases}$$

$$q_2^*(t) = \begin{cases} q_2 + \int_0^t (\lambda_2(s) - z_{22}^*(s)\mu_{22})\, ds, & t \in [0, \tau^*), \\ q_2^*(\tau^*) + \int_{\tau^*}^t (\lambda_2(s) - s_2\mu_{22})\, ds, & t \in [\tau^*, \tau_2^*), \\ 0, & t \in [\tau_2^*, \infty). \end{cases}$$

Note that it may be that $z_{12}^*(t) < s_2$ for $t \in [0, \tau^*]$, if $q_1(t) = 0$ and $\lambda_1(t) < s_1\mu_{11} + s_2\mu_{12}$. In this case, $z_{22}^*(t) = s_2 - z_{12}^*(t)$ (since $q_2(t) > 0$ by assumption). However, it is always the case that $z_{11}^*(t) = s_1$ for $t \in [0, \tau^*]$, since either (i) $q_1(t) > 0$ or (ii) $q_1(t) = 0$ and $t < \kappa_1$, so that $\lambda_1(t) \geq s_1\mu_{11}$.

Assuming $\tau^* > 0$, we now partition the interval $[0, \tau^*)$ into subintervals $I_1, \cdots, I_n$ where $n \geq 1$, $I_i = [A_{i-1}, A_i)$ and $0 = A_0 < A_1 < \cdots < A_n = \tau^*$, as follows. In the interior $t \in (A_{i-1}, A_i)$ of each subinterval, either (i) $q_1(t) > 0$ and $q_2(t) > 0$, in which case we say that I_i is an interior subinterval, or (ii) $q_1(t) = 0$ and $q_2(t) > 0$, in which case we say that I_i is a boundary subinterval. Note that it is not possible that $q_1(t) > 0$ and $q_2(t) = 0$ in some subinterval, because $z_{22}^*(t) = 0$ during this time and $\lambda_2(t) > 0$. The subintervals $I_1, \cdots, I_n$ do not necessarily alternate between interior and boundary subintervals: it is possible that I_k and I_{k+1} are both interior subintervals, with $q_1(t)$ hitting zero at the single point A_k.

We next define the adjoint vector

$$p_2^*(t) = \begin{cases} h_2(\tau_2^* - t), & t \in [0, \tau_2^*), \\ 0, & t \in [\tau_2^*, \infty), \end{cases}$$

and

$$p_1^*(t) = \begin{cases} h_1(\tau_1^* - t), & t \in [\tau^*, \tau_1^*), \\ 0, & t \in [\tau_1^*, \infty). \end{cases}$$

For $t < \tau^*$, with $p_1^*(A_n) = p_1^*(\tau^*)$ defined, moving backwards in time, we recursively define $p_1^*(t)$ for $t \in [0, A_n)$. We will do this in such a way that (i) the jumps of p_1^*, if any, occur only when $q_1^*(t) = 0$ and the jumps are positive; (ii) in any interior subinterval I_i,

$$p_1^*(t)\mu_{12} - \phi_{12} - p_2^*(t)\mu_{22} \geq 0; \tag{A.16}$$

and (iii) in any boundary subinterval I_i,

$$p_1^*(t)\mu_{12} - \phi_{12} - p_2^*(t)\mu_{22} = 0. \tag{A.17}$$

Note that $p_1^*(\tau^*)\mu_{12} - \phi_{12} - p_2^*(\tau^*)\mu_{22} = 0$ if $\tau^* > 0$.

More specifically, suppose $p_1^*(A_k)$ has been defined for some k, with $p_1^*(A_k)\mu_{12} - \phi_{12} - p_2^*(A_k)\mu_{22} \geq 0$. If I_k is an interior subinterval, we set

$$p_1^*(t) = h_1(A_k - t) + p_1^*(A_k)$$

for $t \in [A_{k-1}, A_k)$. That is, p_1^* is continuous at A_k and has slope $-h_1$ in the subinterval I_k. Thus, $p_1^*(t)\mu_{12} - \phi_{12} - p_2^*(t)\mu_{22}$ has slope $h_2\mu_{22} - h_1\mu_{12} \leq 0$, which implies that $p_1^*(A_{k-1})\mu_{12} - \phi_{12} - p_2^*(A_{k-1})\mu_{22} \geq 0$. If I_k is a boundary subinterval, we set $p_1^*(A_{k-1}) = p_2^*(A_{k-1})\mu_{22}/\mu_{12} + \phi_{12}/\mu_{12}$ and $p_1^*(t) = p_1^*(A_{k-1}) - \frac{h_2\mu_{22}}{\mu_{12}}(t - A_{k-1})$ for $t \in (A_{k-1}, A_k)$. That is, p_1^* has a jump at A_k and has slope $-\frac{h_2\mu_{22}}{\mu_{12}}$ in the subinterval I_k. This ensures that $\phi_{12} - p_1^*(t)\mu_{12} = -p_2^*(t)\mu_{22}$ everywhere in I_k. The size of the jump at A_k is $p_1^*(A_k) - p_2^*(A_k)\mu_{22}/\mu_{12} - \phi_{12}/\mu_{12} \geq 0$, which is non-negative because $p_1^*(A_k)\mu_{12} - \phi_{12} - p_2^*(A_k)\mu_{22} \geq 0$. This way, we have defined p_1^* for $t \in [0, \tau^*)$ that satisfies conditions (i), (ii) and (iii).

Lastly, define the multipliers

$$\eta_1^*(t) = \begin{cases} 0, & t \in I_k \text{ and } I_k \text{ is an interior subinterval,} \\ h_1 - \frac{h_2\mu_{22}}{\mu_{12}}, & t \in I_k \text{ and } I_k \text{ is a boundary subinterval,} \\ 0, & t \in [\tau^*, \tau_1^*), \\ h_1, & t \in [\tau_1^*, \infty), \end{cases}$$

$$\eta_2^*(t) = \begin{cases} 0, & t \in [0, \tau_2^*), \\ h_2, & t \in [\tau_2^*, \infty), \end{cases}$$

$$\gamma_1^*(t) = \begin{cases} p_1^*(t)\mu_{11}, & t \in [0, \tau_1^*), \\ 0, & t \in [\tau_1^*, \infty), \end{cases}$$

$$\gamma_2^*(t) = \begin{cases} p_1^*(t)\mu_{12} - \phi_{12}, & t \in [0, \tau^*) \\ p_2^*(t)\mu_{22}, & t \in [\tau^*, \tau_2^*), \\ 0, & t \in [\tau_2^*, \infty), \end{cases}$$

$$\xi_{12}^*(t) = \begin{cases} 0, & t \in [0, \tau^*), \\ \phi_{12} - p_1^*(t)\mu_{12} + p_2^*(t)\mu_{22}, & t \in [\tau^*, \infty), \end{cases}$$

$$\xi_{22}^*(t) = \begin{cases} p_1^*(t)\mu_{12} - \phi_{12} - p_2^*(t)\mu_{22}, & t \in [0, \tau^*), \\ 0, & t \in [\tau^*, \infty), \end{cases}$$

and $\xi_{11}^*(t) = 0$ for all $t \geq 0$. Note that if $\tau^* > 0$, $p_1^*(t)\mu_{12} - \phi_{12} - p_2^*(t)\mu_{22} \geq 0$ for $t \in [0, \tau^*)$, and

$p_1^*(t)\mu_{12} - \phi_{12} - p_2^*(t)\mu_{22} \le 0$ for $t \in [\tau^*, \infty)$. Thus, ξ_{12}^* and ξ_{22}^* are non-negative.

The conditions (ODE), (ADJ), (J), and (H) are straightforwardly verified, i.e., by construction.

For (C), we only need to check that when $z_{22}^*(t) > 0$ in a boundary subinterval $[A_{k-1}, A_k)$, $\xi_{22}^*(t) = 0$. This holds because of (A.17). (Note that $z_{22}^*(A_k) = 0$.)

For (T), we have that

$$\nabla_{z_{11}} L(q^*(t), z^*(t), p^*(t), \eta^*(t), \gamma^*(t), \xi^*(t)) = -p_1^*(t)\mu_{11} + \gamma_1^*(t) - \xi_{11}^*(t) = 0$$

because $\xi_{11}^*(t) = 0$ and $\gamma_1^*(t) = p_1^*(t)\mu_{11}$. Next,

$$\nabla_{z_{22}} L(q^*(t), z^*(t), p^*(t), \eta^*(t), \gamma^*(t), \xi^*(t)) = -p_2^*(t)\mu_{22} + \gamma_2^*(t) - \xi_{22}^*(t) = 0$$

because $\xi_{22}^*(t) = \gamma_2^*(t) - p_2^*(t)\mu_{22}$ for $t \in [0, \tau^*)$, and $\xi_{22}^*(t) = 0$ and $\gamma_2^*(t) = p_2^*(t)\mu_{22}$ for $t \ge \tau^*$. Finally,

$$\nabla_{z_{12}} L(q^*(t), z^*(t), p^*(t), \eta^*(t), \gamma^*(t), \xi^*(t)) = \phi_{12} - p_1^*(t)\mu_{12} + \gamma_2^*(t) - \xi_{12}^*(t) = 0$$

because $\xi_{12}^*(t) = 0$ and $\gamma_2^*(t) = p_1^*(t)\mu_{12} - \phi_{12}$ for $t \in [0, \tau^*)$, and $\xi_{12}^*(t) = \phi_{12} - p_1^*(t)\mu_{12} + \gamma_2^*(t)$ for $t \ge \tau^*$.

Lastly, for (M), it is easy to see that $z_{11}^*(t)$ should always be maximal. Note that even when $q_1^*(t) = 0$ for some $t < \tau^*$, (M) follows because under the constraint that $z_{11}^*(t)\mu_{11} + z_{12}^*(t)\mu_{12} \le \lambda_1(t)$, the coefficients of $z_{11}^*(t)$ and $z_{12}^*(t)$ are $-p_1^*(t)\mu_{11}$ and $\phi_{12} - p_1^*(t)\mu_{12}$. For $t < \tau^*$, the coefficients of $z_{12}^*(t)$ and $z_{22}^*(t)$ are respectively $\phi_{12} - p_1^*(t)\mu_{12}$ and $-p_2^*(t)\mu_{22}$. Since $p_1^*(t)\mu_{12} - \phi_{12} \ge p_2^*(t)\mu_{22}$, it is optimal to have $z_{12}^*(t)$ being maximal. When $t \ge \tau^*$, $p_1^*(t)\mu_{12} - \phi_{12} \le p_2^*(t)\mu_{22}$, and so it is optimal to have $z_{22}^*(t)$ being maximal. This in turn implies $z_{12}^*(t) = 0$ for $t \in [\tau^*, \tau_2^*)$ is optimal (pool 2 has no spare capacity to help class 1). When $t \ge \tau_2^*$, $\phi_{12} - p_1^*(t)\mu_{12} \le 0$, so again $z_{12}^*(t) = 0$ is optimal.

Case II: $h_1\mu_{12} < h_2\mu_{22}$. Let $\tau_i = G_i^0(q_i(0))$ for $i = 1, 2$. In this case, the policy is that each

pool serves only its own class for $t \in [0, \tau_2)$. Under Assumption 1, $\tau_2 \geq \kappa_2$. Thus, $\lambda_2(t) < s_2\mu_{22}$ and $q_2(t) = 0$ for $t \geq \tau_2$. Then, pool 2 gives partial help to class 1 for $t \in [\tau_2, \tau_2 + \tau^*)$, where

$$\tau^* = \inf\{t \geq 0 : h_1\mu_{12}G_1^{\tau_2+t}(q_1(\tau_2 + t)) \leq \phi_{12}\}.$$

If $h_1\mu_{12}G_1^{\tau_2}(q_1(\tau_2)) \leq \phi_{12}$, $\tau^* = 0$. For $t \geq \tau_2 + \tau^*$, the inequality in (IIa) does not hold. Thus, each pool serves its own class only.

Note that if $\tau_1 \leq \tau_2$, we have $\tau^* = 0$, which will be discussed below. Suppose for now $\tau_1 > \tau_2$. Let $\tau_1^* = \tau_2 + \tau^* + G_1^{\tau_2+\tau^*}(q_1^*(\tau_2 + \tau^*))$ be the time at which queue 1 empties. Note that if $\tau^* > 0$, then $h_1\mu_{12}G_1^{\tau_2+\tau^*}(q_1(\tau_2 + \tau^*)) = \phi_{12}$ by continuity, so that $\tau_1^* = \tau_2 + \tau^* + \frac{\phi_{12}}{h_1\mu_{12}}$. Then, we consider the following queue length trajectory:

$$q_1^*(t) = \begin{cases} q_1 + \int_0^t (\lambda_1(s) - s_1\mu_{11})\, ds, & t \in [0, \tau_2), \\ q_1^*(\tau_2) + \int_{\tau_2}^t (\lambda_1(s) - s_1\mu_{11} - (s_2 - \lambda_2(s)/\mu_{22})\mu_{12})\, ds, & t \in [\tau_2, \tau_2 + \tau^*), \\ q_1^*(\tau_2 + \tau^*) + \int_{\tau_2+\tau^*}^t (\lambda_1(s) - s_1\mu_{11})\, ds, & t \in [\tau_2 + \tau^*, \tau_1^*) \\ 0, & t \in [\tau_1^*, \infty), \end{cases}$$

$$q_2^*(t) = \begin{cases} q_2 + \int_0^t (\lambda_2(s) - s_2\mu_{22})\, ds, & t \in [0, \tau_2), \\ 0, & t \in [\tau_2, \infty). \end{cases}$$

Note that the expression for $q_2^*(t)$ holds because, under Assumption 1, queue 2 will only be emptied once. Also, Assumption 1 implies that $q_1^*(t) > 0$ for $t \in [\tau_2, \tau_2+\tau^*)$, so that $z_{12}^*(t) = s_2 - \lambda_2(s)/\mu_{22}$ and $z_{11}^*(t) = s_1\mu_{11}$. Finally, Assumption 1 implies that $q_1^*(t) > 0$ for $t \in [0, \tau_2)$, except possibly for an initial interval containing 0 in which $\lambda_1(t) = s_1\mu_{11}$ if $q_1(0) = 0$. Thus, $z_{11}^*(t) = s_1\mu_{11}$ for $t \in [0, \tau_2)$.

Next, define the adjoint vectors

$$
\begin{aligned}
p_1^*(t) &= \begin{cases} h_1(\tau_1^* - t), & t \in [0, \tau_1^*), \\ 0, & t \in [\tau_1^*, \infty), \end{cases} \\
p_2^*(t) &= \begin{cases} h_2(\tau_2 - t) + h_1 \frac{\mu_{12}}{\mu_{22}} \tau^*, & t \in [0, \tau_2), \\ h_1 \frac{\mu_{12}}{\mu_{22}} (\tau_2 + \tau^* - t), & t \in [\tau_2, \tau_2 + \tau^*), \\ 0, & t \in [\tau_2 + \tau^*, \infty). \end{cases}
\end{aligned}
$$

Lastly, define the multipliers

$$
\begin{aligned}
\eta_1^*(t) &= \begin{cases} 0, & t \in [0, \tau_1^*), \\ h_1, & t \in [\tau_1^*, \infty), \end{cases} \\
\eta_2^*(t) &= \begin{cases} 0, & t \in [0, \tau_2), \\ h_2 - h_1 \frac{\mu_{12}}{\mu_{22}}, & t \in [\tau_2, \tau_2 + \tau^*) \\ h_2, & t \in [\tau_2 + \tau^*, \infty) \end{cases}
\end{aligned}
$$

$$
\begin{aligned}
\gamma_1^*(t) &= \begin{cases} p_1^*(t) \mu_{11}, & t \in [0, \tau_1^*), \\ 0, & t \in [\tau_1^*, \infty), \end{cases} \\
\gamma_2^*(t) &= \begin{cases} p_2^*(t) \mu_{22}, & t \in [0, \tau_2 + \tau^*), \\ 0, & t \in [\tau_2 + \tau^*, \infty) \end{cases}
\end{aligned}
$$

$$\xi_{12}^*(t) = \begin{cases} \phi_{12} - p_1^*(t)\mu_{12} + p_2^*(t)\mu_{22}, & t \in [0, \tau_2) \\ 0, & t \in [\tau_2, \tau_2 + \tau^*), \\ \phi_{12} - p_1^*(t)\mu_{12}, & t \in [\tau_2 + \tau^*, \infty) \end{cases}$$

and $\xi_{11}^*(t) = \xi_{22}^*(t) = 0$ for all $t \geq 0$. Note that $\eta_2^*(t) \geq 0$ because $h_2\mu_{22} \geq h_1\mu_{12}$ by assumption. In addition, because $h_1\mu_{12} \leq h_2\mu_{22}$, $\xi_{12}^*(t) = \phi_{12} - h_1\mu_{12}(\tau_1^* - t) + h_2\mu_{22}(\tau_2 - t) + h_1\mu_{12}\tau^*$ is non-increasing on $[0, \tau_2)$. If $\tau^* > 0$, $\xi_{12}^*(t) \to 0$ as $t \to \tau_2$ because $h_1\mu_{12}(\tau_1^* - \tau^* - \tau_2) = \phi_{12}$. If $\tau^* = 0$, $\xi_{12}^*(t) \to \phi_{12} - h_1\mu_{12}G_1^{\tau_2}(q_1(\tau_2)) \geq 0$ as $t \to \tau_2$. Thus,

$$\phi_{12} - h_1\mu_{12}(\tau_1^* - t) + h_2\mu_{22}(\tau_2 - t) + h_1\mu_{12}\tau^* \geq 0 \tag{A.18}$$

and $\xi_{12}^*(t) \geq 0$.

The conditions (ODE), (ADJ), (C), (J), and (H) are verified straightforwardly by construction. For (T),

$$\nabla_{z_{11}} L(q^*(t), z^*(t), p^*(t), \eta^*(t), \gamma^*(t), \xi^*(t)) = -p_1^*(t)\mu_{11} + \gamma_1^*(t) - \xi_{11}^*(t) = 0$$

because $\xi_{11}^*(t) = 0$ and $\gamma_1^*(t) = p_1^*(t)\mu_{11}$. Similarly,

$$\nabla_{z_{22}} L(q^*(t), z^*(t), p^*(t), \eta^*(t), \gamma^*(t), \xi^*(t)) = -p_2^*(t)\mu_{22} + \gamma_2^*(t) - \xi_{22}^*(t) = 0$$

because $\xi_{22}^*(t) = 0$ and $\gamma_2^*(t) = p_2^*(t)\mu_{22}$. Finally,

$$\nabla_{z_{12}} L(q^*(t), z^*(t), p^*(t), \eta^*(t), \gamma^*(t), \xi^*(t)) = \phi_{12} - p_1^*(t)\mu_{12} + \gamma_2^*(t) - \xi_{12}^*(t)$$

For $t \geq \tau_2 + \tau^*$, $\phi_{12} - p_1^*(t)\mu_{12} + \gamma_2^*(t) - \xi_{12}^*(t) = 0$ because the $\gamma_2^*(t) = 0$ and $\xi_{12}^*(t) = \phi_{12} - p_1^*(t)\mu_{12}$.

For $t \in [\tau_2, \tau_2 + \tau^*)$,

$$\phi_{12} - p_1^*(t)\mu_{12} + p_2^*(t)\mu_{22} = \phi_{12} - h_1\mu_{12}\left((\tau_1^* - t) - (\tau_2 + \tau^* - t)\right) = \phi_{12} - h_1\mu_{12}G_1^{\tau_2+\tau^*}(q_1^*(\tau_2+\tau^*))$$

is zero. Finally, for $t \in [0, \tau_2)$, $\phi_{12} - p_1^*(t)\mu_{12} + \gamma_2^*(t) - \xi_{12}^*(t) = 0$ because $\xi_{12}^*(t) = \phi_{12} - p_1^*(t)\mu_{12} + \gamma_2^*(t)$.

Lastly, for (M), it is easy to see that $z_{11}^*(t)$ should always be maximal. The coefficients of $z_{12}^*(t)$ and $z_{22}^*(t)$ are respectively $\phi_{12} - p_1^*(t)\mu_{12}$ and $-p_2^*(t)\mu_{22}$. When $t \in [0, \tau_2)$, $p_1^*(t)\mu_{12} - \phi_{12} \leq p_2^*(t)\mu_{22}$ (see (A.18)), and it can be verified that the inequality holds for all other t with equality for $t \in [\tau_2, \tau_2 + \tau^*)$. Thus, it is optimal to have $z_{22}^*(t)$ being maximal for $t \geq 0$. When $t < \tau_2$, $q_2^*(t) > 0$, and so $z_{12}^*(t) = 0$ is optimal (there is no spare capacity for pool 2 to help class 1). When $t \in [\tau_2, \tau_2 + \tau^*)$, $\phi_{12} - p_1^*(t)\mu_{12} \leq 0$, so it is optimal to maximize $z_{12}^*(t)$ in the sense of partial sharing, i.e. $z_{12}^*(t) = s_2 - z_{22}^*(t)$. When $t \geq \tau_2 + \tau^*$, $\phi_{12} - p_1^*(t)\mu_{12} \geq 0$, so it is optimal to have $z_{12}^*(t) = 0$. This completes the proof. □

A.5 Proof of Asymptotic Optimality

In this section, we provide the proof of Theorem 3. Lemma 8 establishes the first part of the theorem. For the second part of the theorem, we take the following steps:

1. We first show in Theorem 12 that there exists a fluid limit under any admissible control.
2. We then show in Theorem 13 that the fluid limit under the fluid translated control $\{\tilde{\nu}^n\}_{n\geq 1}$ follows the optimal fluid trajectory given in Section 3.
3. The key to verifying Theorem 13 is the continuity in the G and $\tilde{G}$ factors, which is established in Lemma 9 and Lemma 10, respectively.

For notational convenience, we define the scaled version of the estimated arrival rate:

$$\tilde{\lambda}_i^n(t) := \frac{\Lambda_i^n(t)}{n} = \lambda_i(t) + \epsilon_i^n(t), \text{ where } \epsilon_i^n(t) = E_i^n(t)/n.$$

Then, we can rewrite $\tilde{G}^t_{i,n}(nx)$ as $\tilde{G}^t_{i,n}(nx) = \inf\left\{\Delta \geq 0 : \int_t^{t+\Delta}(s_i\mu_{ii} - \tilde{\lambda}^n_i(s))\,ds = x\right\}$. With a little abuse of notation, we redefine the input of the function as x, instead of nx, i.e.,

$$\tilde{G}^t_{i,n}(x) = \inf\left\{\Delta \geq 0 : \int_t^{t+\Delta}(s_i\mu_{ii} - \tilde{\lambda}^n_i(s))\,ds = x\right\}. \tag{A.19}$$

We start from proving the following lemma, which establishes the results in the first part of Theorem 3.

Lemma 8. *For any admissible control π^n for system n, $\bar{V}^{n,\pi^n}(x) \geq \bar{V}^*(x)$.*

Proof. Proof. We suppress the superscript π^n from the corresponding processes to simplify the notation. Let $g^n_i(t,x) = \lambda^n_i(t) - \sum_j(\pi^n_t(x))_{ij}n\mu_{ij}$ for $t \in [0, T(x)], x \in \mathbb{N}^2$. Note that

$$M_i(\cdot) := X^n_i(\cdot) - nx_i - \int_0^{\cdot} g^n_i(s, X^n(s))\,ds$$

is a zero-mean martingale by the Dynkin formula. Taking expectation gives

$$\mathbb{E}_\pi[X^n_i(t)] = nx_i + \int_0^t \mathbb{E}_\pi[g^n_i(s, X^n(s))]\,ds = nx_i + \int_0^t \left(n\lambda_i(s) - n\sum_j \mathbb{E}_\pi[Z^n_{ij}(s)]\mu_{ij}\right) ds \tag{A.20}$$

for $t \in [0, T(x)]$.

Consider the (fluid) policy u: $u_t(\mathbb{E}_\pi[X^n(t)/n]) = \mathbb{E}_\pi[Z^n(t)]$, i.e., if at time t we have $q(t) = \mathbb{E}_\pi[X^n(t)/n]$, then $z(t) = \mathbb{E}_\pi[Z^n(t)]$. $u_t(x)$ for other values of t and x can be defined arbitrarily. Note that for each j, $\sum_i z_{ij}(t) = \sum_i \mathbb{E}[Z^n_{ij}(t)] \leq s_j$, and (A.20) implies that

$$0 \leq \mathbb{E}_\pi[X^n_i(t)/n] = x_i + \int_0^t \left(\lambda_i(s) - \sum_j \mathbb{E}_\pi[Z^n_{ij}(s)]\mu_{ij}\right) ds$$

for $t \in [0, T(x)]$. Thus, u is an admissible control for the fluid problem and the corresponding fluid dynamics takes the form:

$$q_i(t) = \mathbb{E}_\pi[X^n_i(t)/n], \quad z_{ij}(t) = \mathbb{E}_\pi[Z^n_{ij}(t)] \quad \text{for } t \in [0, T(x)].$$

Then,

$$\begin{aligned}\bar{V}^{n,\pi^n}(x) &= \mathbb{E}_\pi\left[\int_0^{T(x)}\left(\sum_i \frac{h_i}{n}X_i^n(t) + \sum_{i\neq j}\phi_{ij}Z_{ij}^n(t)\right)dt\right] \\ &= \int_0^{T(x)}\left(\sum_i h_i q_i(t) + \sum_{i\neq j}\phi_{ij}z_{ij}(t)\right)dt \\ &\geq \bar{V}^*(x).\end{aligned}$$

□

To prove the second part of Theorem 3, we first introduce a notion of a fluid limit and show in Theorem 12 below that there exists a fluid limit under any admissible control.

Theorem 12. *There exists almost surely a subsequence* $\{n_k : k \in \mathbb{N}\}$ *such that* $(\bar{X}^{n_k}, \bar{Y}^{n_k}) \to (\bar{X}, \bar{Y})$ *uniformly on compact intervals (u.o.c.) as* $n \to \infty$. *Moreover,* $(\bar{X}, \bar{Y})$ *is Lipschitz continuous and satisfies*

(a) $\bar{X}(0) = x$, $\bar{X}(t) \geq 0$ *for* $t \geq 0$;

(b) $\bar{X}_i(t) = \bar{X}_i(0) + \int_0^t \lambda_i(s)\,ds - \sum_j \bar{Y}_{ij}(t)\mu_{ij}$;

(c) $\bar{Y}(\cdot)$ *is non-decreasing with* $\bar{Y}_{ij}(0) = 0$;

(d) $\sum_i(\bar{Y}_{ij}(t) - \bar{Y}_{ij}(s)) \leq s_j(t-s)$ *for* $j = 1, 2$ *and* $0 \leq s < t$.

Proof. Proof. By Strong Law of Large Numbers, the scaled number of arrivals

$$\bar{A}_i^n(t) := \frac{1}{n}S_i\left(\int_0^t \lambda_i^n(s)\,ds\right) = \frac{1}{n}S_i\left(n\int_0^t \lambda_i(s)\,ds\right),$$

where S_i is a rate-1 Poisson process, satisfies

$$\bar{A}_i^n(t) \to \int_0^t \lambda_i(s)\,ds$$

uniformly on compact sets (u.o.c.) as $n \to \infty$. The rest of the proof follows from Theorem 6.5 in [80]. □

From Theorem 12, there exists a fluid limit for the sequence of systems under the fluid translated control $\{\tilde{v}^n\}_{n\geq 1}$ of Theorem 3. We next show that any fluid limit of the sequence of systems under policy $\{\tilde{v}^n\}_{n\geq 1}$ is equal to the optimal fluid trajectory. Let (q^*, y^*) denote the optimal fluid trajectory, i.e., (q^*, z^*) is defined in Theorem 1 and $y^*_{ij}(t) = \int_0^t z^*_{ij}(s)\, ds$.

Theorem 13. *Let $(\bar{Q}, \bar{Y}) = (\bar{q}_1, \bar{q}_2, \bar{Y}_{11}, \bar{Y}_{12}, \bar{Y}_{22})$ be a fluid limit for the sequence of systems under policy $\{\tilde{v}_n\}_{n\geq 1}$. Then, $(\bar{Q}, \bar{Y}) = (q^*, y^*)$.*

Before we prove Theorem 13, we first present two auxiliary lemmas that will be used in the proof.

Lemma 9. *If $t \mapsto \bar{x}_i(t)$ is continuous, $t \mapsto G_i^t(\bar{x}_i(t))$ is also continuous.*

Proof. Proof. We will show that $t \mapsto G_i^t(x)$ is continuous for any fixed $x \geq 0$. To simplify notation, let $a_i(t) = s_i\mu_{ii} - \lambda_i(t)$. Fix $t > 0$ and let $\epsilon > 0$ be an arbitrarily small constant.

Case I: $t > \kappa_i$. We may assume that $t - \epsilon > \kappa_i$. Note that $\int_{t-\delta}^{t+G_i^t(x)} a_i(s)\, ds \geq x$ if $\delta \leq \epsilon$. Thus,

$$G_i^{t-\delta}(x) \leq G_i^t(x) + \delta \leq G_i^t(x) + \epsilon.$$

Next, note that $\xi_1(\delta) := \int_{t-\delta}^{t-\delta-\epsilon+G_i^t(x)} a_i(s)\, ds$ is a continuous function of δ. As $\xi_1(0) < x$, there exists $\delta_1 > 0$ such that for all $0 \leq \delta < \delta_1$, $\xi_1(\delta) < x$. Thus,

$$G_i^{t-\delta}(x) \geq G_i^t(x) - \epsilon.$$

Above all, for $0 \leq \delta < \epsilon \wedge \delta_1$, $|G_i^{t-\delta}(x) - G_i^t(x)| < \epsilon$, i.e., we have left-continuity.

For the right continuity, we first note that for $0 \leq \delta < \epsilon$, $\int_{t+\delta}^{t+G_i^t(x)} a_i(s)\, ds < x$. Thus,

$$G_i^{t+\epsilon}(x) > G_i^t(x) - \delta > G_i^t(x) - \epsilon.$$

Next, note that $\xi_2(\delta) := \int_{t+\delta}^{t+\delta+G_i^t(x)+\epsilon} a_i(s)\,ds$ is a continuous function of δ. As $\xi_2(0) > x$, there exists $\delta_2 > 0$ such that for all $0 \le \delta < \delta_2$, $\xi_2(\delta) > x$. Thus,

$$G_i^{t+\delta}(x) \le G_i^t(x) + \epsilon.$$

Above all, for $0 \le \delta < \epsilon \wedge \delta_2$, $|G_i^{t+\delta}(x) - G_i^t(x)| < \epsilon$, i.e., we have right-continuity.

Case II: $t < \kappa_i$. Note that for $0 \le \delta \le \kappa_i - t$, $\int_{t+\delta}^{t+G_i^t(x)} a_i(s)\,ds \ge x$. Thus,

$$G_i^{t+\delta}(x) \le G_i^t(x) - \delta \le G_i^t(x).$$

Next, note that $\xi_3(\delta) = \int_{t+\delta}^{t+G_i^t(x)-\epsilon/2} a_i(s)\,ds$ is a continuous function of δ. As $\xi_3(0) < x$, there exists $\delta_3 > 0$, such that for $0 \le \delta \le \delta_3$ $\xi_3(\delta) < x$ Thus,

$$G_i^{t+\delta}(x) \ge G_i^t(x) - \epsilon/2 - \delta.$$

Above all, for $0 \le \delta \le \delta_3 \wedge \epsilon/2$, $|G_i^{t+\delta}(x) - G_i^t(x)| = G_i^t(x) - G_i^{t+\delta}(x) \le \epsilon$, i.e., we have right-continuity.

For the left continuity, we first note that for $0 \le \delta \le \kappa_i - t$, $\int_{t-\delta}^{t+G_i^t(x)} a_i(s)\,ds \le x$. Thus,

$$G_i^{t-\delta}(x) \ge G_i^t(x) + \delta \ge G_i^t(x).$$

Next, note that $\xi_4(\delta) := \int_{t-\delta}^{t+G_i^t(x)+\epsilon/2} a_i(s)\,ds$ is a continuous function of δ. As $\xi_4(0) > x$, there exists $\delta_4 > 0$ such that for $0 \le \delta \le \delta_4$, $\xi_4(\delta) > x$ Thus,

$$G_i^{t-\delta}(x) \le G_i^t(x) + \epsilon/2 + \delta.$$

Above all, for $0 \le \delta \le \delta_4 \wedge \epsilon/2$, $|G_i^{t-\delta}(x) - G_i^t(x)| = G_i^{t-\delta}(x) - G_i^t(x) \le \epsilon$, i.e., we have left-continuity.

Case III: $t = \kappa_i$. The right-continuity follows the right-continuity argument of case I and the

left-continuity follows the left-continuity argument of case II.

The proof that $t \mapsto G_i^t(\bar{q}_i(t))$ is continuous in t follows similarly. □

For the next lemma, recall the definition of $\tilde{G}_{i,n}^t(x)$ in (A.19).

Lemma 10. *If $X_1(t)$ is bounded on $[0, \kappa_1]$, $X_1^n(t) \geq 0$ and $X_1^n(t) \to X_1(t)$ uniformly on $t \in [0, \kappa_1]$, then*

$$\tilde{G}_{1,n}^t(X_1^n(t)) \to G_1^t(X_1(t))$$

uniformly on $t \in [0, \kappa_1]$ as $n \to \infty$. The same is true on the interval $[\kappa_1, A]$ for any $A > \kappa_1$. The same is also true for class 2 on any closed bounded interval.

Proof. Proof of Lemma 10. Note that the assumptions imply that there exist $N_0 > 0$ and $B > 0$ such that $X_1^n(t) \leq B$ for all $n > N_0$ and all $t \in [0, \kappa_1]$. Let $\alpha > 0$ (we use α instead of ϵ to avoid confusion with the error function $\epsilon^n(\cdot)$). It suffices, then, to show that there exist $N_1 > 0$ and $\delta > 0$ such that

$$|G_1^t(x_1) - \tilde{G}_{1,n}^t(x_2)| \leq \alpha$$

for all $n > N_1$, $t \in [0, \kappa_1]$ and $0 \leq x_1, x_2 \leq B$ and $|x_1 - x_2| \leq \delta$.

Note that because $G_1^t(B)$ is a continuous function of t, $G_1^t(x)$ is bounded, say by C, for $t \in [0, \kappa_1]$ and $0 \leq x \leq B$. Let

$$D(\alpha) = \inf_{0 \leq s \leq C} \int_{\kappa_1 + s}^{\kappa_1 + s + \alpha} (s_1 \mu_{11} - \lambda_1(u))\, du.$$

Note that $D(\alpha) > 0$ because $s_1 \mu_{11} > \lambda_1(u)$ for $u > \kappa_1$.

Now, observe that

$$\begin{aligned}
&\int_t^{t+G_1^t(x_1)+\alpha} (s_1\mu_{11} - \lambda_1(u) - \epsilon_1^n(u))\, du \\
\geq &\int_t^{t+G_1^t(x_1)+\alpha} (s_1\mu_{11} - \lambda_1(u) - |\epsilon_1^n(u)|)\, du \\
= &\int_t^{t+G_1^t(x_1)} (s_1\mu_{11} - \lambda_1(u))\, du + \int_{t+G_1^t(x_1)}^{t+G_1^t(x_1)+\alpha} (s_1\mu_{11} - \lambda_1(u))\, du - \int_t^{t+G_1^t(x_1)+\alpha} |\epsilon_1^n(u)|\, du \\
\geq &x_1 + D(\alpha) - (C+\alpha) \sup_{t\leq u\leq t+C+\alpha} |\epsilon_1^n(u)| \\
\geq &x_1 + D(\alpha)/2
\end{aligned}$$

for $n > N_2$ large enough, since $\epsilon^n(\cdot) \to 0$ u.o.c. by assumption. Therefore, if $x_2 < x_1 + D(\alpha)/2$, then

$$\tilde{G}_{1,n}^t(x_2) < G_1^t(x_1) + \alpha.$$

Next, observe that

$$\begin{aligned}
&\int_t^{t+G_1^t(x_1)-\alpha} (s_1\mu_{11} - \lambda_1(u) - \epsilon_1^n(u))\, du \\
\leq &\int_t^{t+G_1^t(x_1)-\alpha} (s_1\mu_{11} - \lambda_1(u) + |\epsilon_1^n(u)|)\, du \\
= &\int_t^{t+G_1^t(x_1)} (s_1\mu_{11} - \lambda_1(u))\, du - \int_{t+G_1^t(x_1)-\alpha}^{t+G_1^t(x_1)} (s_1\mu_{11} - \lambda_1(u))\, du + \int_t^{t+G_1^t(x_1)-\alpha} |\epsilon_1^n(u)|\, du \\
\leq &x_1 - D(\alpha) + (C-\alpha) \sup_{t\leq u\leq t+C-\alpha} |\epsilon_1^n(u)| \\
\leq &x_1 - D(\alpha)/2
\end{aligned}$$

for $n > N_2$ large enough, as before. Therefore, if $x_2 > x_1 - D(\alpha)/2$, then

$$\tilde{G}_{1,n}^t(x_2) > G_1^t(x_1) - \alpha.$$

(We have assumed above that $t + G_1^t(x_1) - \alpha \geq \kappa_1$, since otherwise it is trivial that $\tilde{G}_{1,n}^t(x_2) >$

$G_1^t(x_1) - \alpha$.) Hence, if $|x_2 - x_1| < D(\alpha)/2$, then

$$|G_1^t(x_1) - \tilde{G}_{1,n}^t(x_2)| \leq \alpha$$

for $n > N_2$, as required. The proof for $[\kappa_1, A]$ and for class 2 are similar. □

Next, we prove Theorem 13.

Proof. Proof of Theorem 13. We divide the analysis into two cases.

Case I: $h_1\mu_{12} > h_2\mu_{22}$. Let $T_1 \geq 0$ be the time that pool 2 stops helping class 1 under the optimal fluid control. That is,

$$h_1\mu_{12}G_1^t(q_1^*(t)) - \phi_{12} > h_2\mu_{22}G_2^t(q_2^*(t))$$

for $t \in [0, T_1)$ and

$$h_1\mu_{12}G_1^t(q_1^*(t)) - \phi_{12} < h_2\mu_{22}G_2^t(q_2^*(t))$$

for $t > T_1$.

Supppose $T_1 > 0$, so that $h_1\mu_{12}G_1^{T_1}(q_1^*(T_1)) - \phi_{12} = h_2\mu_{22}G_2^{T_1}(q_2^*(T_1))$. We partition $[0, T_1)$ into finitely many subintervals $[A_i, A_{i+1})$ $(i = 0, \cdots, n)$ such that $0 = A_0 < \cdots < A_{n+1} = T_1$, and on each open subinterval $t \in (A_i, A_{i+1})$, either $q_1^*(t) = 0$ only or $q_1^*(t) > 0$ only. The fact that are finitely many such intervals comes from piecewise monotonicity assumption, i.e., Assumption 5 and the proof of Theorem 1.

We next show inductively that $(\bar{Q}, \bar{Y}) = (q^*, y^*)$ on each $[A_i, A_{i+1})$. Suppose that $\bar{Q}(A_i) = q^*(A_i)$ and $\bar{Y}(A_i) = y^*(A_i)$ for some i.

We first consider the case that $q_1^*(t) > 0$ on (A_i, A_{i+1}). Because $q_1^*(t)$ decreases at the maximum possible rate for $t < T_1$ under the optimal fluid control, we have that $\bar{Q}_1(t) \geq q_1^*(t) > 0$ and $\bar{Q}_2(t) \leq q_2^*(t)$ for all $t \in (A_i, A_{i+1})$. Hence, for $s \in (A_i, A_{i+1})$, $h_1\mu_{12}G_1^s(\bar{Q}_1(s)) - \phi_{12} >$

$h_2\mu_{22}G_2^s(\bar{Q}_2(s))$. By continuity of $t \mapsto G_i^t(\bar{Q}_i(t))$ (Lemma 9), we have

$$h_1\mu_{12}G_1^t(\bar{Q}_1(t)) - \phi_{12} > \delta + h_2\mu_{22}G_2^t(\bar{Q}_2(t))$$

for all $t \in [s-\epsilon, s+\epsilon]$, for some $\epsilon, \delta > 0$. Since $(Q_1^n(t)/n, Q_2^n(t)/n)) \to (\bar{Q}_1(t), \bar{Q}_2(t))$ u.o.c., we have by Lemma 10 that

$$h_1\mu_{12}\tilde{G}_{1,n}^t(Q_1^n(t)/n) - \phi_{12} > \delta/2 + h_2\mu_{22}\tilde{G}_{2,n}^t(Q_2^n(t)/n) \text{ and } Q_1^n(t) > s_1 + s_2$$

for all $t \in [s-\epsilon, s+\epsilon]$, for n large enough. According to the scheduling policy, for each such nth system and $t \in [s-\epsilon, s+\epsilon]$, pool 2 prioritizes class 1, so that $d\bar{Y}^n(t)/dt = (s_1, s_2, 0)$. In addition, since $h_1\mu_{12}G_1^t(q_1^*(t)) - \phi_{12} > h_2\mu_{22}G_2^t(q_2^*(t))$,

$$h_1\mu_{12}G_1^t(\bar{Q}_1(t)) - \phi_{12} > h_2\mu_{22}G_2^t(\bar{Q}_2(t))$$

for all $t \in (A_i, A_{i+1})$. Then, $d\bar{Y}(t)/dt = dy^*(t)/dt$ for all (regular) $t \in (A_i, A_{i+1})$, which implies that $(\bar{Q}, \bar{Y}) = (q^*, y^*)$ on $t \in [A_i, A_{i+1}]$. In particular, $\bar{Q}(A_{i+1}) = q^*(A_{i+1})$ and $\bar{Y}(A_{i+1}) = y^*(A_{i+1})$. This technique – applying Lemma 10 to derive inequalities involving $\tilde{G}_{i,n}^t(Q_i^n(t)/n)$ based on inequalities involving $G_i^t(\bar{Q}_i(t))$ – is also used in subsequent cases in the proof.

We next consider the case $q_1^*(t) = 0$ on (A_i, A_{i+1}). Suppose that $\bar{Q}_1(t) > 0$ for some $t \in (A_i, A_{i+1})$. Let $S = \sup\{s \le t : \bar{Q}_1(s) = 0\}$. Note that $\bar{Q}_1(S) = 0$ by continuity, and that $S \ge A_i$ because $\bar{Q}_1(A_i) = 0$. By definition $\bar{Q}_1(s) > 0$ for all $s \in (S, t]$. Then, following similar lines of analysis as in the case when $q_1^*(t) > 0$, $d\bar{Y}^n(t)/dt = (s_1, s_2, 0)$ for large enough n and $d\bar{Q}_1(t)/dt = \lambda_1(t) - s_1\mu_{11} - s_2\mu_{12} \le 0$. In this case, $\bar{Q}_1(s)$ is non-increasing in $(S, t]$. Thus, $\bar{Q}_1(s) \ge \bar{Q}_1(t) > 0$ for $s \in (S, t]$. This implies that $\bar{Q}_1(s)$ is not continuous at $s = S$, a contradiction. This implies that $\bar{Q}_1(t) = 0$ for $t \in (A_i, A_{i+1})$. In addition, by Assumption 1, $\bar{Q}_2(t) > 0$ on (A_i, A_{i+1}), which implies that $d(\bar{Y}_{12}(t) + \bar{Y}_{22}(t))/dt = s_2$ on (A_i, A_{i+1}). As

$d\bar{Y}_{11}(t)/dt = s_1$ and $\bar{Q}_1(t) = 0$, we have

$$\frac{d\bar{Y}_{12}(t)}{dt} = \frac{\lambda_1(t) - s_1\mu_{11}}{\mu_{12}} = \frac{dy^*_{12}(t)}{dt} \text{ for } t \in (A_i, A_{i+1}).$$

Thus, $(\bar{Q}, \bar{Y}) = (q^*, y^*)$ on $[A_i, A_{i+1}]$.

By induction, $(\bar{Q}, \bar{Y}) = (q^*, y^*)$ on $[0, T_1)$.

Lastly, we analyze $(\bar{Q}(t), \bar{Y}(t))$ for $t > T_1$. Note that for $T_1 > 0$, $q^*_1(T_1) > 0$. Let $T_2 > T_1$ be the first time q^*_1 empties, i.e., $T_2 = \inf\{t \geq T_1 : q^*_1(t) = 0\}$. Let $S_2 > T_1$ be the first time $\bar{Q}_1$ empties, i.e., $S_2 = \inf\{t \geq T_1 : \bar{Q}_1(t) = 0\}$. For $T_1 < t < S_2$, $\bar{Q}_1(t) > 0$. By the same reasoning as before, we have that there exists $\epsilon > 0$ such that $Q^n_1(s) > s_1$ for $s \in [t - \epsilon, t + \epsilon]$, for n sufficiently large. Thus, according to our scheduling policy, $d\bar{Y}_{11}(t)/dt = s_1$. This implies that $h_1\mu_{12}G^t_1(\bar{Q}_1(t)) - \phi_{12}$ decreases at rate at least $h_1\mu_{12}$, whereas $h_2\mu_{22}G^t_2(\bar{Q}_2(t))$ decreases at rate at most $h_2\mu_{22}$. Since $h_1\mu_{12} > h_2\mu_{22}$ and $h_1\mu_{12}G^{T_1}_1(\bar{Q}_1(T_1)) - \phi_{12} \leq h_2\mu_{22}G^{T_1}_2(\bar{Q}_2(T_1))$ (strict inequality is possible if $T_1 = 0$), we have that for all $T_1 < t < S_2$,

$$h_1\mu_{12}G^t_1(\bar{Q}_1(t)) - \phi_{12} < h_2\mu_{22}G^t_2(\bar{Q}_2(t)).$$

Following similar lines of argument as before, we can show that for each $t \in (T_1, S_2)$ and large enough n, pool 2 only serves class 2 in the nth system at time t. Therefore, $\bar{Q}_1(t) = q^*_1(t)$ for $t \in (T_1, S_2)$ and $S_2 = T_2$. For $t > T_2$, $\bar{Q}_1(t) = 0$. Hence, $\bar{Q}_1(t) = q^*_1(t)$ for all t.

The above also establishes that $\bar{Q}_2 = q^*_2$, $\bar{Y}_{12} = y^*_{12}$ and $\bar{Y}_{22} = y^*_{22}$ on (T_1, T_3), where $T_3 = \inf\{t \geq T_1 : q^*_2(t) = 0\}$ is the common emptying time of the class 2 queue for both q^*_2 and $\bar{Q}_2$. For $t > T_3$, we again have that

$$h_1\mu_{12}G^t_1(\bar{Q}_1(t)) - \phi_{12} < h_2\mu_{22}G^t_2(\bar{Q}_2(t)),$$

as shown above if $t < S_2$, and trivially if $t \geq S_2$. Therefore $\bar{Y}_{12} = y^*_{12}$ for $t > T_3$, and pool 2 only serves its own class for large enough systems. Since $\bar{Q}_1 = q^*_1$, this also implies that $\bar{Y}_{11} = y^*_{11}$.

Finally, for $t > S_2$, since $\bar{Q}_2'(t) \leq 0$ whenever $\bar{Q}_2(t) > 0$, $\bar{Q}_2(t) = q_2^*(t) = 0$ and $\bar{Y}_{22}(t) = y_{22}^*(t)$.

Case II: $h_1\mu_{12} < h_2\mu_{22}$. The case of interest is the one where pool 2 provides partial help to queue 1 in q^*, i.e., after queue 2 has emptied, queue 1 is still large. If no partial help occurs, the result will follow from the analysis in case I.

Let $T_1 = \inf\{t \geq \kappa_2 : q_2^*(t) = 0\}$. Similarly, let $S_1 = \inf\{t \geq \kappa_2 : \bar{Q}_2(t) = 0\}$. For $t < S_1 \wedge T_1$, we have $\bar{Q}_i(t) > 0$. By the same reasoning as in Case I, $d\bar{Y}_{ii}(t)/dt = s_i = dy_{ii}^*(t)/dt$. This implies $S_1 = T_1$. Thus, $(\bar{Q}, \bar{Y}) = (q^*, y^*)$ on $t \in [0, T_1]$. For $t > T_1$, we can show as in Case I that $\bar{Q}_2'(t) \leq 0$ if $\bar{Q}_2(t) > 0$, so that $\bar{Q}_2(t) = 0 = q_2^*(t)$. Hence also $\bar{Y}_{22}(t) = y_{22}^*(t)$ for $t > T_1$.

Next, let $T_2 \geq T_1$ be the time at which partial help by pool 2 ends under the optimal fluid control. For $t \in [T_1, T_2)$, $h_1\mu_{12}G_1^t(q_1^*(t)) > \phi_{12}$ and $q_2^*(t) = 0$. Note that $\bar{Q}_2(t) = 0$ for $t \in [T_1, T_2)$ as well. Because the optimal fluid control minimizes $q_1(t)$ for $t \in [T_1, T_2)$ while keeping $q_2(t)$ at zero, we have that $\bar{Q}_1(t) \geq q_1^*(t) > 0$ for $t \in [T_1, T_2)$. Hence, $h_1\mu_{12}G_1^t(\bar{Q}_1(t)) > \phi_{12}$ for $t \in [T_1, T_2)$. Thus, $d(\bar{Y}_{12}(t) + \bar{Y}_{22}(t))/dt = s_2 = d(y_{12}^*(t) + y_{22}^*(t))/dt$ for $t \in (T_1, T_2)$. Since $\bar{Y}_{22} = y_{22}^*$, this implies that $\bar{Y}_{12}(t) = y_{12}^*(t)$ for $t \in [T_1, T_2)$. Hence also $\bar{Q}_1(t) = q_1^*(t)$ for $t \in [T_1, T_2)$.

For $t > T_2$, $h_1\mu_{12}G_1^t(\bar{Q}_1(t)) < \phi_{12}$. Following similar lines of argument as before, $\bar{Y}_{12}(t) = y_{12}^*(t) = y_{12}^*(T_2)$, $\bar{Y}_{11}(t) = y_{11}^*(t)$ and $\bar{Q}_{11}(t) = q_{11}^*(t)$. □

With Theorem 13, we are now ready to prove the second part of Theorem 3.

Proof. Proof. Recall that $(\bar{X}^n, \bar{Y}^n) \to (q^*, y^*)$ uniformly on $[0, T(x)]$ almost surely, which implies that $\bar{X}_i^n(t)$ is uniformly bounded in n and t. Also, note that $\bar{Y}_{12}^n(T(x)) \leq s_2 T(x)$ is bounded. We have

$$
\begin{aligned}
\bar{V}^{n,\tilde{\nu}^n}(x) =& \mathbb{E}\left[\int_0^{T(x)} \sum_i h_i \bar{X}_i^n(t)\, dt + \phi_{12}\bar{Y}_{12}^n(T(x))\right] \\
=& \int_0^{T(x)} \sum_i h_i \mathbb{E}[\bar{X}_i^n(t)]\, dt + \phi_{12}\mathbb{E}[\bar{Y}_{12}^n(T(x))] \text{ since } \bar{X}_i^n(t) \geq 0 \\
\to& \int_0^{T(x)} \sum_i h_i q_i^*(t)\, dt + \phi_{12} y_{12}^*(T(x)) \text{ as } n \to \infty \text{ by bounded convergence} \\
=& \bar{V}^*(x).
\end{aligned}
$$

□

A.6 Proof of the optimal fluid control for the N-model with multiple demand surges

Proof. Proof of Theorem 2. We will construct the optimal primal and dual trajectories under the policy characterized in Theorem 2 and show that the conditions in Theorem 11 are satisfied.

Case I: $h_1\mu_{12} > h_2\mu_{22}$**.** In this case,

$$
\begin{aligned}
H(q(t), z(t), p(t), t) =& h_1 q_1(t) + h_2 q_2(t) + \phi_{12} z_{12}(t) \\
&+ p_1(t)\left(\lambda_1(t) - \mu_{11} z_{11}(t) - \mu_{12} z_{12}(t)\right) + p_2(t)\left(\lambda_2(t) - \mu_{22} z_{22}(t)\right)
\end{aligned}
$$

and

$$
\begin{aligned}
L(q(t), z(t), p(t), \eta(t), \xi(t), \gamma(t), t) =& H(q(t), z(t), p(t), t) - \eta_1(t) q_1(t) - \eta_2(t) q_2(t) \\
&- \xi_{11}(t) z_{11}(t) - \xi_{12}(t) z_{12}(t) - \xi_{22}(t) z_{22}(t) \\
&+ \gamma_1(t)(z_{11}(t) - s_1) + \gamma_2(t)(z_{12}(t) + z_{22}(t) - s_2).
\end{aligned}
$$

There are three scenarios to consider, depending on the queue lengths at time κ_b (i.e., the beginning of the second demand surge).

Scenario A: $q_1^*(\kappa_b) = q_2^*(\kappa_b) = 0$. That is, both queues have been emptied by the start of the second demand surge. Following the proof of Theorem 1, we obtain $q_i^*, p_i^*, z_{ij}^*, \eta_i^*, \xi_{ij}^*, \gamma_j^*$

for $t \in [0, \kappa_b)$. We can then solve an "independent" optimal control problem using the initial state $(0, 0)$ to obtain the values of $q_i^*, p_i^*, z_{ij}^*, \eta_i^*, \xi_{ij}^*, \gamma_j^*$ for $t \in [\kappa_b, \infty)$. The verification of the conditions in Theorem 11 follows exactly the same lines of analysis as the proof of Theorem 1.

Scenario B: $q_1^*(\kappa_b) > 0$. This implies that $q_1^*(t) > 0$ and $t + G_1^t(q_1^*(t)) > \kappa_b$ for $t \in [0, \kappa_b)$ (except possibly $q_1^*(t) = 0$ in an initial interval containing zero). In this case, $G_1^t(q_1^*(t))$ decreases at rate at least one until it hits zero. As such, pool 2 does not resume helping class 1 once it stops helping class 1. Pool 2 gives priority to class 1 for an initial time

$$\tau^* = \inf\{t \geq 0 : h_1\mu_{12}G_1^t(q_1(t)) - \phi_{12} \leq h_2\mu_{22}G_2^t(q_2(t))\}. \tag{A.21}$$

Thereafter, each queue is served by its primary server pool only, and is emptied at time $\tau_i^* = \tau^* + G_i^{\tau^*}(q_i^*(\tau^*))$. Note that $\tau_1^* \notin (\kappa_b, \kappa_c]$, because the class 1 queue cannot be emptied at time $t \in (\kappa_b, \kappa_c]$ without help from pool 2.

The optimal queue length trajectory follows:

$$q_1^*(t) = \begin{cases} q_1 + \int_0^t (\lambda_1(s) - s_1\mu_{11} - z_{12}^*(s)\mu_{12})\, ds, & t \in [0, \tau^*), \\ q_1^*(\tau^*) + \int_{\tau_*}^t (\lambda_1(s) - s_1\mu_{11})\, ds, & t \in [\tau^*, \tau_1^*), \\ 0, & t \in [\tau_1^*, \infty), \end{cases}$$

$$q_2^*(t) = \begin{cases} q_2 + \int_0^t (\lambda_2(s) - z_{22}^*(s)\mu_{22})\, ds, & t \in [0, \tau^*), \\ q_2^*(\tau^*) + \int_{\tau^*}^t (\lambda_2(s) - s_2\mu_{22})\, ds, & t \in [\tau^*, \tau_2^*), \\ 0, & t \in [\tau_2^*, \infty). \end{cases}$$

Note that $q_i^*(t)$'s have exactly the same dynamics as $q_i^*(t)$'s in Case I in the proof of Theorem 1. Thus, the proof of this scenario follows exactly the same lines of analysis as Case I in Theorem 1 (i.e., the two demand surges can be treat as a single demand surge).

Scenario C: $q_1^*(\kappa_b) = 0$ and $q_2^*(\kappa_b) > 0$. Pool 2 gives priority to class 1 for an initial time

$$\tau^* = \inf\{t \geq 0 : h_1\mu_{12}G_1^t(q_1(t)) - \phi_{12} \leq h_2\mu_{22}G_2^t(q_2(t))\}. \tag{A.22}$$

At time $\tau_1^* = \tau^* + G_1^{\tau^*}(q_1^*(\tau^*)) \leq \kappa_b$, pool 1 is emptied.

Next, at time κ_b, $G_1^t(q_1(t))$ jumps from zero to a positive number due to the second demand surge, and hence pool 2 may resume helping class 1. Let

$$\tau' = \inf\{t \geq \kappa_b : h_1\mu_{12}G_1^t(q_1(t)) - \phi_{12} \leq h_2\mu_{22}G_2^t(q_2(t))\} \tag{A.23}$$

be the time this helping period ends. In addition, let

$$\tau_i' = \tau' + G_i^{\tau_d}(q_i^*(\tau_d))$$

be the subsequent time that class i, $i = 1, 2$, queue is emptied.

The optimal queue length trajectory follows:

$$q_1^*(t) = \begin{cases} q_1 + \int_0^t (\lambda_1(s) - s_1\mu_{11} - z_{12}^*(s)\mu_{12})\,ds, & t \in [0, \tau^*), \\ q_1^*(\tau^*) + \int_{\tau_*}^t (\lambda_1(s) - s_1\mu_{11})\,ds, & t \in [\tau^*, \tau_1^*), \\ 0, & t \in [\tau_1^*, \kappa_b), \\ \int_{\kappa_b}^t (\lambda_1(s) - s_1\mu_{11} - z_{12}^*(s)\mu_{12})\,ds, & t \in [\kappa_b, \tau'), \\ \int_{\tau'}^t (\lambda_1(s) - s_1\mu_{11})\,ds, & t \in [\tau', \tau_1'), \\ 0, & t \in [\tau_1', \infty), \end{cases}$$

$$q_2^*(t) = \begin{cases} q_2 + \int_0^t (\lambda_2(s) - z_{22}^*(s)\mu_{22})\,ds, & t \in [0, \tau^*), \\ q_2^*(\tau^*) + \int_{\tau^*}^t (\lambda_2(s) - s_2\mu_{22})\,ds, & t \in [\tau^*, \kappa_b), \\ q_2^*(\kappa_b) + \int_{\kappa_b}^t (\lambda_2(s) - z_{22}^*(s)\mu_{22})\,ds, & t \in [\kappa_b, \tau'), \\ q_2^*(\tau') + \int_{\tau_d}^t (\lambda_2(s) - s_2\mu_{22})\,ds, & t \in [\tau', \tau_2'), \\ 0, & t \in [\tau_2', \infty). \end{cases}$$

Note that it may be that $z_{12}^*(t) < s_2$ for $t \in [0, \tau^*)$ or $t \in [\kappa_b, \tau')$, if $q_1(t) = 0$ and $\lambda_1(t) < s_1\mu_{11} + s_2\mu_{12}$. In this case, $z_{22}^*(t) = s_2 - z_{12}^*(t)$ (since $q_2(t) > 0$ by assumption). However, it is always the case that $z_{11}^*(t) = s_1$ for $t \in [0, \tau^*]$.

Assume $\tau^* > 0$. We now partition the interval $[0, \tau^*)$ into subintervals $I_1, \cdots, I_n$ where $n \geq 1$, $I_i = [A_{i-1}, A_i)$ and $0 = A_0 < A_1 < \cdots < A_n = \tau^*$. The subintervals are defined such that in the interior of each subinterval, i.e., $t \in (A_{i-1}, A_i)$, either (i) $q_1(t) > 0$ and $q_2(t) > 0$, in which case we say that I_i is an interior subinterval, or (ii) $q_1(t) = 0$ and $q_2(t) > 0$, in which case we say that I_i is a boundary subinterval. Note that it is not possible that $q_1(t) > 0$ and $q_2(t) = 0$ for $t \in I_i$, because when $q_1(t) > 0$, $z_{22}^*(t) = 0$ and $\lambda_2(t) > 0$ during this time. The subintervals $I_1, \cdots, I_n$ do not necessarily alternate between interior and boundary subintervals: it is possible that I_k and I_{k+1} are both interior subintervals, with $q_1(t)$ hitting zero at the single point A_k.

Define the adjoint vector

$$p_2^*(t) = \begin{cases} h_2(\tau_2^* - t), & t \in [0, \tau_2'), \\ 0, & t \in [\tau_2', \infty). \end{cases}$$

We also define

$$p_1^*(t) = \begin{cases} h_1(\tau_1^* - t), & t \in [\tau^*, \tau_1^*), \\ 0, & t \in [\tau_1^*, \kappa_b). \end{cases}$$

With $p_1^*(A_n) = p_1^*(\tau^*)$ defined, we recursively define $p_1^*(t)$ for $t \in [0, A_n)$. We will do this in such a way that (i) the jumps of p_1^*, if any, occur only when $q_1^*(t) = 0$ and are positive; (ii) in interior subintervals I_i,

$$p_1^*(t)\mu_{12} - \phi_{12} - p_2^*(t)\mu_{22} \geq 0; \tag{A.24}$$

and (iii) in boundary subintervals I_i,

$$p_1^*(t)\mu_{12} - \phi_{12} - p_2^*(t)\mu_{22} = 0. \tag{A.25}$$

Note that this is done exactly as in the proof of Theorem 1.

Likewise, define

$$p_1^*(t) = \begin{cases} h_1(\tau_1' - t), & t \in [\tau', \tau_1'), \\ 0, & t \in [\tau_1', \kappa_b). \end{cases}$$

With $p_1^*(\tau')$ defined, we can again recursively define $p_1^*(t)$ for $t \in [\kappa_b, \tau')$ such that (i) the jumps of p_1^*, if any, occur only when $q_1^*(t) = 0$ and are positive; (ii) in interior subintervals I_i of $[\kappa_b, \tau_d)$,

$$p_1^*(t)\mu_{12} - \phi_{12} - p_2^*(t)\mu_{22} \geq 0; \tag{A.26}$$

and (iii) in boundary subintervals I_i of $[\kappa_b, \tau_d)$,

$$p_1^*(t)\mu_{12} - \phi_{12} - p_2^*(t)\mu_{22} = 0. \tag{A.27}$$

Note that while $p_2^*(t)$ decreases linearly to zero, $p_1^*(t)$ may not always be decreasing as it has a jump at time κ_b.

Define the multipliers

$$\eta_1^*(t) = \begin{cases} 0, & t \in I_k \text{ and } I_k \text{ is an interior subinterval,} \\ h_1 - \frac{h_2\mu_{22}}{\mu_{12}}, & t \in I_k \text{ and } I_k \text{ is a boundary subinterval,} \\ 0, & t \in [\tau^*, \tau_1^*) \bigcup [\tau_d, \tau_1'), \\ h_1, & t \in [\tau_1^*, \kappa_b) \bigcup [\tau_1', \infty), \end{cases}$$

$$\eta_2^*(t) = \begin{cases} 0, & t \in [0, \tau_2'), \\ h_2, & t \in [\tau_2', \infty) \end{cases}$$

$$\gamma_1^*(t) = \begin{cases} p_1^*(t)\mu_{11}, & t \in [0, \tau_1^*) \bigcup [\kappa_b, \tau_1'), \\ 0, & t \in [\tau_1^*, \kappa_b) \bigcup [\tau_1', \infty), \end{cases}$$

$$\gamma_2^*(t) = \begin{cases} p_1^*(t)\mu_{12} - \phi_{12}, & t \in [0, \tau^*) \bigcup [\kappa_b, \tau') \\ p_2^*(t)\mu_{22}, & t \in [\tau^*, \kappa_b) \bigcup [\tau', \tau_2'), \\ 0, & t \in [\tau_2', \infty) \end{cases}$$

$$\xi^*_{12}(t) = \begin{cases} 0, & t \in [0, \tau^*) \bigcup [\kappa_b, \tau'), \\ \phi_{12} - p^*_1(t)\mu_{12} + p^*_2(t)\mu_{22}, & t \in [\tau^*, \kappa_b) \bigcup [\tau', \infty), \end{cases}$$

$$\xi^*_{22}(t) = \begin{cases} p^*_1(t)\mu_{12} - \phi_{12} - p^*_2(t)\mu_{22}, & t \in [0, \tau^*) \bigcup [\kappa_b, \tau'), \\ 0, & t \in [\tau^*, \kappa_b) \bigcup [\tau', \infty) \end{cases}$$

and $\xi^*_{11}(t) = 0$ for all $t \geq 0$. Note that if $\tau^* > 0$, $p^*_1(t)\mu_{12} - \phi_{12} - p^*_2(t)\mu_{22} \geq 0$ for $t \in [0, \tau^*)$ by construction, and $p^*_1(t)\mu_{12} - \phi_{12} - p^*_2(t)\mu_{22} \leq 0$ for $t \in [\tau^*, \infty)$ since $p^*_1(\tau^*)\mu_{12} - \phi_{12} - p^*_2(\tau^*)\mu_{22} \leq 0$ (it is worth noting that strict inequality can occur if $q^*_1(t)$ hits zero exactly at time κ_b, since then $G^t_1(q^*_1(t))$ will jump at time τ^*_1) and $h_1\mu_{12} - h_2\mu_{22} \geq 0$. Thus, ξ^*_{12} and ξ^*_{22} are non-negative on $[0, \kappa_b)$. Similarly, they are non-negative on $[\kappa_b, \infty)$.

The conditions (ODE), (ADJ), (J), and (H) are easily verified. For (C), we only need to check that when $z^*_{22}(t) > 0$ in boundary subintervals $t \in [A_{k-1}, A_k)$, $\xi^*_{22}(t) = 0$. This holds because of (A.25) and (A.27). We now verify (T). Note that

$$\nabla_{z_{11}} L(q^*(t), z^*(t), p^*(t), \eta^*(t), \gamma^*(t), \xi^*(t)) = -p^*_1(t)\mu_{11} + \gamma^*_1(t) - \xi^*_{11}(t) = 0$$

because $\xi^*_{11}(t) = 0$ and $\gamma^*_1(t) = p^*_1(t)\mu_{11}$. Next,

$$\nabla_{z_{22}} L(q^*(t), z^*(t), p^*(t), \eta^*(t), \gamma^*(t), \xi^*(t)) = -p^*_2(t)\mu_{22} + \gamma^*_2(t) - \xi^*_{22}(t) = 0$$

because $\xi^*_{22}(t) = \gamma^*_2(t) - p^*_2(t)\mu_{22}$ for $t \in [0, \tau^*) \bigcup [\kappa_b, \tau_d)$, and $\xi^*_{22}(t) = 0$ and $\gamma^*_2(t) = p^*_2(t)\mu_{22}$ otherwise. Finally,

$$\nabla_{z_{12}} L(q^*(t), z^*(t), p^*(t), \eta^*(t), \gamma^*(t), \xi^*(t)) = \phi_{12} - p^*_1(t)\mu_{12} + \gamma^*_2(t) - \xi^*_{12}(t) = 0$$

because $\xi^*_{12}(t) = 0$ and $\gamma^*_2(t) = p^*_1(t)\mu_{12} - \phi_{12}$ for $t \in [0, \tau^*) \bigcup [\kappa_b, \tau_d)$, and $\xi^*_{12}(t) = \phi_{12} -$

$p_1^*(t)\mu_{12} + \gamma_2^*(t)$ otherwise.

It remains to verify (M). It is clear that $z_{11}^*(t)$ should always be maximal. (This is slightly less clear if $q_1^*(t) = 0$ for some $t < \tau^*$, since there is a constraint $z_{11}^*(t)\mu_{11} + z_{12}^*(t)\mu_{12} \leq \lambda_1(t)$ when $q_1^*(t) = 0$. In this case, (M) follows because the coefficients of $z_{11}^*(t)$ and $z_{12}^*(t)$ are $-p_1^*(t)\mu_{11}$ and $\phi_{12} - p_1^*(t)\mu_{12}$.) For $t \in [0, \tau^*) \bigcup [\kappa_b, \tau_d)$, the coefficients of $z_{12}^*(t)$ and $z_{22}^*(t)$ are respectively $\phi_{12} - p_1^*(t)\mu_{12}$ and $-p_2^*(t)\mu_{22}$. Since $p_1^*(t)\mu_{12} - \phi_{12} \geq p_2^*(t)\mu_{22}$, it is optimal to have $z_{12}^*(t)$ maximal. For other t, the reverse inequality is true, and so it is optimal to have $z_{22}^*(t)$ maximal. This in turn implies that $z_{12}^*(t) = 0$ for $t < \tau_2'$ is optimal (pool 2 has no spare capacity to help class 1). When $t \geq \tau_2'$, $\phi_{12} - p_1^*(t)\mu_{12} \leq 0$, and again $z_{12}^*(t) = 0$ is optimal.

Case II: $h_1\mu_{12} \leq h_2\mu_{22}$. The proof is similar to that of Theorem 1 and we provide a roadmap here only. In this case, pool 2 will serve only its own class until the class 2 queue is emptied. Thereafter, it may provide partial help to class 1 for up to two different intervals, one for each demand surge period of class 1.

If $q_1^*(\kappa_b) > 0$, the two demand surges for class 1 behave as a single demand surge, and the proof of Theorem 1 applies directly. If $G_2^0(q_2(0)) \geq \kappa_b$, there is at most one demand surge for class 1 after pool 2 is ready to provide partial help. The proof of Theorem 1 again applies directly. Suppose instead $q_1^*(\kappa_b) = 0$ and $G_2^0(q_2(0)) < \kappa_b$. In this case, we can apply the proof of Theorem 1 separately to each of the two intervals $[0, \kappa_b)$ and $[\kappa_b, \infty)$. Noting that in this case, $q_1^*(\kappa_b) = q_2^*(\kappa_b) = 0$. □

A.7 Optimal control for the X-Model

Proof. Proof of Theorem 4. To prove that the policy characterized in Theorem 4 is optimal, we shall construct the optimal primal and dual trajectories and show that the conditions in Theorem 11 are satisfied.

Let $q_1(0) = q_1$ and $q_2(0) = q_2$. For the X-model, the Hamiltonian takes the form:

$$H(q(t), z(t), p(t), t) = \sum_i h_i q_i(t) + \phi_{12} z_{12}(t) + \phi_{21} z_{21}(t) + \sum_i p_i(t) \left(\lambda_i(t) - \sum_j \mu_{ij} z_{ij}(t) \right).$$

The augmented Hamiltonian takes the form:

$$\begin{aligned} & L(q(t), z(t), p(t), \eta(t), \gamma(t), \xi(t), t) \\ = \; & H(q(t), z(t), p(t), t) - \sum_i \eta_i(t) q_i(t) + \gamma_1(t)(z_{11}(t) + z_{21}(t) - s_1) \\ & + \gamma_2(t)(z_{12}(t) + z_{22}(t) - s_2) - \sum_{i,j} \xi_{ij}(t) z_{ij}(t). \end{aligned}$$

Consider first the case $h_1\mu_{12} > h_2\mu_{22}$. We further consider two sub-cases, depending on whether pool 2 initially prioritizes class 1, i.e., whether (2.11) holds at $t = 0$.

Case I: Pool 2 does not initially prioritize class 1, i.e., (2.11) does not hold at $t = 0$. In this case, the policy is that each pool serves only its own class for $t < \tau_1 := G_1^0(q_1(0))$. Then, pool 1 gives partial help to class 2 for $t \in [\tau_1, \tau_1 + \tau^*)$, where

$$\tau^* = \inf\{t \geq 0 : h_2\mu_{21} G_2^{\tau_1 + t}(q_2(\tau_1 + t)) \leq \phi_{21}\}.$$

At all subsequent times, each pool serves only its own class again. In what follows, intervals of the form $[a, b)$ for $b \leq a$ are empty.

Let τ_2^* denote the first time at which queue 2 empties. That is, $\tau_2^* = \tau_2 := G_2^0(q_2(0))$ if $\tau_2 \leq \tau_1$, and $\tau_2^* = \tau_1 + \tau^* + G_2^{\tau_1 + \tau^*}(q_2^*(\tau_1 + \tau^*))$ if $\tau_2 > \tau_1$. Note that if $\tau^* > 0$, then $h_2\mu_{21} G_2^{\tau_1 + \tau^*}(q_2^*(\tau_1 + \tau^*)) = \phi_{21}$ by continuity, so that $\tau_2^* = \tau_1 + \tau^* + \frac{\phi_{21}}{h_2\mu_{21}}$.

The optimal queue length trajectory follows:

$$q_1^*(t) = \begin{cases} q_1 + \int_0^t (\lambda_1(s) - s_1\mu_{11})\,ds, & t \in [0, \tau_1), \\ 0, & t \in [\tau_1, \infty), \end{cases}$$

$$q_2^*(t) = \begin{cases} q_2 + \int_0^t (\lambda_2(s) - s_2\mu_{22})\,ds, & t \in [0, \tau_1 \wedge \tau_2^*), \\ q_2^*(\tau_1) + \int_{\tau_1}^t (\lambda_2(s) - s_2\mu_{22} - (s_1 - \lambda_1(s)/\mu_{11})\mu_{21})\,ds, & t \in [\tau_1, \tau_1 + \tau^*), \\ q_2^*(\tau_1 + \tau^*) + \int_{\tau_1+\tau^*}^t (\lambda_2(s) - s_2\mu_{22})\,ds, & t \in [\tau_1 + \tau^*, \tau_2^*) \\ 0, & t \in [\tau_2^*, \infty). \end{cases}$$

Define the adjoint vectors

$$p_1^*(t) = \begin{cases} h_1(\tau_1 - t) + h_2\frac{\mu_{21}}{\mu_{11}}\tau^*, & t \in [0, \tau_1), \\ h_2\frac{\mu_{21}}{\mu_{11}}(\tau_1 + \tau^* - t), & t \in [\tau_1, \tau_1 + \tau^*), \\ 0, & t \in [\tau_1 + \tau^*, \infty), \end{cases}$$

$$p_2^*(t) = \begin{cases} h_2(\tau_2^* - t), & t \in [0, \tau_2^*), \\ 0, & t \in [\tau_2^*, \infty). \end{cases}$$

Define the Lagrangian multipliers

$$\eta_1^*(t) = \begin{cases} 0, & t \in [0, \tau_1), \\ h_1 - h_2\frac{\mu_{21}}{\mu_{11}}, & t \in [\tau_1, \tau_1 + \tau^*), \\ h_1, & t \in [\tau_1 + \tau^*, \infty), \end{cases}$$

$$\eta_2^*(t) = \begin{cases} 0, & t \in [0, \tau_2^*), \\ h_2, & t \in [\tau_2^*, \infty). \end{cases}$$

$$\gamma_1^*(t) = \begin{cases} p_1^*(t)\mu_{11}, & t \in [0, \tau_1 + \tau^*), \\ 0, & t \in [\tau_1 + \tau^*, \infty), \end{cases}$$

$$\gamma_2^*(t) = \begin{cases} p_2^*(t)\mu_{22}, & t \in [0, \tau_2^*), \\ 0, & t \in [\tau_2^*, \infty). \end{cases}$$

$$\xi_{12}^*(t) = \begin{cases} \phi_{12} - p_1^*(t)\mu_{12} + p_2^*(t)\mu_{22}, & t \in [0, \tau_1 + \tau^*), \\ \phi_{12} + p_2^*(t)\mu_{22}, & t \in [\tau_1 + \tau^*, \infty), \end{cases}$$

$$\xi_{21}^*(t) = \begin{cases} \phi_{21} - p_2^*(t)\mu_{21} + p_1^*(t)\mu_{11}, & t \in [0, \tau_1), \\ 0, & t \in [\tau_1, \tau_1 + \tau^*), \\ \phi_{21} - p_2^*(t)\mu_{21}, & t \in [\tau_1 + \tau^*, \infty), \end{cases}$$

and $\xi_{11}^*(t) = \xi_{22}^*(t) = 0$ for all $t \geq 0$. Note that $\eta_1^*(t) \geq 0$ because $h_1\mu_{11} \geq h_2\mu_{21}$ by assumption.

To see that $\xi_{21}^*(t) \geq 0$, note that because $h_2\mu_{21} \leq h_1\mu_{11}$, $\xi_{21}^*(t)$ is non-increasing on $[0, \tau_1)$. Moreover, if $\tau^* > 0$, it approaches zero as $t \to \tau_1$ (because $h_2\mu_{21}(\tau_2^* - \tau^* - \tau_1) = \phi_{21}$), while if $\tau^* = 0$, it approaches $\phi_{21} - h_2\mu_{21}G_2^{\tau_1}(q_2(\tau_1)) \geq 0$ instead, even if $\tau_2^* \leq \tau_1$.

To see that $\xi_{12}^*(t) \geq 0$, note that it is non-decreasing on $[0, \tau_1)$ (because $h_1\mu_{12} \geq h_2\mu_{22}$), it is monotone on $[\tau_1, \tau_1 + \tau^*)$, and it attains the value $\phi_{12} + p_2^*(\tau_1 + \tau^*)\mu_{22} \geq 0$ at $\tau_1 + \tau^*$. Thus, it suffices to check that $\xi_{12}^*(0) \geq 0$. Meanwhile $\xi_{12}^*(0) \geq 0$ is equivalent to (2.11) is violated, which is assumed in this case.

The conditions (ODE), (ADJ), (C), (J), and (H) can be straightforwardly verified by construction.

We now verify (T). We have for $i = 1, 2$ that

$$\nabla_{z_{ii}} L(q^*(t), z^*(t), p^*(t), \eta^*(t), \gamma^*(t), \xi^*(t), t) = -p_i^*(t)\mu_{ii} + \gamma_i^*(t) - \xi_{ii}^*(t) = 0$$

because $\xi_{ii}^*(t) = 0$ and $\gamma_i^*(t) = p_i^*(t)\mu_{ii}$. Next,

$$\nabla_{z_{12}} L(q^*(t), z^*(t), p^*(t), \eta^*(t), \gamma^*(t), \xi^*(t), t) = \phi_{12} - p_1^*(t)\mu_{12} + \gamma_2^*(t) - \xi_{12}^*(t) = 0$$

because $\xi_{12}^*(t) = \phi_{12} - p_1^*(t)\mu_{12} + p_2^*(t)\mu_{22}$. Finally,

$$\nabla_{z_{21}} L(q^*(t), z^*(t), p^*(t), \eta^*(t), \gamma^*(t), \xi^*(t), t) = \phi_{21} - p_2^*(t)\mu_{21} + \gamma_1^*(t) - \xi_{21}^*(t).$$

For $t \geq \tau_1 + \tau^*$, $\phi_{21} - p_2^*(t)\mu_{21} + \gamma_1^*(t) - \xi_{21}^*(t) = 0$ because the $\gamma_1^*(t) = 0$ and $\xi_{21}^*(t) = \phi_{21} - p_2^*(t)\mu_{21}$. For $t \in [\tau_1, \tau_1 + \tau^*)$, we get

$$\phi_{21} - p_2^*(t)\mu_{21} + p_1^*(t)\mu_{11} = \phi_{21} - h_2\mu_{21}\left((\tau_2^* - t) - (\tau_1 + \tau^* - t)\right) = \phi_{21} - h_2\mu_{21} G_2^{\tau_1+\tau^*}(q_2^*(\tau_1 + \tau^*))$$

which is zero. Finally, for $t \in [0, \tau_1)$, $\phi_{21} - p_2^*(t)\mu_{21} + \gamma_1^*(t) - \xi_{21}^*(t) = 0$ because $\xi_{21}^*(t) = \phi_{21} - p_2^*(t)\mu_{21} + \gamma_1^*(t)$.

It remains to verify (M). The coefficients of $z_{11}^*(t)$ and $z_{21}^*(t)$ are respectively $-p_1^*(t)\mu_{11}$ and $\phi_{21} - p_2^*(t)\mu_{21}$. Note that $\phi_{21} - p_2^*(t)\mu_{21} + p_1^*(t)\mu_{11} = \xi_{21}^*(t) \geq 0$ for all t, so the Hamiltonian is minimized by setting $z_{11}^*(t)$ maximal, i.e. pool 1 prioritizing class 1. For $t < \tau_1$, $G_1^t(q_1(t)) > 0$, so pool 1 has no capacity to help class 2, i.e. $z_{21}^*(t) = 0$. For $t \in [\tau_1, \tau_1 + \tau^*)$, $\phi_{21} - p_2^*(t)\mu_{21} = 0$, so it is Hamiltonian-minimal (i.e. minimizes the Hamiltonian) for pool 1 to partially help class 2 (not helping is also Hamiltonian-minimal), i.e. $z_{21}^*(t) = N_1 - z_{11}^*(t)$. Finally, for $t \geq \tau_1 + \tau^*$, $\phi_{21} - p_2^*(t)\mu_{21} \geq 0$, and so it is Hamiltonian-minimal for pool 2 to serve only its own class.

Next, the coefficients of $z_{12}^*(t)$ and $z_{22}^*(t)$ are respectively $\phi_{12} - p_1^*(t)\mu_{12}$ and $-p_2^*(t)\mu_{22}$. Note that $\phi_{12} - p_1^*(t)\mu_{12} + p_2^*(t)\mu_{22} = \xi_{12}^*(t) \geq 0$, for all t, so the Hamiltonian is minimized by setting

$z^*_{22}(t)$ maximal, i.e., pool 2 prioritizing class 2. Thus, for $t < \tau^*_2$, it is Hamiltonian-minimal to have $z^*_{22}(t) = N_2$ and $z^*_{12}(t) = 0$. If $\tau^*_2 < \tau_1$ and $t \geq \tau^*_2$, $\phi_{12} - p^*_1(t)\mu_{12} = \xi^*_{12}(t) \geq 0$, and so it is again Hamiltonian-minimal to have $z^*_{12}(t) = 0$. This completes the proof for case I.

Case II: Pool 2 initially prioritizes class 1, i.e., (2.11) holds at $t = 0$. Let $T_1 > 0$ be the length of time pool 2 initially prioritizes class 1. By continuity, equality holds for (2.11) at $t = T_1$. In the next period $[T_1, T_2)$, each pool serves its own primary class until queue 1 empties at time T_2. Next, in $[T_2, T_3)$, pool 1 partially helps class 2, i.e. $z^*_{21}(t) = s_1 - z^*_{11}(t)$. Finally, for all remaining time $t \geq T_3$, each pool again serves only its own primary class. Here, $0 < T_1 \leq T_2 \leq T_3$, with $T_2 = T_3$ if $G_2^{T_2}(q^*_2(T_2)) \leq \frac{\phi_{21}}{h_2\mu_{21}}$. Also, $T_1 = T_2$ is only possible if $\phi_{12} = 0$.

Let T_4 be the time other than zero that queue 2 empties after its demand surge ends, i.e. $T_4 = \inf\{t > 0 : G_2^t(q^*_2(t)) = 0\}$. It is possible that $T_4 \leq T_2$ or $T_4 > T_2$. If $T_4 > T_2$, then $T_4 = T_3 + G_2^{T_3}(q^*_2(T_3))$. Note that the restriction that $t > 0$ is necessary because it is possible that $G_2^0(q_2(0)) = 0$ if $q_2(0) = 0$ and $\kappa_2 = 0$.

The optimal queue length trajectory follows:

$$q^*_1(t) = \begin{cases} q_1 + \int_0^t (\lambda_1(s) - s_1\mu_{11} - z^*_{12}(s)\mu_{12})\, ds, & t \in [0, T_1), \\ q^*_1(\tau_1) + \int_{\tau_1}^t (\lambda_1(s) - s_1\mu_{11})\, ds, & t \in [T_1, T_2), \\ 0, & t \in [T_2, \infty), \end{cases}$$

$$q^*_2(t) = \begin{cases} q_2 + \int_0^t (\lambda_2(s) - z^*_{22}(s)\mu_{22})\, ds, & t \in [0, T_1), \\ q^*_2(\tau_1) + \int_{\tau_1}^t (\lambda_2(s) - s_2\mu_{22})\, ds, & t \in [T_1, T_2 \wedge T_4), \\ q^*_2(\tau_2) + \int_{\tau_2}^t (\lambda_2(s) - s_2\mu_{22} - (s_1 - \lambda_1(s)/\mu_{11})\mu_{21})\, ds, & t \in [T_2, T_3), \\ q^*_2(T_3) + \int_{T_3}^t (\lambda_2(s) - s_2\mu_{22})\, ds, & t \in [T_3, T_4), \\ 0, & t \in [T_4, \infty). \end{cases}$$

Note that it is possible that $z^*_{12}(t) < s_2$ and $z^*_{22}(t) = s_2 - z^*_{12}(t) > 0$ for $t \in [0, T_1)$ when pool 2 prioritizes class 1, because it may be that $q^*_1(t) = 0$ and $\lambda_1(t) < s_1\mu_{11} + s_2\mu_{12}$ but $G_1^t(q^*_1(t)) > 0$.

Define the adjoint vector

$$p_2^*(t) = \begin{cases} h_2(T_4 - t), & t \in [0, T_4), \\ 0, & t \in [T_4, \infty). \end{cases}$$

We also define

$$p_1^*(t) = \begin{cases} h_1(T_2 - t) + h_2 \frac{\mu_{21}}{\mu_{11}}(T_3 - T_2), & t \in [T_1, T_2), \\ h_2 \frac{\mu_{21}}{\mu_{11}}(T_3 - t), & t \in [T_2, T_3), \\ 0, & t \in [T_3, \infty). \end{cases}$$

We also need to define $p_1^*(t)$ for $t \in [0, T_1)$. We partition the interval $[0, T_1)$ into subintervals $I_1, \cdots, I_n$ where $n \geq 1$, $I_i = [A_{i-1}, A_i)$ and $0 = A_0 < A_1 < \cdots < A_n = T_1$, as follows. In the interior $t \in (A_{i-1}, A_i)$ of each subinterval, either (i) $q_1^*(t) > 0$ and $q_2^*(t) > 0$, in which case we say that I_i is an interior subinterval, or (ii) $q_1^*(t) = 0$ and $q_2^*(t) > 0$, in which case we say that I_i is a boundary subinterval. Note that it is not possible that $q_1^*(t) > 0$ and $q_2^*(t) = 0$ in some subinterval, because $z_{22}^*(t) = 0$ during this time and $\lambda_2(t) > 0$. Also, Assumption 5 rules out the case $q_1^*(t) = q_2^*(t) = 0$ (such a subinterval cannot occur after $\kappa_1 \vee \kappa_2$, because then $G_i^t(q_i(t)) = 0$ for $i = 1, 2$ and (2.11) cannot hold).

The subintervals $I_1, \cdots, I_n$ do not necessarily alternate between interior and boundary subintervals: it is possible that I_k and I_{k+1} are both interior subintervals, with $q_1(t)$ hitting zero at the single point A_k. The fact that there are finitely many such subintervals follows from piecewise monotonicity in Assumption 5, because the class 1 queue length can only leave zero once during each monotone period.

With $p_1^*(A_n) = p_1^*(T_1)$ defined, we recursively define $p_1^*(t)$ for $t \in [0, A_n)$. We will do this in such a way that (i) the jumps of p_1^*, if any, occur only when $q_1^*(t) = 0$ and are positive; (ii) in interior subintervals I_i,

$$p_1^*(t)\mu_{12} - \phi_{12} - p_2^*(t)\mu_{22} \geq 0; \tag{A.28}$$

and (iii) in boundary subintervals I_i,

$$p_1^*(t)\mu_{12} - \phi_{12} - p_2^*(t)\mu_{22} = 0. \tag{A.29}$$

Note that

$$p_1^*(T_1)\mu_{12} - \phi_{12} - p_2^*(T_1)\mu_{22} = 0. \tag{A.30}$$

Indeed, this statement is equivalent to equality for (2.11) at $t = T_1$, which follows from continuity.

Suppose $p_1^*(A_k)$ has been defined for some k, with $p_1^*(A_k)\mu_{12} - \phi_{12} - p_2^*(A_k)\mu_{22} \geq 0$. If I_k is an interior subinterval, we set

$$p_1^*(t) = h_1(A_k - t) + p_1^*(A_k)$$

for $t \in [A_{k-1}, A_k)$. That is, p_1^* is continuous at A_k and has slope $-h_1$ in the subinterval I_k. Thus, $p_1^*(t)\mu_{12} - \phi_{12} - p_2^*(t)\mu_{22}$ has slope $h_2\mu_{22} - h_1\mu_{12} \leq 0$, which implies that $p_1^*(A_{k-1})\mu_{12} - \phi_{12} - p_2^*(A_{k-1})\mu_{22} \geq 0$.

Suppose instead I_k is a boundary subinterval. We set $p_1^*(A_{k-1}) = p_2^*(A_{k-1})\mu_{22}/\mu_{12} + \phi_{12}/\mu_{12}$ and $p_1^*(t) = p_1^*(A_{k-1}) - \frac{h_2\mu_{22}}{\mu_{12}}(t - A_{k-1})$ for $t \in (A_{k-1}, A_k)$. That is, p_1^* has a jump at A_k and has slope $-\frac{h_2\mu_{22}}{\mu_{12}}$ in the subinterval I_k. This ensures that $\phi_{12} - p_1^*(t)\mu_{12} = -p_2^*(t)\mu_{22}$ everywhere in I_k. The size of the jump at A_k is $p_1^*(A_k) - p_2^*(A_k)\mu_{22}/\mu_{12} - \phi_{12}/\mu_{12} \geq 0$, which is non-negative because $p_1^*(A_k)\mu_{12} - \phi_{12} - p_2^*(A_k)\mu_{22} \geq 0$. Thus, we have defined p_1^* for $t \in [0, T_1)$ satisfying conditions (i), (ii) and (iii).

Define the multipliers

$$\eta_1^*(t) = \begin{cases} 0, & t \in I_k \text{ and } I_k \text{ is an interior subinterval,} \\ h_1 - h_2\frac{\mu_{22}}{\mu_{12}}, & t \in I_k \text{ and } I_k \text{ is a boundary subinterval,} \\ 0, & t \in [T_1, T_2), \\ h_1 - h_2\frac{\mu_{21}}{\mu_{11}}, & t \in [T_2, T_3), \\ h_1, & t \in [T_3, \infty), \end{cases}$$

$$\eta_2^*(t) = \begin{cases} 0, & t \in [0, T_4), \\ h_2, & t \in [T_4, \infty). \end{cases}$$

Note that $h_1\mu_{12} \geq h_2\mu_{22}$ and $h_1\mu_{11} \geq h_2\mu_{21}$, so that $\eta_1^*(t) \geq 0$. Define also

$$\gamma_1^*(t) = \begin{cases} p_1^*(t)\mu_{11}, & t \in [0, T_3), \\ 0, & t \in [T_3, \infty), \end{cases}$$

$$\gamma_2^*(t) = \begin{cases} p_1^*(t)\mu_{12} - \phi_{12}, & t \in [0, T_1) \\ p_2^*(t)\mu_{22}, & t \in [T_1, T_4), \\ 0, & t \in [T_4, \infty). \end{cases}$$

$$
\begin{aligned}
\xi_{12}^*(t) &= \begin{cases} 0, & t \in [0, T_1), \\ \phi_{12} - p_1^*(t)\mu_{12} + p_2^*(t)\mu_{22}, & t \in [T_1, \infty), \end{cases} \\
\xi_{21}^*(t) &= \begin{cases} \phi_{21} - p_2^*(t)\mu_{21} + p_1^*(t)\mu_{11}, & t \in [0, T_2), \\ 0, & t \in [T_2, T_3), \\ \phi_{21} - p_2^*(t)\mu_{21}, & t \in [T_3, \infty). \end{cases} \\
\xi_{22}^*(t) &= \begin{cases} p_1^*(t)\mu_{12} - \phi_{12} - p_2^*(t)\mu_{22}, & t \in [0, T_1), \\ 0, & t \in [T_1, \infty), \end{cases}
\end{aligned}
$$

and $\xi_{11}^*(t) = 0$ for all $t \geq 0$. Note that $p_1^*(t)\mu_{12} - \phi_{12} - p_2^*(t)\mu_{22} \geq 0$ for $t \in [0, T_1)$ by construction. Also, $\phi_{12} - p_1^*(T_1)\mu_{12} + p_2^*(T_1)\mu_{22} = 0$ from (A.30). Since $h_1\mu_{12} - h_2\mu_{22} \geq 0$, $\phi_{12} - p_1^*(t)\mu_{12} + p_2^*(t)\mu_{22}$ is non-decreasing for $t \in [T_1, T_3)$, after which point it equals $\phi_{12} + p_2^*(t)\mu_{22} \geq 0$. Thus, ξ_{12}^* and ξ_{22}^* are non-negative.

To see that $\xi_{21}^*(t) \geq 0$, note that because $h_2\mu_{21} \leq h_1\mu_{11}$, $\xi_{21}^*(t)$ is non-increasing on $[0, T_2)$ (its slope is at most $h_2\mu_{21} - h_1\mu_{11}$ for $t \in [0, T_1)$). If $T_3 > T_2$, $\xi_{21}^*(T_2-) = 0$ and $\xi_{21}^*(t)$ is non-decreasing from its value of zero for $t \in [T_3, \infty)$, so that $\xi_{21}^*(t) \geq 0$ everywhere. If instead $T_3 = T_2$, we have $\xi_{21}^*(T_2-) \geq 0$ instead, and the same result holds.

The conditions (ODE), (ADJ), (J), and (H) can be straightforwardly verified by our construction. For (C), we only need to check that when $z_{22}^*(t) > 0$ in boundary subintervals $t \in [A_{k-1}, A_k)$, $\xi_{22}^*(t) = 0$. This holds because of (A.29). (Note that $z_{22}^*(A_k) = 0$.)

We now verify (T). We have that

$$
\nabla_{z_{11}} L(q^*(t), z^*(t), p^*(t), \eta^*(t), \gamma^*(t), \xi^*(t), t) = -p_1^*(t)\mu_{11} + \gamma_1^*(t) - \xi_{11}^*(t) = 0
$$

because $\xi_{11}^*(t) = 0$ and $\gamma_1^*(t) = p_1^*(t)\mu_{11}$. Next,

$$
\nabla_{z_{22}} L(q^*(t), z^*(t), p^*(t), \eta^*(t), \gamma^*(t), \xi^*(t), t) = -p_2^*(t)\mu_{22} + \gamma_2^*(t) - \xi_{22}^*(t) = 0
$$

because $\xi^*_{22}(t) = \gamma^*_2(t) - p^*_2(t)\mu_{22}$ for $t \in [0, T_1)$, and $\xi^*_{22}(t) = 0$ and $\gamma^*_2(t) = p^*_2(t)\mu_{22}$ for $t \geq T_1$. Next,

$$\nabla_{z_{12}} L(q^*(t), z^*(t), p^*(t), \eta^*(t), \gamma^*(t), \xi^*(t), t) = \phi_{12} - p^*_1(t)\mu_{12} + \gamma^*_2(t) - \xi^*_{12}(t) = 0$$

because $\xi^*_{12}(t) = 0$ and $\gamma^*_2(t) = p^*_1(t)\mu_{12} - \phi_{12}$ for $t \in [0, T_1)$, and $\xi^*_{12}(t) = \phi_{12} - p^*_1(t)\mu_{12} + \gamma^*_2(t)$ for $t \geq T_1$. Finally,

$$\nabla_{z_{21}} L(q^*(t), z^*(t), p^*(t), \eta^*(t), \gamma^*(t), \xi^*(t), t) = \phi_{21} - p^*_2(t)\mu_{21} + \gamma^*_1(t) - \xi^*_{21}(t) = 0$$

because $\xi^*_{21}(t) = \phi_{21} - p^*_2(t)\mu_{21} + \gamma^*_1(t)$ for all $t \geq 0$.

It remains to verify (M). The coefficients of $z^*_{11}(t)$ and $z^*_{21}(t)$ are respectively $-p^*_1(t)\mu_{11}$ and $\phi_{21} - p^*_2(t)\mu_{21}$. Note that $\phi_{21} - p^*_2(t)\mu_{21} + p^*_1(t)\mu_{11} = \xi^*_{21}(t) \geq 0$ for all t, so the Hamiltonian is minimized by setting $z^*_{11}(t)$ maximal, i.e. pool 1 prioritizing class 1. For $t < T_2$, $G^t_1(q_1(t)) > 0$, so pool 1 has no capacity to help class 2, i.e. $z^*_{21}(t) = 0$. For $t \in [T_2, T_3)$, $\phi_{21} - p^*_2(t)\mu_{21} = 0$, so it is Hamiltonian-minimal for pool 1 to partially help class 2 (not helping is also Hamiltonian-minimal), i.e. $z^*_{21}(t) = s_1 - z^*_{11}(t)$. Finally, for $t \geq T_3$, $\phi_{21} - p^*_2(t)\mu_{21} \geq 0$, and so it is Hamiltonian-minimal for pool 2 to serve only its own class.

Next, the coefficients of $z^*_{12}(t)$ and $z^*_{22}(t)$ are respectively $\phi_{12} - p^*_1(t)\mu_{12}$ and $-p^*_2(t)\mu_{22}$. For $t < T_1$, $\phi_{12} - p^*_1(t)\mu_{12} + p^*_2(t)\mu_{22} = -\xi^*_{22}(t) \leq 0$, so it is Hamiltonian-minimal to have $z^*_{12}(t)$ maximal, i.e. pool 2 prioritizing class 1. Since $p^*_2(t) \geq 0$, it is also Hamiltonian-minimal to have any remaining pool 2 servers serve its own class, i.e. $z^*_{22}(t) = s_2 - z^*_{12}(t)$. For $t \geq T_1$, $\phi_{12} - p^*_1(t)\mu_{12} + p^*_2(t)\mu_{22} = \xi^*_{12}(t) \geq 0$, so it is Hamiltonian-minimal to have $z^*_{22}(t)$ maximal, i.e. pool 2 prioritizing class 2. Thus, for $t < T_4$, it is Hamiltonian-minimal to have $z^*_{22}(t) = s_2$ and $z^*_{12}(t) = 0$. If $T_4 < T_2$ and $t \geq T_2$, $\phi_{12} - p^*_1(t)\mu_{12} = \xi^*_{12}(t) \geq 0$, and so it is again Hamiltonian-minimal to have $z^*_{12}(t) = 0$. This completes the proof.

Consider next the other case $h_1\mu_{12} < h_2\mu_{22}$ and $h_2\mu_{21} < h_1\mu_{11}$. Let $\tau_i = G^0_i(q_i(0))$ for $i = 1, 2$ be the time for each queue to empty using its own pool. By symmetry, we may assume

without loss of generality that $\tau_1 \leq \tau_2$. Thus, the trajectory under the stated policy is as follows. First, in $[0, \tau_1]$, each pool serves only its own class until queue 1 empties. Let

$$\tau^* = \inf \left\{ t \geq 0 : G_2^{\tau_1+t}(q_2(\tau_1 + t)) \leq \frac{\phi_{21}}{h_2 \mu_{21}} \right\}.$$

Then, pool 1 will partially help class 2 for $t \in [\tau_1, \tau_1 + \tau^*)$, after which helping stops and both pools again serve only their own class, until queue 2 is also emptied.

Let $\tau_2^* = \tau_1 + \tau^* + G_2^{\tau_1+\tau^*}(q_2^*(\tau_1 + \tau^*))$ be the time until queue 2 empties. Note that if $\tau^* > 0$, then $h_2 \mu_{21} G_2^{\tau_1+\tau^*}(q_2^*(\tau_1 + \tau^*)) = \phi_{21}$ by continuity, so that $\tau_2^* = \tau_1 + \tau^* + \frac{\phi_{21}}{h_2 \mu_{21}}$.

The optimal queue length trajectory follows:

$$q_1^*(t) = \begin{cases} q_1 + \int_0^t (\lambda_1(s) - s_1 \mu_{11})\, ds, & t \in [0, \tau_1), \\ 0, & t \in [\tau_1, \infty), \end{cases}$$

$$q_2^*(t) = \begin{cases} q_2 + \int_0^t (\lambda_2(s) - s_2 \mu_{22})\, ds, & t \in [0, \tau_1), \\ q_2^*(\tau_1) + \int_{\tau_1}^t (\lambda_2(s) - s_2 \mu_{22} - (s_1 - \lambda_1(s)/\mu_{11})\mu_{21})\, ds, & t \in [\tau_1, \tau_1 + \tau^*), \\ q_2^*(\tau_1 + \tau^*) + \int_{\tau_1+\tau^*}^t (\lambda_2(s) - s_2 \mu_{22})\, ds, & t \in [\tau_1 + \tau^*, \tau_2^*) \\ 0, & t \in [\tau_2^*, \infty). \end{cases}$$

Define the adjoint vectors

$$p_1^*(t) = \begin{cases} h_1(\tau_1 - t) + h_2 \frac{\mu_{21}}{\mu_{11}} \tau^*, & t \in [0, \tau_1), \\ h_2 \frac{\mu_{21}}{\mu_{11}} (\tau_1 + \tau^* - t), & t \in [\tau_1, \tau_1 + \tau^*), \\ 0, & t \in [\tau_1 + \tau^*, \infty), \end{cases}$$

$$p_2^*(t) = \begin{cases} h_2(\tau_2^* - t), & t \in [0, \tau_2^*), \\ 0, & t \in [\tau_2^*, \infty). \end{cases}$$

Define the Lagrangian multipliers

$$\eta_1^*(t) = \begin{cases} 0, & t \in [0, \tau_1), \\ h_1 - h_2 \frac{\mu_{21}}{\mu_{11}}, & t \in [\tau_1, \tau_1 + \tau^*), \\ h_1, & t \in [\tau_1 + \tau^*, \infty), \end{cases}$$

$$\eta_2^*(t) = \begin{cases} 0, & t \in [0, \tau_2^*), \\ h_2, & t \in [\tau_2^*, \infty). \end{cases}$$

$$\gamma_1^*(t) = \begin{cases} p_1^*(t)\mu_{11}, & t \in [0, \tau_1 + \tau^*), \\ 0, & t \in [\tau_1 + \tau^*, \infty), \end{cases}$$

$$\gamma_2^*(t) = \begin{cases} p_2^*(t)\mu_{22}, & t \in [0, \tau_2^*), \\ 0, & t \in [\tau_2^*, \infty). \end{cases}$$

$$\xi_{12}^*(t) = \begin{cases} \phi_{12} - p_1^*(t)\mu_{12} + p_2^*(t)\mu_{22}, & t \in [0, \tau_1 + \tau^*), \\ \phi_{12} + p_2^*(t)\mu_{22}, & t \in [\tau_1 + \tau^*, \infty), \end{cases}$$

$$\xi_{21}^*(t) = \begin{cases} \phi_{21} - p_2^*(t)\mu_{21} + p_1^*(t)\mu_{11}, & t \in [0, \tau_1), \\ 0, & t \in [\tau_1, \tau_1 + \tau^*), \\ \phi_{21} - p_2^*(t)\mu_{21}, & t \in [\tau_1 + \tau^*, \infty), \end{cases}$$

and $\xi_{11}^*(t) = \xi_{22}^*(t) = 0$ for all $t \geq 0$. Note that $\eta_1^*(t) \geq 0$ because $h_1\mu_{11} \geq h_2\mu_{21}$ by assumption.

To see that $\xi_{21}^*(t) \geq 0$, note that because $h_2\mu_{21} \leq h_1\mu_{11}$, $\xi_{21}^*(t)$ is non-increasing on $[0, \tau_1)$. Moreover, if $\tau^* > 0$, it approaches zero as $t \to \tau_1$ (because $h_2\mu_{21}(\tau_2^* - \tau^* - \tau_1) = \phi_{21}$), while if $\tau^* = 0$, it approaches $\phi_{21} - h_2\mu_{21}G_2^{\tau_1}(q_2(\tau_1)) \geq 0$ instead, even if $\tau_2^* \leq \tau_1$.

Next, note that $\xi^*_{12}(t)$ is non-increasing on $[0, \tau_1)$ (because $h_1\mu_{12} \le h_2\mu_{22}$) and decreasing on $[\tau_1, \tau_1 + \tau^*)$, at which point it attains the value $\phi_{12} + p^*_2(\tau_1 + \tau^*)\mu_{22} \ge 0$. Thus, $\xi^*_{12}(t) \ge 0$ for all t.

The conditions (ODE), (ADJ), (C), (J), and (H) can be straightforwardly verified by construction.

We now verify (T). We have for $i = 1, 2$ that

$$\nabla_{z_{ii}} L(q^*(t), z^*(t), p^*(t), \eta^*(t), \gamma^*(t), \xi^*(t)) = -p^*_i(t)\mu_{ii} + \gamma^*_i(t) - \xi^*_{ii}(t) = 0$$

because $\xi^*_{ii}(t) = 0$ and $\gamma^*_i(t) = p^*_i(t)\mu_{ii}$. Next,

$$\nabla_{z_{12}} L(q^*(t), z^*(t), p^*(t), \eta^*(t), \gamma^*(t), \xi^*(t)) = \phi_{12} - p^*_1(t)\mu_{12} + \gamma^*_2(t) - \xi^*_{12}(t) = 0$$

because $\xi^*_{12}(t) = \phi_{12} - p^*_1(t)\mu_{12} + p^*_2(t)\mu_{22}$. Finally,

$$\nabla_{z_{21}} L(q^*(t), z^*(t), p^*(t), \eta^*(t), \gamma^*(t), \xi^*(t)) = \phi_{21} - p^*_2(t)\mu_{21} + \gamma^*_1(t) - \xi^*_{21}(t).$$

For $t \ge \tau_1 + \tau^*$, $\phi_{21} - p^*_2(t)\mu_{21} + \gamma^*_1(t) - \xi^*_{21}(t) = 0$ because the third term is zero and $\xi^*_{21}(t) = \phi_{21} - p^*_2(t)\mu_{21}$. For $t \in [\tau_1, \tau_1 + \tau^*)$, we get

$$\phi_{21} - p^*_2(t)\mu_{21} + p^*_1(t)\mu_{11} = \phi_{21} - h_2\mu_{21}\left((\tau^*_2 - t) - (\tau_1 + \tau^* - t)\right) = \phi_{21} - h_2\mu_{21} G_2^{\tau_1+\tau^*}(q^*_2(\tau_1 + \tau^*))$$

which is zero. Finally, for $t \in [0, \tau_1)$, $\phi_{21} - p^*_2(t)\mu_{21} + \gamma^*_1(t) - \xi^*_{21}(t) = 0$ because $\xi^*_{21}(t) = \phi_{21} - p^*_2(t)\mu_{21} + \gamma^*_1(t)$.

It remains to verify Hamiltonian minimization. The coefficients of $z^*_{11}(t)$ and $z^*_{21}(t)$ are respectively $-p^*_1(t)\mu_{11}$ and $\phi_{21} - p^*_2(t)\mu_{21}$. Note that $\phi_{21} - p^*_2(t)\mu_{21} + p^*_1(t)\mu_{11} = \xi^*_{21}(t) \ge 0$ for all t, so the Hamiltonian is minimized by setting $z^*_{11}(t)$ maximal, i.e. pool 1 prioritizing class 1. For $t < \tau_1$, $G^t_1(q_1(t)) > 0$, so pool 1 has no capacity to help class 2, i.e. $z^*_{21}(t) = 0$. For $t \in [\tau_1, \tau_1 + \tau^*)$, $\phi_{21} - p^*_2(t)\mu_{21} = 0$, so it is Hamiltonian-minimal for pool 1 to partially help class 2 (not helping is

also Hamiltonian-minimal), i.e. $z^*_{21}(t) = s_1 - z^*_{11}(t)$. Finally, for $t \geq \tau_1 + \tau^*$, $\phi_{21} - p^*_2(t)\mu_{21} \geq 0$, and so it is Hamiltonian-minimal for pool 2 to serve only its own class.

Next, the coefficients of $z^*_{12}(t)$ and $z^*_{22}(t)$ are respectively $\phi_{12} - p^*_1(t)\mu_{12}$ and $-p^*_2(t)\mu_{22}$. Note that $\phi_{12} - p^*_1(t)\mu_{12} + p^*_2(t)\mu_{22} = \xi^*_{12}(t) \geq 0$, for all t, so the Hamiltonian is minimized by setting $z^*_{22}(t)$ maximal, i.e. pool 2 prioritizing class 2. Thus, for $t < \tau^*_2$, it is Hamiltonian-minimal to have $z^*_{22}(t) = s_2$ and $z^*_{12}(t) = 0$. If $\tau^*_2 < \tau_1$ and $t \geq \tau^*_2$, $\phi_{12} - p^*_1(t)\mu_{12} = \xi^*_{12}(t) \geq 0$, and so it is again Hamiltonian-minimal to have $z^*_{12}(t) = 0$. This completes the proof. □

A.8 Optimal control for the exN1-Model

Proof. Proof of Theorem 9. We will construct the optimal primal and dual trajectories under the policy characterized in Theorem 9 and show that the conditions in Theorem 11 are satisfied.

Case I: $h_1\mu_{12} \geq h_2\mu_{22}$ and $h_1\mu_{13} \geq h_3\mu_{33}$. Let $q^*(t), z^*(t)$ be the trajectories under the given control. The trajectory is such that each pool $i = 2, 3$ gives priority to class 1 for some (possibly zero) time, then only helps its own class thereafter. To see this, suppose without loss of generality that pool 2 is the first pool to stop giving priority to class 1. After this point, $\bar{G}^t_{exN1,1,3}(q(t))$ decreases at rate 1 while pool 3 continues to give priority to class 1, and $G^t_2(q_2(t))$ decreases at rate 1 as well. Since $h_1\mu_{12} \geq h_2\mu_{22}$, (A.1) does not hold at all subsequent times. When pool 3 stops helping class 1, $\bar{G}^t_{exN1,1,3}(q(t)) = G^t_1(q_1(t)$, and again, because $h_1\mu_{12} \geq h_2\mu_{22}$, the second inequality in (A.2) is never subsequently triggered.

Define $\lambda_{1,3}(t) = \lambda_1(t) - z^*_{13}(t)\mu_{13}$ to be the class 1 arrival rate 'seen' by pool 2, after accounting for the effects of pool 3's help. We claim that $(q^*_1(t), q^*_2(t), z^*_{11}(t), z^*_{12}(t)), z^*_{22}(t))$ corresponds to that of Theorem 1 for the N-model, where the arrival rate of class 1 is replaced by $\lambda_{1,3}(t)$. To see this, consider the two cases: (i) pool 2 stops helping class 1 after pool 3, and (ii) pool 2 stops helping class 1 before pool 3. If (i), then pool 2 stops helping class 1 when (A.2) is violated, which is precisely the same condition as in the N-model. If (ii), note that pool 2 stops helping class 1 when (A.1) is violated. Recall the definition of $F^t_3(q)$. Note that when pool 2 stops helping class 1 at time t, pool 3 continues to help class 1 for time $F^t_3(q)$, by construction. After which, pool 1 will

take a further time $G_1^{t+G_3^t(q)+P_3^t(q)}(\tilde{q}_1(t+G_3^t(q)+P_3^t(q)))$ to empty. As such,

$$\bar{G}^t_{exN1,1,3}(q(t)) = G^t_1(q_1(t)),$$

where in the definition of G^t_1, the arrival rate of class 1 is replaced by $\lambda_{1,3}(t)$. This proves the claim.

From the proof of Theorem 1, we obtain $p^*_i(t), \eta^*_i(t), \gamma^*_i(t), \xi^*_{1i}(t), \xi^*_{ii}(t)$ for $i = 1, 2$. Repeating the above procedure of considering $\lambda_{1,2}(t) = \lambda_1(t) - z_{12}(t)\mu_{12}$, we obtain $p^*_i(t), \eta^*_i(t), \gamma^*_i(t)$, $\xi^*_{1i}(t), \xi^*_{ii}(t)$ for $i = 1, 3$. Note that the obtained values of $p^*_1(t), \eta^*_1(t), \gamma^*_1(t), \xi^*_{11}(t)$ are the same.

The conditions (ODE), (ADJ), (T), (C) and (J) follow directly from the N-model analysis. It remains to verify (M).

The Hamiltonian is

$$H(q(t), z(t), p(t), t) \quad = \quad \sum_i h_i q_i(t) + \sum_{j\neq 1} \phi_{1j} z_{1j}(t) + \sum_i p_i(t) \left(\lambda_i(t) - \sum_j \mu_{ij} z_{ij}(t) \right) \tag{A.31}$$

For $i = 2, 3$, the coefficients of $z_{1i}(t)$ and $z_{ii}(t)$ are $\phi_{1i} - p^*_1(t)\mu_{1i}$ and $-p^*_i(t)\mu_{ii}$. By the proof of the corresponding N-model, $\phi_{1i} - p^*_1(t)\mu_{1i} \leq -p^*_i(t)\mu_{ii} \leq 0$ whenever pool i is prioritizing class 1, $\phi_{1i} - p^*_1(t)\mu_{1i} \geq -p^*_i(t)\mu_{ii}$ whenever pool i is prioritizing its own class and $G^t_i(q_i(t)) > 0$ and $\phi_{1i} - p^*_1(t)\mu_{1i} \geq 0$ whenever pool $G^t_i(q_i(t)) = 0$. Moreover, by Assumption 6, $q_1(t) > 0$ whenever pool i is prioritizing class 1. This establishes (M).

Case II: $h_1\mu_{12} < h_2\mu_{22}$ and $h_1\mu_{13} > h_3\mu_{33}$.

The argument is similar to that of Case I. Let $q^*(t), z^*(t)$ be the trajectories under the given control. Define $\lambda_{1,3}(t) = \lambda_1(t) - z^*_{13}(t)\mu_{13}$ to be the class 1 arrival rate 'seen' by pool 2, after accounting for the effects of pool 3's help. We claim that $(q^*_1(t), q^*_2(t), z^*_{11}(t), z^*_{12}(t)), z^*_{22}(t))$ corresponds to that of Theorem 1 for the N-model, where the arrival rate of class 1 is replaced by $\lambda_{1,3}(t)$. To see this, note that pool 2 stops partial helping class 1 when the inequality in (A.5) is violated. At this time, pool 3 will continue to prioritize class 1 until (A.6) is violated (if it has not

yet stopped prioritizing class 1). Thus, the class 1 queue will take an additional

$$\bar{G}^t_{\text{exN1,1,3}}(q) = G^t_3(q_3) + P^t_3(q) + G_1^{t+G^t_3(q)+P^t_3(q)}(\tilde{q}_1(t + G^t_3(q) + P^t_3(q)))$$

time to empty, as required.

From the proof of Theorem 1, we obtain $p^*_i(t), \eta^*_i(t), \gamma^*_i(t), \xi^*_{1i}(t), \xi^*_{ii}(t)$ for $i = 1, 2$. Repeating the above procedure of considering $\lambda_{1,2}(t) = \lambda_1(t) - z_{12}(t)\mu_{12}$, we obtain $p^*_i(t), \eta^*_i(t), \gamma^*_i(t)$, $\xi^*_{1i}(t), \xi^*_{ii}(t)$ for $i = 1, 3$. Note that the obtained values of $p^*_1(t), \eta^*_1(t), \gamma^*_1(t), \xi^*_{11}(t)$ are the same.

The conditions (ODE), (ADJ), (T), (C), and (J) all follow directly from the N-model analysis. It remains to verify (M).

The Hamiltonian is

$$H(q(t), z(t), p(t), t) \quad = \quad \sum_i h_i q_i(t) + \sum_{j\neq 1} \phi_{1j} z_{1j}(t) + \sum_i p_i(t) \left(\lambda_i(t) - \sum_j \mu_{ij} z_{ij}(t) \right) \tag{A.32}$$

For $i = 2, 3$, the coefficients of $z_{1i}(t)$ and $z_{ii}(t)$ are $\phi_{1i} - p^*_1(t)\mu_{1i}$ and $-p^*_i(t)\mu_{ii}$. By the proof of the corresponding N-model (consisting of classes 1 and 3), $\phi_{13} - p^*_1(t)\mu_{13} \leq -p^*_3(t)\mu_{33} \leq 0$ whenever pool 3 is prioritizing class 1, $\phi_{13} - p^*_1(t)\mu_{13} \geq -p^*_3(t)\mu_{33}$ whenever pool 3 is prioritizing its own class and $G^t_3(q_3(t)) > 0$ and $\phi_{13} - p^*_1(t)\mu_{13} \geq 0$ whenever pool $G^t_3(q_3(t)) = 0$. Also, by the proof of the corresponding N-model (consisting of classes 1 and 2), $\phi_{12} - p^*_1(t)\mu_{12} \geq -p^*_2(t)\mu_{22}$ for all t, so it is optimal for pool 2 to prioritize its own class for all t. It also follows from the proof of the N-model that $\phi_{12} - p^*_1(t)\mu_{12} \leq 0$ when pool 2 is partially helping class 2, and $\phi_{12} - p^*_1(t)\mu_{12} \geq 0$ otherwise. Moreover, by Assumption 6, $q_1(t) > 0$ whenever pool i is providing help to class 1. This establishes (M).

Case III: $h_1\mu_{12} < h_2\mu_{22}$ and $h_1\mu_{13} < h_3\mu_{33}$. The argument is similar to the previous two cases. Let $q^*(t), z^*(t)$ be the trajectories under the given control. Define $\lambda_{1,3}(t) = \lambda_1(t) - z^*_{13}(t)\mu_{13}$ to be the class 1 arrival rate 'seen' by pool 2, after accounting for the effects of pool 3's help. It follows similarly to the other cases that $(q^*_1(t), q^*_2(t), z^*_{11}(t), z^*_{12}(t)), z^*_{22}(t))$ corresponds to that of Theorem 1 for the N-model, where the arrival rate of class 1 is replaced by $\lambda_{1,3}(t)$.

From the proof of Theorem 1, we obtain $p_i^*(t), \eta_i^*(t), \gamma_i^*(t), \xi_{1i}^*(t), \xi_{ii}^*(t)$ for $i = 1, 2$. Repeating the above procedure of considering $\lambda_{1,2}(t) = \lambda_1(t) - z_{12}(t)\mu_{12}$, we obtain $p_i^*(t), \eta_i^*(t), \gamma_i^*(t)$, $\xi_{1i}^*(t), \xi_{ii}^*(t)$ for $i = 1, 3$. Note that the obtained values of $p_1^*(t), \eta_1^*(t), \gamma_1^*(t), \xi_{11}^*(t)$ are the same.

The conditions (ODE), (ADJ), (T), (C), and (J) all follow directly from the N-model analysis. It remains to verify (M).

The Hamiltonian is

$$H(q(t), z(t), p(t), t) = \sum_i h_i q_i(t) + \sum_{j \neq 1} \phi_{1j} z_{1j}(t) + \sum_i p_i(t) \left(\lambda_i(t) - \sum_j \mu_{ij} z_{ij}(t) \right) \quad \text{(A.33)}$$

For $i = 2, 3$, the coefficients of $z_{1i}(t)$ and $z_{ii}(t)$ are $\phi_{1i} - p_1^*(t)\mu_{1i}$ and $-p_i^*(t)\mu_{ii}$. By the proof of the corresponding N-model, $\phi_{1i} - p_1^*(t)\mu_{1i} \geq -p_i^*(t)\mu_{ii}$ for $i = 2, 3$ and all t, so it is optimal for pool i to prioritize its own class for all t. It also follows from the proof of the N-model that $\phi_{1i} - p_i^*(t)\mu_{1i} \leq 0$ when pool i is partially helping class 2, and $\phi_{1i} - p_1^*(t)\mu_{1i} \geq 0$ otherwise. Moreover, by Assumption 6, $q_1(t) > 0$ whenever pool i is providing help to class 1. This establishes (M). □

A.9 Optimal control for the exN2-Model

Proof. Proof of Theorem 10. We will construct the optimal primal and dual trajectories under the policy characterized in Theorem 10 and show that the conditions in Theorem 11 are satisfied.

Let $q_1(0) = q_1$ and $q_2(0) = q_2$. For the exN2-model, the Hamiltonian takes the form:

$$H(q(t), z(t), p(t), t) = \sum_i h_i q_i(t) + \sum_{j=2}^{3} \phi_{j1} z_{j1}(t) + \sum_i p_i(t) \left(\lambda_i(t) - \sum_j \mu_{ij} z_{ij}(t) \right) \quad \text{(A.34)}$$

The augmented Hamiltonian takes the form:

$$\begin{aligned} & L(q(t), z(t), p(t), \eta(t), \gamma(t), \xi(t), t) \\ = \; & H(q(t), z(t), p(t), t) - \sum_i \eta_i(t) q_i(t) + \gamma_1(t)(z_{11}(t) + z_{21}(t) + z_{31}(t) - s_1) \\ & + \gamma_2(t)(z_{22}(t) - s_2) + \gamma_3(t)(z_{33}(t) - s_3) - \sum_{i,j} \xi_{ij}(t) z_{ij}(t). \end{aligned}$$

Case I: $h_2\mu_{21} \geq h_3\mu_{31} \geq h_1\mu_{11}$. In this case, the policy is that pool 1 first fully serves class 2 for a time $\tau_1 \geq 0$, then fully serves class 3 for a time $\tau_2 \geq 0$, then serves only its own class thereafter. To see this, note first that

$$h_2\mu_{21}G_2^{\tau_1}(q_2(\tau_1)) \leq h_1\mu_{11}\bar{G}_{exN2,1}(q(\tau_1)) + (h_3\mu_{31} - h_1\mu_{11})F^t(q(t)) + \phi_{21}$$

where equality holds by continuity if $\tau_1 > 0$. Subsequently, $h_2\mu_{21}G_2^t(q_2(t))$ decreases at rate $h_2\mu_{21}$, while the RHS decreases at rate $h_3\mu_{31}$ when pool 1 fully helps class 3 and at rate $h_1\mu_{11}$ when pool 1 serves its own class. Because $h_2\mu_{21} \geq h_3\mu_{31} \geq h_1\mu_{11}$, the inequality (A.9) never holds subsequently, and so pool 1 will not fully serve class 2 after time τ_1.

Next, note that

$$h_3\mu_{31}G_3^{\tau_1+\tau_2}(q_3(\tau_1+\tau_2)) - \phi_{31} \leq h_1\mu_{11}G_1^{\tau_1+\tau_2}(q_1(\tau_1+\tau_2))$$

with equality holding by continuity if $\tau_2 > 0$. Subsequently, when pool 1 serves its own class, $h_3\mu_{31}G_3^t(q_3(t))$ decreases at rate $h_3\mu_{31}$ while $h_1\mu_{11}G_1^t(q_1(t))$ decreases at rate $h_1\mu_{11}$, and since $h_3\mu_{31} \geq h_1\mu_{11}$, the inequality (A.10) never holds subsequently, and so pool 1 will not fully serve class 3 after time $\tau_1 + \tau_2$.

The times to deplete the three queues are

$$\begin{aligned}
\tau_1^* &= \tau_1 + \tau_2 + G_1^{\tau_1+\tau_2}(q_1^*(\tau_1+\tau_2)), \\
\tau_2^* &= \tau_1 + G_2^{\tau_1}(q_2^*(\tau_1)), \\
\tau_3^* &= \min\left\{G_3^0(q_3(0)), \tau_1 + \tau_2 + G_3^{\tau_1+\tau_2}(q_3^*(\tau_1+\tau_2))\right\}.
\end{aligned}$$

The optimal queue length trajectory follows:

$$
\begin{aligned}
q_1^*(t) &= \begin{cases} q_1 + \int_0^t (\lambda_1(s) - z_{11}^*(s)\mu_{11})\, ds, & t \in [0, \tau_1 + \tau_2), \\ q_1^*(\tau_1+\tau_2) + \int_{\tau_1+\tau_2}^t (\lambda_1(s) - s_1\mu_{11})\, ds, & t \in [\tau_1+\tau_2, \tau_1^*), \\ 0, & t \in [\tau_1^*, \infty), \end{cases} \\
q_2^*(t) &= \begin{cases} q_2 + \int_0^t (\lambda_2(s) - s_2\mu_{22} - z_{21}^*(s)\mu_{21})\, ds, & t \in [0, \tau_1), \\ q_2^*(\tau_1) + \int_{\tau_1}^t (\lambda_2(s) - s_2\mu_{22})\, ds, & t \in [\tau_1, \tau_2^*), \\ 0, & t \in [\tau_2^*, \infty), \end{cases} \\
q_3^*(t) &= \begin{cases} q_3 + \int_0^t (\lambda_3(s) - s_3\mu_{33} - z_{31}^*(s)\mu_{31})\, ds, & t \in [0, \min\{\tau_3^*, \tau_1+\tau_2\}), \\ q_3^*(\tau_1+\tau_2) + \int_{\tau_1+\tau_2}^t (\lambda_3(s) - s_3\mu_{33})\, ds, & t \in [\tau_1+\tau_2, \tau_3^*), \\ 0, & t \in [\tau_3^*, \infty). \end{cases}
\end{aligned}
$$

Note that by Assumption 7, $q_2^*(t) > 0$ for $t \in [0, \tau_1)$ and $q_3^*(t) > 0$ for $t \in [0, \tau_1 + \tau_2)$. Thus, when pool 1 is fully helping class 2, $z_{21}^*(t) = s_1$ and similarly, when pool 1 is fully helping class 3, $z_{31}^*(t) = s_1$.

Define the adjoint vectors, for $i = 1, 2, 3$,

$$
p_i^*(t) = \begin{cases} h_i(\tau_i^* - t), & t \in [0, \tau_i^*), \\ 0, & t \in [\tau_i^*, \infty). \end{cases}
$$

Define the multipliers

$$\eta_1^*(t) = \begin{cases} 0, & t \in [0, \tau_1^*), \\ h_1, & t \in [\tau_1^*, \infty), \end{cases}$$

$$\eta_2^*(t) = \begin{cases} 0, & t \in [0, \tau_2^*), \\ h_2, & t \in [\tau_2^*, \infty) \end{cases}$$

$$\eta_3^*(t) = \begin{cases} 0, & t \in [0, \tau_3^*), \\ h_3, & t \in [\tau_3^*, \infty) \end{cases}$$

$$\gamma_1^*(t) = \begin{cases} p_2^*(t)\mu_{21} - \phi_{21}, & t \in [0, \tau_1), \\ p_3^*(t)\mu_{31} - \phi_{31}, & t \in [\tau_1, \tau_1 + \tau_2), \\ p_1^*(t)\mu_{11}, & t \in [\tau_1 + \tau_2, \tau_1^*), \\ 0, & t \in [\tau_1^*, \infty), \end{cases}$$

$$\gamma_2^*(t) = \begin{cases} p_2^*(t)\mu_{22}, & t \in [0, \tau_2^*), \\ 0, & t \in [\tau_2^*, \infty), \end{cases}$$

$$\gamma_3^*(t) = \begin{cases} p_3^*(t)\mu_{33}, & t \in [0, \tau_3^*), \\ 0, & t \in [\tau_3^*, \infty) \end{cases}$$

$$\xi_{21}^*(t) = \begin{cases} 0, & t \in [0, \tau_1), \\ \phi_{21} - p_2^*(t)\mu_{21} + \gamma_1^*(t), & t \in [\tau_1, \infty) \end{cases}$$

$$\xi_{31}^*(t) = \begin{cases} 0, & t \in [\tau_1, \tau_1 + \tau_2), \\ \phi_{31} - p_3^*(t)\mu_{31} + \gamma_1^*(t), & t \notin [\tau_1, \tau_1 + \tau_2) \end{cases}$$

$$\xi_{11}^*(t) = \begin{cases} \gamma_1^*(t) - p_1^*(t)\mu_{11}, & t \in [0, \tau_1 + \tau_2), \\ 0, & t \in [\tau_1 + \tau_2, \infty) \end{cases}$$

and $\xi_{22}^*(t) = \xi_{33}^*(t) = 0$ for all $t \geq 0$. We next show $\gamma_1^*(t)$ and $\xi_{ij}^*(t)$ are non-negative. Suppose first that $\tau_1 > 0$ and $\tau_2 > 0$. By construction of the policy, we have

$$h_2\mu_{21}(\tau_2^* - \tau_1) - \phi_{21} = h_1\mu_{11}(\tau_1^* - \tau_1 - \tau_2) + h_3\mu_{31}\tau_2$$

and

$$h_3\mu_{31}(\tau_3^* - \tau_1 - \tau_2) - \phi_{31} = h_1\mu_{11}(\tau_1^* - \tau_1 - \tau_2).$$

Then,

$$p_2^*(\tau_1)\mu_{21} - \phi_{21} = p_3^*(\tau_1)\mu_{31} - \phi_{31}$$

and

$$p_3^*(\tau_1 + \tau_2)\mu_{31} - \phi_{31} = p_1^*(\tau_1 + \tau_2)\mu_{11}.$$

In particular, $\gamma_1^*(t)$ is continuous. Since $\gamma_1^*(t)$ is decreasing in each of the intervals $[0, \tau_1)$, $[\tau_1, \tau_1 + \tau_2)$ and $[\tau_1 + \tau_2, \tau_1^*)$, before reaching zero, it is non-negative. Moreover, because $h_2\mu_{21} \geq h_3\mu_{31} \geq h_1\mu_{11}$, $\gamma_1^*(t)$ decreases at a rate that is at least the rate at which $p_1^*(t)\mu_{11}$ changes in $[0, \tau_1 + \tau_2)$, $\gamma_1^*(t) \geq p_1^*(t)\mu_{11}$ in $[0, \tau_1 + \tau_2)$, i.e., $\xi_{11}^*(t) \geq 0$.

Next, from the above discussion, we have that $\xi_{21}^*(\tau_1) = 0$. Because $h_2\mu_{21} \geq h_3\mu_{31} \geq h_1\mu_{11}$,

$\xi^*_{21}(t)$ is non-decreasing for $t \in [\tau_1, \tau_2^*)$, and is non-negative for $t \geq \tau_2^*$ because $p^*_2(t) = 0$. Thus, $\xi^*_{21}(t) \geq 0$. We also have that $\xi^*_{31}(\tau_1-) = 0 = \xi^*_{31}(\tau_1 + \tau_2)$. A similar reasoning shows that $\xi^*_{31}(t)$ is non-increasing in $[0, \tau_1)$ and non-decreasing in $[\tau_1 + \tau_2, \tau_3^*)$, and so $\xi^*_{31}(t) \geq 0$ for all t.

The analysis for the cases involving $\tau_1 = 0$ and $\tau_2 = 0$ follows similarly.

The conditions (ODE), (ADJ), (C), (J), and (H) can be straightforwardly verified by construction.

We now verify (T). For $i = 2, 3$,

$$\nabla_{z_{ii}} L(q^*(t), z^*(t), p^*(t), \eta^*(t), \gamma^*(t), \xi^*(t), t) = -p^*_i(t)\mu_{ii} + \gamma^*_i(t) - \xi^*_{ii}(t) = 0$$

because $\xi^*_{ii}(t) = 0$ and $\gamma^*_i(t) = p^*_i(t)\mu_{ii}$. Next,

$$\nabla_{z_{11}} L(q^*(t), z^*(t), p^*(t), \eta^*(t), \gamma^*(t), \xi^*(t), t) = -p^*_1(t)\mu_{11} + \gamma^*_1(t) - \xi^*_{11}(t) = 0$$

because $\xi^*_{11}(t) = \gamma^*_1(t) - p^*_1(t)\mu_{11}$ for $t \in [0, \tau_1 + \tau_2)$, and $\xi^*_{11}(t) = 0$ and $\gamma^*_1(t) = p^*_1(t)\mu_{11}$ for $t \geq \tau_1 + \tau_2$. Next,

$$\nabla_{z_{21}} L(q^*(t), z^*(t), p^*(t), \eta^*(t), \gamma^*(t), \xi^*(t), t) = \phi_{21} - p^*_2(t)\mu_{21} + \gamma^*_1(t) - \xi^*_{21}(t) = 0$$

because $\xi^*_{21}(t) = 0$ and $\gamma^*_1(t) = p^*_2(t)\mu_{21} - \phi_{21}$ for $t \in [0, \tau_1)$, and $\xi^*_{21}(t) = \phi_{21} - p^*_2(t)\mu_{21} + \gamma^*_1(t)$ for $t \geq \tau_1$. Finally,

$$\nabla_{z_{31}} L(q^*(t), z^*(t), p^*(t), \eta^*(t), \gamma^*(t), \xi^*(t), t) = \phi_{31} - p^*_3(t)\mu_{31} + \gamma^*_1(t) - \xi^*_{31}(t) = 0$$

because $\xi^*_{31}(t) = 0$ and $\gamma^*_1(t) = p^*_3(t)\mu_{31} - \phi_{31}$ for $t \in [\tau_1, \tau_1 + \tau_2)$, and $\xi^*_{31}(t) = \phi_{31} - p^*_3(t)\mu_{31} + \gamma^*_1(t)$ for $t \notin [\tau_1, \tau_1 + \tau_2)$.

It remains to verify (M). It is easy to see that $z^*_{22}(t)$ and $z^*_{33}(t)$ should always be maximal. The coefficients of $z^*_{11}(t)$, $z^*_{21}(t)$ and $z^*_{31}(t)$ are respectively $-p^*_1(t)\mu_{11}$, $\phi_{21} - p^*_2(t)\mu_{21}$ and $\phi_{31} -$

$p_3^*(t)\mu_{31}$. For $t < \tau_1$, we have that $p_2^*(t)\mu_{21} - \phi_{21} \geq p_3^*(t)\mu_{31} - \phi_{31} \geq p_1^*(t)\mu_{11}$ (this follows from the earlier discussion of $\gamma_1^*(t)$), it is optimal to have $z_{21}^*(t)$ maximal. When $t \in [\tau_1, \tau_1 + \tau_2)$, we have $p_3^*(t)\mu_{31} - \phi_{31} \geq \max(p_2^*(t)\mu_{21} - \phi_{21}, p_1^*(t)\mu_{11})$, so it is optimal to have $z_{31}^*(t)$ maximal. Finally, when $t \geq \tau_1 + \tau_2$, we have $p_1^*(t)\mu_{11} \geq p_3^*(t)\mu_{31} - \phi_{31} \geq p_2^*(t)\mu_{21} - \phi_{21}$, so it is optimal to have pool 1 give class 1 priority. When $p_1^*(t) = 0$ so that $q_1^*(t) = 0$, we have that $p_i^*(t)\mu_{i1} - \phi_{i1} \leq 0$ for $i = 2, 3$, so it is optimal for pool 1 to not partially help classes 2 and 3.

Case II: $h_2\mu_{21} \geq h_1\mu_{11} > h_3\mu_{31}$. In this case, the policy is that pool 1 first fully serves class 2 for a time $\tau_1 \geq 0$, then serves only its own class 1 for time $\tau_2 = G_1^{\tau_1}(q_1(\tau_1))$ until it empties, then partially helps class 3 for some time $\tau_3 \geq 0$, then serves only its own class thereafter. To see this, note first that

$$h_2\mu_{21}G_2^{\tau_1}(q_2(\tau_1)) \leq h_1\mu_{11}G_1^{\tau_1}(q_1(\tau_1)) + h_3\mu_{31}P^{\tau_1}(q(\tau_1)) + \phi_{21}$$

where equality holds by continuity if $\tau_1 > 0$. Subsequently, $h_2\mu_{21}G_2^t(q_2(t))$ decreases at rate $h_2\mu_{21}$, while the RHS decreases at rate $h_1\mu_{11}$ when pool 1 serves only its own class and at rate $h_3\mu_{31}$ when pool 1 partially helps class 3. Because $h_2\mu_{21} \geq h_1\mu_{11} > h_3\mu_{31}$, the inequality (A.11) never holds subsequently, and so pool 1 will not fully serve class 2 after time τ_1.

The times to deplete the three queues are

$$\begin{aligned}
\tau_1^* &= \tau_1 + G_1^{\tau_1}(q_1^*(\tau_1)), \\
\tau_2^* &= \tau_1 + G_2^{\tau_1}(q_2^*(\tau_1)), \\
\tau_3^* &= \min\left\{G_3^0(q_3(0)), \tau_1^* + \tau_3 + G_3^{\tau_1^*+\tau_3}(q_3(\tau_1^* + \tau_3))\right\}.
\end{aligned}$$

The optimal queue length trajectory follows:

$$q_1^*(t) = \begin{cases} q_1 + \int_0^t (\lambda_1(s) - z_{11}^*(s)\mu_{11})\, ds, & t \in [0, \tau_1), \\ q_1^*(\tau_1) + \int_{\tau_1}^t (\lambda_1(s) - s_1\mu_{11})\, ds, & t \in [\tau_1, \tau_1^*), \\ 0, & t \in [\tau_1^*, \infty), \end{cases}$$

$$q_2^*(t) = \begin{cases} q_2 + \int_0^t (\lambda_2(s) - s_2\mu_{22} - z_{21}^*(s)\mu_{21})\, ds, & t \in [0, \tau_1), \\ q_2^*(\tau_1) + \int_{\tau_1}^t (\lambda_2(s) - s_2\mu_{22})\, ds, & t \in [\tau_1, \tau_2^*), \\ 0, & t \in [\tau_2^*, \infty). \end{cases}$$

In addition, if $\tau_3^* > 0$,

$$q_3^*(t) = \begin{cases} q_3 + \int_0^t (\lambda_3(s) - s_3\mu_{33})\, ds, & t \in [0, \tau_1^*), \\ q_3^*(\tau_1^*) + \int_{\tau_1^*}^t (\lambda_3(s) - s_3\mu_{33} - (s_1 - \lambda_1(s)/\mu_{11})\mu_{31})\, ds, & t \in [\tau_1^*, \tau_1^* + \tau_3), \\ q_3^*(\tau_1^* + \tau_3) + \int_{\tau_1^*}^t (\lambda_3(s) - s_3\mu_{33})\, ds, & t \in [\tau_1^* + \tau_3, \tau_3^*), \\ 0, & t \in [\tau_3^*, \infty), \end{cases}$$

otherwise,

$$q_3^*(t) = \begin{cases} q_3 + \int_0^t (\lambda_3(s) - s_3\mu_{33})\, ds, & t \in [0, \tau_3^*), \\ 0, & t \in [\tau_3^*, \infty). \end{cases}$$

Assumption 7 ensures that $q_2^*(t) > 0$ for $t \in [0, \tau_1)$. Thus, when pool 1 is fully helping class 2, $z_{21}^*(t) = s_1$.

Define the adjoint vectors, for $i = 2, 3$,

$$p_i^*(t) = \begin{cases} h_i(\tau_i^* - t), & t \in [0, \tau_i^*), \\ 0, & t \in [\tau_i^*, \infty). \end{cases}$$

Define also

$$p_1^*(t) = \begin{cases} h_1(\tau_1^* - t) + h_3\frac{\mu_{31}}{\mu_{11}}\tau_3, & t \in [0, \tau_1^*) \\ h_3\frac{\mu_{31}}{\mu_{11}}(\tau_1^* + \tau_3 - t), & t \in [\tau_1^*, \tau_1^* + \tau_3), \\ 0, & t \in [\tau_1^* + \tau_3, \infty). \end{cases}$$

Define the multipliers

$$\eta_1^*(t) = \begin{cases} 0, & t \in [0, \tau_1^*), \\ h_1 - h_3\frac{\mu_{31}}{\mu_{11}}, & t \in [\tau_1^*, \tau_1^* + \tau_3), \\ h_1, & t \in [\tau_1^* + \tau_3, \infty), \end{cases}$$

$$\eta_2^*(t) = \begin{cases} 0, & t \in [0, \tau_2^*), \\ h_2, & t \in [\tau_2^*, \infty) \end{cases}$$

$$\eta_3^*(t) = \begin{cases} 0, & t \in [0, \tau_3^*), \\ h_3, & t \in [\tau_3^*, \infty) \end{cases}$$

$$\gamma_1^*(t) = \begin{cases} p_2^*(t)\mu_{21} - \phi_{21}, & t \in [0, \tau_1), \\ p_1^*(t)\mu_{11}, & t \in [\tau_1, \tau_1^* + \tau_3), \\ 0, & t \in [\tau_1^* + \tau_3, \infty), \end{cases}$$

$$\gamma_2^*(t) = \begin{cases} p_2^*(t)\mu_{22}, & t \in [0, \tau_2^*), \\ 0, & t \in [\tau_2^*, \infty), \end{cases}$$

$$\gamma_3^*(t) = \begin{cases} p_3^*(t)\mu_{33}, & t \in [0, \tau_3^*), \\ 0, & t \in [\tau_3^*, \infty) \end{cases}$$

$$\xi_{21}^*(t) = \begin{cases} 0, & t \in [0, \tau_1), \\ \phi_{21} - p_2^*(t)\mu_{21} + \gamma_1^*(t), & t \in [\tau_1, \infty) \end{cases}$$

$$\xi_{31}^*(t) = \begin{cases} 0, & t \in [\tau_1^*, \tau_1^* + \tau_3), \\ \phi_{31} - p_3^*(t)\mu_{31} + \gamma_1^*(t), & t \notin [\tau_1^*, \tau_1^* + \tau_3) \end{cases}$$

$$\xi_{11}^*(t) = \begin{cases} \gamma_1^*(t) - p_1^*(t)\mu_{11}, & t \in [0, \tau_1), \\ 0, & t \in [\tau_1, \infty) \end{cases}$$

and $\xi_{22}^*(t) = \xi_{33}^*(t) = 0$ for all $t \geq 0$. Note that $\eta_1^* \geq 0$ because $h_1\mu_{11} > h_3\mu_{31}$. We next show $\gamma_1^*(t)$ and $\xi_{ij}^*(t)$ are non-negative. Suppose first that $\tau_1 > 0$. Note that

$$h_2\mu_{21}(\tau_2^* - \tau_1) - \phi_{21} = h_1\mu_{11}(\tau_1^* - \tau_1) + h_3\mu_{31}\tau_3,$$

from which it follows that

$$p_2^*(\tau_1)\mu_{21} - \phi_{21} = p_1^*(\tau_1)\mu_{11}.$$

In particular, $\gamma_1^*(t)$ is continuous. Since $\gamma_1^*(t)$ is decreasing in each of the intervals $[0, \tau_1)$ and $[\tau_1, \tau_1^* + \tau_3)$ before reaching zero, it is non-negative. Moreover, because $h_2\mu_{21} \geq h_1\mu_{11} > h_3\mu_{31}$, $\gamma_1^*(t)$ decreases at a rate that is at least the rate at which $p_1^*(t)\mu_{11}$ changes in $[0, \tau_1)$, $\gamma_1^*(t) \geq p_1^*(t)\mu_{11}$ in $[0, \tau_1)$, i.e., $\xi_{11}^*(t) \geq 0$.

Next, from the above discussion, we have that $\xi_{21}^*(\tau_1) = 0$. Because $h_2\mu_{21} \geq h_1\mu_{11} > h_3\mu_{31}$, $\xi_{21}^*(t)$ is non-decreasing for $t \in [\tau_1, \tau_2^*)$, and is non-negative for $t \geq \tau_2^*$ because $p_2^*(t) = 0$. Thus, $\xi_{21}^*(t) \geq 0$. Next, $\xi_{31}^*(t)$ is zero if $\tau_3 = 0$; suppose instead $\tau_3 > 0$. For $t \in [\tau_1^*, \tau_1^* + \tau_3)$, we have

$$\xi_{31}^*(t) = \phi_{31} - p_3^*(t)\mu_{31} + p_1^*(t)\mu_{11} = \phi_{31} - h_3\mu_{31}(\tau_1^* - t) + h_3\frac{\mu_{31}}{\mu_{11}}\mu_{11}(\tau_1^* + \tau_3 - t) = 0,$$

because $\tau_3^* = \tau_1^* + \tau_3 + \frac{\phi_{31}}{h_3\mu_{31}}$.

The analysis for the cases involving $\tau_1 = 0$ and $\tau_2 = 0$ follows similarly.

The conditions (ODE), (ADJ), (C), (J), and (H) can be straightforwardly verified by construction.

We now verify (T). For $i = 2, 3$,

$$\nabla_{z_{ii}} L(q^*(t), z^*(t), p^*(t), \eta^*(t), \gamma^*(t), \xi^*(t), t) = -p_i^*(t)\mu_{ii} + \gamma_i^*(t) - \xi_{ii}^*(t) = 0$$

because $\xi_{ii}^*(t) = 0$ and $\gamma_i^*(t) = p_i^*(t)\mu_{ii}$. Next,

$$\nabla_{z_{11}} L(q^*(t), z^*(t), p^*(t), \eta^*(t), \gamma^*(t), \xi^*(t), t) = -p_1^*(t)\mu_{11} + \gamma_1^*(t) - \xi_{11}^*(t) = 0$$

because $\xi_{11}^*(t) = \gamma_1^*(t) - p_1^*(t)\mu_{11}$ for $t \in [0, \tau_1)$, and $\xi_{11}^*(t) = 0$ and $\gamma_1^*(t) = p_1^*(t)\mu_{11}$ for $t \geq \tau_1$. Next,

$$\nabla_{z_{21}} L(q^*(t), z^*(t), p^*(t), \eta^*(t), \gamma^*(t), \xi^*(t), t) = \phi_{21} - p_2^*(t)\mu_{21} + \gamma_1^*(t) - \xi_{21}^*(t) = 0$$

because $\xi^*_{21}(t) = 0$ and $\gamma^*_1(t) = p^*_2(t)\mu_{21} - \phi_{21}$ for $t \in [0, \tau_1)$, and $\xi^*_{21}(t) = \phi_{21} - p^*_2(t)\mu_{21} + \gamma^*_1(t)$ for $t \geq \tau_1$. Finally,

$$\nabla_{z_{31}} L(q^*(t), z^*(t), p^*(t), \eta^*(t), \gamma^*(t), \xi^*(t), t) = \phi_{31} - p^*_3(t)\mu_{31} + \gamma^*_1(t) - \xi^*_{31}(t).$$

When $t \notin [\tau^*_1, \tau^*_1 + \tau_3)$, $\phi_{31} - p^*_3(t)\mu_{31} + \gamma^*_1(t) - \xi^*_{31}(t) = 0$ because $\xi^*_{31}(t) = \phi_{31} - p^*_3(t)\mu_{31} + \gamma^*_1(t)$. For $t \in [\tau^*_1, \tau^*_1 + \tau_3)$, $\xi^*_{31}(t) = 0$ and we have

$$\phi_{31} - p^*_3(t)\mu_{31} + p^*_1(t)\mu_{11} = \phi_{31} - h_3\mu_{31}(\tau^*_1 - t) + h_3\frac{\mu_{31}}{\mu_{11}}\mu_{11}(\tau^*_1 + \tau_3 - t) = 0,$$

because $\tau^*_3 = \tau^*_1 + \tau_3 + \frac{\phi_{31}}{h_3\mu_{31}}$.

It remains to verify (M). It is clear that $z^*_{22}(t)$ and $z^*_{33}(t)$ should always be maximal. The coefficients of $z^*_{11}(t)$, $z^*_{21}(t)$ and $z^*_{31}(t)$ are respectively $-p^*_1(t)\mu_{11}$, $\phi_{21} - p^*_2(t)\mu_{21}$ and $\phi_{31} - p^*_3(t)\mu_{31}$. From the earlier discussion, we have that $p^*_2(t)\mu_{21} - \phi_{21} \geq p^*_1(t)\mu_{11}$ for $t \in [0, \tau_1)$, and that $p^*_1(t)\mu_{11} = p^*_3(t)\mu_{31} - \phi_{31}$ for $t \in [\tau^*_1, \tau^*_1 + \tau_3)$. Because $h_2\mu_{21} \geq h_1\mu_{11} > h_3\mu_{31}$, we have that $p^*_1(t)\mu_{11} > p^*_3(t)\mu_{31} - \phi_{31}$ for $t \in [\tau_1, \tau^*_1)$, and hence also $p^*_2(t)\mu_{21} - \phi_{21} > p^*_1(t)\mu_{11}$ for $t \in [0, \tau_1)$. As such, for $t \in [0, \tau_1)$, it is optimal to have $z^*_{21}(t)$ maximal. When $t \in [\tau_1, \tau^*_1)$, we have $p^*_1(t)\mu_{11} \geq \max(p^*_2(t)\mu_{21} - \phi_{21}, p^*_3(t)\mu_{31} - \phi_{31})$, so it is optimal to have pool 1 serve only class 1. When $t \in [\tau^*_1, \tau^*_1 + \tau_3)$, we have $p^*_1(t)\mu_{11} = p^*_3(t)\mu_{31} - \phi_{31} \geq p^*_2(t)\mu_{21} - \phi_{21}$, so it is optimal to have pool 1 to partially help class 3. When $t \geq \tau^*_1 + \tau_3$, $p^*_1(t) = 0$, and we have that $p^*_i(t)\mu_{i1} - \phi_{i1} \leq 0$ for $i = 2, 3$, so it is optimal for pool 1 to not partially help classes 2 and 3.

Case III: $h_1\mu_{11} > h_2\mu_{21} \geq h_3\mu_{31}$. In this case, the policy is that pool 1 first serves only its own class 1 for a time $\tau^*_1 = G^0_1(q_1(0)) \geq 0$ until it empties, then partially helps class 2 for time $\tau_2 \geq 0$, then partially helps class 3 for some time $\tau_3 \geq 0$, then serves only its own class thereafter. To see this, note first that

$$h_2\mu_{21}G_2^{\tau^*_1+\tau_2}(q_2(\tau^*_1 + \tau_2)) \leq h_3\mu_{31}P^{\tau^*_1+\tau_2}(q(\tau^*_1 + \tau_2)) + \phi_{21}$$

where equality holds by continuity if $\tau_2 > 0$. Subsequently, $h_2\mu_{21}G_2^t(q_2(t))$ decreases at rate $h_2\mu_{21}$, while the RHS decreases at rate $h_3\mu_{31}$. Because $h_2\mu_{21} \geq h_3\mu_{31}$, the inequality (A.12) never holds subsequently, and so pool 1 will not partially help class 2 after time $\tau_1^* + \tau_2$.

The times to deplete the three queues are

$$\begin{aligned}
\tau_1^* &= G_1^0(q_1(0)), \\
\tau_2^* &= \min\left\{G_2^0(q_2(0)), \tau_1^* + \tau_2 + G_2^{\tau_1^*+\tau_2}(q_2^*(\tau_1^* + \tau_2))\right\}, \\
\tau_3^* &= \min\left\{G_3^0(q_3(0)), \tau_1^* + \tau_2 + \tau_3 + G_3^{\tau_1^*+\tau_2+\tau_3}(q_3^*(\tau_1^* + \tau_2 + \tau_3))\right\}.
\end{aligned}$$

The optimal queue length trajectory follows:

$$q_1^*(t) = \begin{cases} q_1 + \int_0^t (\lambda_1(s) - s_1\mu_{11})\,ds, & t \in [0, \tau_1^*), \\ 0, & t \in [\tau_1^*, \infty), \end{cases}$$

$$q_2^*(t) = \begin{cases} q_2 + \int_0^t (\lambda_2(s) - s_2\mu_{22})\,ds, & t \in [0, \min\{\tau_1^*, \tau_2^*\}), \\ q_2^*(\tau_1^*) + \int_{\tau_1^*}^t (\lambda_2(s) - s_2\mu_{22} - z_{21}^*(s)\mu_{21})\,ds, & t \in [\tau_1^*, \tau_1^* + \tau_2), \\ q_2^*(\tau_1^* + \tau_2) + \int_{\tau_1^*+\tau_2}^t (\lambda_2(s) - s_2\mu_{22})\,ds, & t \in [\tau_1^* + \tau_2, \tau_2^*), \\ 0, & t \in [\tau_2^*, \infty), \end{cases}$$

$$q_3^*(t) = \begin{cases} q_3 + \int_0^t (\lambda_3(s) - s_3\mu_{33})\,ds, & t \in [0, \min\{\tau_1^*, \tau_3^*\}), \\ q_3^*(\tau_1^*) + \int_{\tau_1^*}^t (\lambda_3(s) - s_3\mu_{33} - z_{31}^*(s)\mu_{31})\,ds, & t \in [\tau_1^*, \min\{\tau_1^* + \tau_2 + \tau_3, \tau_3^*\}), \\ q_3^*(\tau_1^* + \tau_2 + \tau_3) + \int_{\tau_1^*+\tau_2+\tau_3}^t (\lambda_3(s) - s_3\mu_{33})\,ds, & t \in [\tau_1^* + \tau_2 + \tau_3, \tau_3^*), \\ 0, & t \in [\tau_3^*, \infty). \end{cases}$$

Note for example that if $\tau_1^* > \tau_2^*$, then $\tau_2 = 0$ and $[\tau_1^*, \tau_1^* + \tau_2)$ is empty and so the corresponding expression for $q_2^*(t)$ can be ignored. Assumption 7 ensures that $q_2^*(t) > 0$ for $t \in [\tau_1^*, \tau_1^* + \tau_2)$ and $q_3^*(t) > 0$ for $t \in [\tau_1^* + \tau_2, \tau_1^* + \tau_2 + \tau_3)$. Thus, when pool 1 is partially helping class $i = 2, 3$, $z_{i1}^*(t) = s_1 - z_{11}^*(t) = s_1 - \lambda_1(s)/\mu_{11}$.

Define the adjoint vectors, for $i = 2, 3$,

$$p_i^*(t) = \begin{cases} h_i(\tau_i^* - t), & t \in [0, \tau_i^*), \\ 0, & t \in [\tau_i^*, \infty). \end{cases}$$

Define also

$$p_1^*(t) = \begin{cases} h_1(\tau_1^* - t) + h_2\frac{\mu_{21}}{\mu_{11}}\tau_2 + h_3\frac{\mu_{31}}{\mu_{11}}\tau_3, & t \in [0, \tau_1^*), \\ h_2\frac{\mu_{21}}{\mu_{11}}(\tau_1^* + \tau_2 - t) + h_3\frac{\mu_{31}}{\mu_{11}}\tau_3, & t \in [\tau_1^*, \tau_1^* + \tau_2), \\ h_3\frac{\mu_{31}}{\mu_{11}}(\tau_1^* + \tau_2 + \tau_3 - t), & t \in [\tau_1^* + \tau_2, \tau_1^* + \tau_2 + \tau_3), \\ 0, & t \in [\tau_1^* + \tau_2 + \tau_3, \infty). \end{cases}$$

Define the multipliers

$$\eta_1^*(t) = \begin{cases} 0, & t \in [0, \tau_1^*), \\ h_1 - h_2\frac{\mu_{21}}{\mu_{11}}, & t \in [\tau_1^*, \tau_1^* + \tau_2), \\ h_1 - h_3\frac{\mu_{31}}{\mu_{11}}, & t \in [\tau_1^* + \tau_2, \tau_1^* + \tau_2 + \tau_3), \\ h_1, & t \in [\tau_1^* + \tau_2 + \tau_3, \infty), \end{cases}$$

$$\eta_2^*(t) = \begin{cases} 0, & t \in [0, \tau_2^*), \\ h_2, & t \in [\tau_2^*, \infty) \end{cases}$$

$$\eta_3^*(t) = \begin{cases} 0, & t \in [0, \tau_3^*), \\ h_3, & t \in [\tau_3^*, \infty) \end{cases}$$

$$\gamma_1^*(t) = \begin{cases} p_1^*(t)\mu_{11}, & t \in [0, \tau_1^* + \tau_2 + \tau_3), \\ 0, & t \in [\tau_1^* + \tau_2 + \tau_3, \infty), \end{cases}$$

$$\gamma_2^*(t) = \begin{cases} p_2^*(t)\mu_{22}, & t \in [0, \tau_2^*), \\ 0, & t \in [\tau_2^*, \infty), \end{cases}$$

$$\gamma_3^*(t) = \begin{cases} p_3^*(t)\mu_{33}, & t \in [0, \tau_3^*), \\ 0, & t \in [\tau_3^*, \infty) \end{cases}$$

$$\xi_{21}^*(t) = \begin{cases} 0, & t \in [\tau_1^*, \tau_1^* + \tau_2), \\ \phi_{21} - p_2^*(t)\mu_{21} + \gamma_1^*(t), & t \notin [\tau_1^*, \tau_1^* + \tau_2) \end{cases}$$

$$\xi_{31}^*(t) = \begin{cases} 0, & t \in [\tau_1^* + \tau_2, \tau_1^* + \tau_2 + \tau_3), \\ \phi_{31} - p_3^*(t)\mu_{31} + \gamma_1^*(t), & t \notin [\tau_1^* + \tau_2, \tau_1^* + \tau_2 + \tau_3) \end{cases}$$

and $\xi_{11}^*(t) = \xi_{22}^*(t) = \xi_{33}^*(t) = 0$ for all $t \geq 0$. Note that $\eta_1^*(t) \geq 0$ because $h_1\mu_{11} > h_2\mu_{21} \geq h_3\mu_{31}$. We next show $\xi_{21}^*(t)$ and $\xi_{31}^*(t)$ are non-negative. Consider first $\xi_{21}^*(t)$. Because $h_1\mu_{11} > h_2\mu_{21} \geq h_3\mu_{31}$, $\gamma_1^*(t) - p_2^*(t)\mu_{21} + \phi_{21}$ is decreasing on $[0, \tau_1^*)$, constant on $[\tau_1^*, \tau_1^* + \tau_2)$ and non-decreasing on $[\tau_1^* + \tau_2, \tau_2^*)$, after which it is positive since $p_2^*(t) = 0$. So, it suffices to show that $\gamma_1^*(t) - p_2^*(t)\mu_{21} + \phi_{21}$ is non-negative at $t = \tau_1^* + \tau_2$. This holds because

$$\phi_{21} - p_2^*(\tau_1^* + \tau_2)\mu_{21} + p_1^*(\tau_1^* + \tau_2)\mu_{11} = \phi_{21} - h_2\mu_{21}(\tau_2^* - \tau_1^* - \tau_2) + h_2\mu_{21}(\tau_1^* - \tau_1^*) + h_3\mu_{31}\tau_3 \geq 0,$$

since $h_2\mu_{21}(\tau_2^* - \tau_1^* - \tau_2) - \phi_{21} \leq h_3\mu_{31}\tau_3$ by construction of the policy (equality holds if $\tau_2 > 0$).

We next turn to $\xi_{31}^*(t)$. Because $h_1\mu_{11} > h_2\mu_{21} \geq h_3\mu_{31}$, $\gamma_1^*(t) - p_3^*(t)\mu_{31} + \phi_{31}$ is non-

increasing on $[0, \tau_1^* + \tau_2)$ and constant on $[\tau_1^* + \tau_2, \tau_1^* + \tau_2 + \tau_3)$, after which it is non-decreasing since $\gamma_1^*(t) = 0$. So, it suffices to show that $\gamma_1^*(t) - p_3^*(t)\mu_{31} + \phi_{31}$ is non-negative at $t^* = \tau_1^* + \tau_2 + \tau_3$. This holds because we have

$$\phi_{31} - p_3^*(t^*)\mu_{31} + p_1^*(t^*)\mu_{11} = \phi_{31} - h_3\mu_{31}(\tau_3^* - t^*) + h_3\mu_{31}(\tau_1^* + \tau_2 + \tau_3 - t^*) \geq 0,$$

because $h_3\mu_{31}(\tau_3^* - \tau_1^* - \tau_2 - \tau_3) \leq \phi_{31}$ by construction of the policy (equality holds if $\tau_3 > 0$).

The conditions (ODE), (ADJ), (C), (J), and (H) can be straightforward verified by construction.

We now verify (T). For $i = 1, 2, 3$,

$$\nabla_{z_{ii}} L(q^*(t), z^*(t), p^*(t), \eta^*(t), \gamma^*(t), \xi^*(t), t) = -p_i^*(t)\mu_{ii} + \gamma_i^*(t) - \xi_{ii}^*(t) = 0$$

because $\xi_{ii}^*(t) = 0$ and $\gamma_i^*(t) = p_i^*(t)\mu_{ii}$. Next,

$$\nabla_{z_{21}} L(q^*(t), z^*(t), p^*(t), \eta^*(t), \gamma^*(t), \xi^*(t), t) = \phi_{21} - p_2^*(t)\mu_{21} + \gamma_1^*(t) - \xi_{21}^*(t).$$

When $t \notin [\tau_1^*, \tau_1^* + \tau_2)$, this is zero because $\xi_{21}^*(t) = \phi_{21} - p_2^*(t)\mu_{21} + \gamma_1^*(t)$. For $t \in [\tau_1^*, \tau_1^* + \tau_2)$, $\xi_{21}^*(t) = 0$ and we have

$$\phi_{21} - p_2^*(t)\mu_{21} + p_1^*(t)\mu_{11} = \phi_{21} - h_2\mu_{21}(\tau_2^* - t) + h_2\mu_{21}(\tau_1^* + \tau_2 - t) + h_3\mu_{31}\tau_3 = 0,$$

because $h_2\mu_{21}(\tau_2^* - \tau_1^* - \tau_2) - \phi_{21} = h_3\mu_{31}\tau_3$, when $\tau_2 > 0$. Finally,

$$\nabla_{z_{31}} L(q^*(t), z^*(t), p^*(t), \eta^*(t), \gamma^*(t), \xi^*(t), t) = \phi_{31} - p_3^*(t)\mu_{31} + \gamma_1^*(t) - \xi_{31}^*(t).$$

When $t \notin [\tau_1^* + \tau_2, \tau_1^* + \tau_2 + \tau_3)$, this is zero because $\xi_{31}^*(t) = \phi_{31} - p_3^*(t)\mu_{31} + \gamma_1^*(t)$. For $t \in [\tau_1^* + \tau_2, \tau_1^* + \tau_2 + \tau_3)$, $\xi_{31}^*(t) = 0$ and we have

$$\phi_{31} - p_3^*(t)\mu_{31} + p_1^*(t)\mu_{11} = \phi_{31} - h_3\mu_{31}(\tau_3^* - t) + h_3\mu_{31}(\tau_1^* + \tau_2 + \tau_3 - t) = 0,$$

because $h_3\mu_{31}(\tau_3^* - \tau_1^* - \tau_2 - \tau_3) = \phi_{31}$, when $\tau_3 > 0$.

It remains to verify (M). It is clear that $z_{22}^*(t)$ and $z_{33}^*(t)$ should always be maximal. The coefficients of $z_{11}^*(t)$, $z_{21}^*(t)$ and $z_{31}^*(t)$ are respectively $-p_1^*(t)\mu_{11}$, $\phi_{21} - p_2^*(t)\mu_{21}$ and $\phi_{31} - p_3^*(t)\mu_{31}$. From the earlier discussion on the non-negativity of $\xi_{21}^*(t)$ and $\xi_{31}^*(t)$, we have that $p_2^*(t)\mu_{21} - \phi_{21} \leq p_1^*(t)\mu_{11}$ for all t, and that $p_3^*(t)\mu_{31} - \phi_{31} \leq p_1^*(t)\mu_{11}$ for all t. Thus, it is optimal for pool 1 to give priority to class 1 at all times. For $t \in [\tau_1^*, \tau_1^* + \tau_2)$, $p_1^*(t)\mu_{11} = p_2^*(t)\mu_{21} - \phi_{21}$, and so it is optimal for pool 1 to partially help class 2. For $t \in [\tau_1^* + \tau_2, \tau_1^* + \tau_2 + \tau_3)$, $p_1^*(t)\mu_{11} = p_3^*(t)\mu_{31} - \phi_{31}$, and so it is optimal for pool 1 to partially help class 3. For $t \geq \tau_1^* + \tau_2 + \tau_3$, $p_1^*(t) = 0$, which implies that $p_i^*(t)\mu_{i1} - \phi_{i1} \leq 0$ for $i = 2, 3$, and so it is optimal for pool 1 to not help classes 2 and 3. □

Appendix B: Proofs of Results in Chapter 3

B.1 Two important stochastic dominance results

In this section, we present two important stochastic dominance results that are useful for our subsequent analysis, e.g., the proofs of Theorem 5, Lemma 2, and Lemma 4. These results build on coupling arguments and can be of independent interest.

Let $\{Y(t) = (Y_1(t), Y_2(t)); t \geq 0\}$ and $\{\tilde{Y}(t) = (\tilde{Y}_1(t), \tilde{Y}_2(t)); t \geq 0\}$ be two positive recurrent birth-and-death processes. The birth (arrival) rates are λ for both Y_i and $\tilde{Y}_i$, $i = 1, 2$. Let $\zeta_i(y)$ be the death (departure) rate of Y_i when $Y(t) = y$. We also define $\zeta_\Sigma(y) = \zeta_1(y) + \zeta_2(y)$, $\zeta_M(y) = \zeta_1(y)1\{y_1 \geq y_2\} + \zeta_2(y)1\{y_1 < y_2\}$, and $\zeta_m(y) = \zeta_1(y)1\{y_1 \leq y_2\} + \zeta_2(y)1\{y_1 > y_2\}$. Similarly, let $\tilde{\zeta}_i(\tilde{y})$ be the death rate of $\tilde{Y}_i$, $i = 1, 2$, when $\tilde{Y}(t) = \tilde{y}$, $\tilde{\zeta}_\Sigma(\tilde{y}) = \tilde{\zeta}_1(\tilde{y}) + \tilde{\zeta}_2(\tilde{y})$, $\tilde{\zeta}_M(y) = \tilde{\zeta}_1(\tilde{y})1\{\tilde{y}_1 \geq \tilde{y}_2\} + \tilde{\zeta}_2(\tilde{y})1\{\tilde{y}_1 < \tilde{y}_2\}$, and $\tilde{\zeta}_m(y) = \tilde{\zeta}_1(\tilde{y})1\{\tilde{y}_1 \leq \tilde{y}_2\} + \tilde{\zeta}_2(\tilde{y})1\{\tilde{y}_1 > \tilde{y}_2\}$.

The following two lemmas provide sufficient conditions to establish stochastic dominance between $Y(\infty)$ and $\tilde{Y}(\infty)$.

Lemma 11. *For $\{Y(t); t \geq 0\}$ and $\{\tilde{Y}(t); t \geq 0\}$, suppose*

P1) $\zeta_\Sigma(y) \geq \tilde{\zeta}_\Sigma(\tilde{y})$ whenever $y_1 + y_2 = \tilde{y}_1 + \tilde{y}_2$ and $y_1 \vee y_2 \leq \tilde{y}_1 \vee \tilde{y}_2$;

P2) $\zeta_M(y) \geq \tilde{\zeta}_M(\tilde{y})$ whenever $y_1 \vee y_2 = \tilde{y}_1 \vee \tilde{y}_2$ and $y_1 + y_2 \leq \tilde{y}_1 + \tilde{y}_2$.

Then, $Y_1(\infty) + Y_2(\infty) \leq_{st} \tilde{Y}_1(\infty) + \tilde{Y}_2(\infty)$ and $Y_1(\infty) \vee Y_2(\infty) \leq_{st} \tilde{Y}_1(\infty) \vee \tilde{Y}_2(\infty)$.

Proof. We prove the lemma by constructing a coupling, under which

$$Y_1(t) + Y_2(t) \leq \tilde{Y}_1(t) + \tilde{Y}_2(t) \text{ and } Y_1(t) \vee Y_2(t) \leq \tilde{Y}_1(t) \vee \tilde{Y}_2(t)$$

for all $t \geq 0$ path-by-path [81].

We start by introducing the coupling. Let $Y(0) = \tilde{Y}(0) = y_0$ for any fixed $y_0 \in \mathbb{N}_0^2$. We denote the k-th potential transition time in both systems by t_k with $t_0 = 0$. In particular, for $Y(t_k) = y$ and $\tilde{Y}(t_k) = \tilde{y}$, let

$$\delta_M = \begin{cases} e_1 & y_1 \geq y_2 \\ e_2 & y_1 < y_2 \end{cases} \text{ and } \delta_m = \begin{cases} e_2 & y_1 \geq y_2 \\ e_1 & y_1 < y_2 \end{cases}.$$

Similarly let

$$\tilde{\delta}_M = \begin{cases} e_1 & \tilde{y}_1 \geq \tilde{y}_2 \\ e_2 & \tilde{y}_1 < \tilde{y}_2 \end{cases} \text{ and } \tilde{\delta}_m = \begin{cases} e_2 & \tilde{y}_1 \geq \tilde{y}_2 \\ e_1 & \tilde{y}_1 < \tilde{y}_2 \end{cases}.$$

We then generate $t_{k+1} - t_k$ from an exponential distribution with rate $\Lambda := 2\lambda + \zeta_\Sigma(y) \vee \tilde{\zeta}_\Sigma(\tilde{y})$. We also generate a random variable U uniformly distributed on $[0, 1]$. We update the states of the two systems according to the following:

$$Y(t_{k+1}) = Y(t_k) + \begin{cases} \delta_M & 0 \leq U \leq \lambda/\Lambda \\ \delta_m & \lambda/\Lambda < U \leq 2\lambda/\Lambda \\ -\delta_M & 2\lambda/\Lambda < U \leq (2\lambda + \zeta_M(y))/\Lambda \\ -\delta_m & (2\lambda + \zeta_M(y))/\Lambda < U \leq (2\lambda + \zeta_\Sigma(y))/\Lambda \\ 0 & \text{Otherwise;} \end{cases}$$

and

$$\tilde{Y}(t_{k+1}) = \tilde{Y}(t_k) + \begin{cases} \tilde{\delta}_M & 0 \leq U \leq \lambda/\Lambda \\ \tilde{\delta}_m & \lambda/\Lambda < U \leq 2\lambda/\Lambda \\ -\tilde{\delta}_M & 2\lambda/\Lambda < U \leq (2\lambda + \tilde{\zeta}_M(\tilde{y}))/\Lambda \\ -\tilde{\delta}_m & (2\lambda + \tilde{\zeta}_M(\tilde{y}))/\Lambda < U \leq (2\lambda + \tilde{\zeta}_\Sigma(\tilde{y}))/\Lambda \\ 0 & \text{Otherwise.} \end{cases}$$

Now, let $S = \{k \in \mathbb{N}_0 : Y_1(t_k) + Y_2(t_k) = \tilde{Y}_1(t_k) + \tilde{Y}_2(t_k)\}$, and let the elements of S, s_i, be

ordered such that $0 = s_0 < s_1 < \cdots$. We will prove by induction that

$$Y_1(t_k) \vee Y_2(t_k) \leq \tilde{Y}_1(t_k) \vee \tilde{Y}_2(t_k) \text{ and } Y_1(t_k) + Y_2(t_k) \leq \tilde{Y}_1(t_k) + \tilde{Y}_2(t_k) \text{ for } 0 \leq k \leq s_i \text{ for any } i \in \mathbb{N}. \tag{B.1}$$

For $i = 0$, we have $Y_1(0) + Y_2(0) = \tilde{Y}_1(0) + \tilde{Y}_2(0)$ and $Y_1(0) \vee Y_2(0) = \tilde{Y}_1(0) \vee \tilde{Y}_2(0)$ by construction.

Suppose (B.1) **holds for some** i**,** $i \in \mathbb{N}_0$**.** We first note that for $k = s_i + 1$, if $s_i + 1 \in S$, we have $Y_1(t_k) + Y_2(t_k) = \tilde{Y}_1(t_k) + \tilde{Y}_2(t_k)$. If $s_i + 1 \notin S$, then by the coupling construction and P1, there must be a departure from Y but not $\tilde{Y}$. Consequently, $Y_1(t_k) + Y_2(t_k) < \tilde{Y}_1(t_k) + \tilde{Y}_2(t_k)$. This also implies that for $s_i + 1 < k < s_{i+1}$ (we set $s_{i+1} = \infty$ if s_i is the last element in S), $Y_1(t_k) + Y_2(t_k) < \tilde{Y}_1(t_k) + \tilde{Y}_2(t_k)$. We next note that for $s_i < k \leq s_{i+1}$, by our coupling construction, if there is an arrival, it either joins the larger queue in both systems or the smaller queue in both systems. Thus, in this case $Y_1(t_k) \vee Y_2(t_k) \leq \tilde{Y}_1(t_k) \vee \tilde{Y}_2(t_k)$. If there is a departure, then we further consider two cases.

Case 1. $Y_1(t_{k-1}) \vee Y_2(t_{k-1}) < \tilde{Y}_1(t_{k-1}) \vee \tilde{Y}_2(t_{k-1})$: since the difference between the two quantities changes by at most 1 at each epoch, we have $Y_1(t_k) \vee Y_2(t_k) \leq \tilde{Y}_1(t_k) \vee \tilde{Y}_2(t_k)$.

Case 2. $Y_1(t_{k-1}) \vee Y_2(t_{k-1}) = \tilde{Y}_1(t_{k-1}) \vee \tilde{Y}_2(t_{k-1})$: by P2, if there is a departure from the larger queue in $\tilde{Y}$, there must be a departure from the larger queue in Y. Moreover, if $Y_1(t_{k-1}) = Y_2(t_{k-1})$, as $Y_1(t_{k-1}) + Y_2(t_{k-1}) \leq \tilde{Y}_1(t_{k-1}) + \tilde{Y}_2(t_{k-1})$, we have $\tilde{Y}_1(t_{k-1}) = \tilde{Y}_2(t_{k-1})$. Thus, $Y_1(t_k) \vee Y_2(t_k) \leq \tilde{Y}_1(t_k) \vee \tilde{Y}_2(t_k)$.

Above all, $Y_1(t) + Y_2(t) \leq \tilde{Y}_1(t) + \tilde{Y}_2(t)$ and $Y_1(t) \vee Y_2(t) \leq \tilde{Y}_1(t) \vee \tilde{Y}_2(t)$ for all $t \geq 0$ under our coupling construction. This further implies the stochastic dominance results for the stationary distributions. □

Lemma 12. *For* $\{Y(t); t \geq 0\}$ *and* $\{\tilde{Y}(t); t \geq 0\}$*, suppose*

P1) $\zeta_\Sigma(y) \leq \tilde{\zeta}_\Sigma(\tilde{y})$ *whenever* $y_1 + y_2 = \tilde{y}_1 + \tilde{y}_2$ *and* $y_1 \wedge y_2 \geq \tilde{y}_1 \wedge \tilde{y}_2$*;*

P2) $\zeta_m(y) \leq \tilde{\zeta}_m(\tilde{y})$ *whenever* $y_1 \wedge y_2 = \tilde{y}_1 \wedge \tilde{y}_2$ *and* $y_1 + y_2 \geq \tilde{y}_1 + \tilde{y}_2$*.*

Then, $Y_1(\infty) + Y_2(\infty) \geq_{st} \tilde{Y}_1(\infty) + \tilde{Y}_2(\infty)$ *and* $Y_1(\infty) \wedge Y_2(\infty) \geq_{st} \tilde{Y}_1(\infty) \wedge \tilde{Y}_2(\infty)$.

Proof. The coupling construction follows a similar coupling idea as the proof of Lemma 11. We highlight the difference here for completeness.

Let $Y(0) = \tilde{Y}(0) = y_0$ for any fixed $y_0 \in \mathbb{N}_0^2$. We denote the k-th potential transition time in both systems by t_k with $t_0 = 0$. In particular, for $Y(t_k) = y$ and $\tilde{Y}(t_k) = \tilde{y}$, we generate $t_{k+1} - t_k$ from an exponential distribution with rate $\Lambda := 2\lambda + \zeta_\Sigma(y) \vee \tilde{\zeta}_\Sigma(\tilde{y})$. We also generate a random variable U uniformly distributed on $[0, 1]$ and update the states of the two systems according to the following:

$$Y(t_{k+1}) = Y(t_k) + \begin{cases} \delta_M & 0 \leq U \leq \lambda/\Lambda \\ \delta_m & \lambda/\Lambda < U \leq 2\lambda/\Lambda \\ -\delta_m & 2\lambda/\Lambda < U \leq (2\lambda + \zeta_m(y))/\Lambda \\ -\delta_M & (2\lambda + \zeta_m(y))/\Lambda < U \leq (2\lambda + \zeta_\Sigma(y))/\Lambda \\ 0 & \text{Otherwise;} \end{cases}$$

and

$$\tilde{Y}(t_{k+1}) = \tilde{Y}(t_k) + \begin{cases} \tilde{\delta}_M & 0 \leq U \leq \lambda/\Lambda \\ \tilde{\delta}_m & \lambda/\Lambda < U \leq 2\lambda/\Lambda \\ -\tilde{\delta}_m & 2\lambda/\Lambda < U \leq (2\lambda + \tilde{\zeta}_m(\tilde{y}))/\Lambda \\ -\tilde{\delta}_M & (2\lambda + \tilde{\zeta}_m(\tilde{y}))/\Lambda < U \leq (2\lambda + \tilde{\zeta}_\Sigma(\tilde{y}))/\Lambda \\ 0 & \text{Otherwise.} \end{cases}$$

We next prove by contradiction that

$$Y_1(t_k) \wedge Y_2(t_k) \geq \tilde{Y}_1(t_k) \wedge \tilde{Y}_2(t_k) \text{ and } Y_1(t_k) + Y_2(t_k) \geq \tilde{Y}_1(t_k) + \tilde{Y}_2(t_k) \text{ for all } k \geq 0. \tag{B.2}$$

Let $k > 0$ be the minimal index such that either (i) $Y_1(t_k) \wedge Y_2(t_k) < \tilde{Y}_1(t_k) \wedge \tilde{Y}_2(t_k)$ or (ii)

$Y_1(t_k) + Y_2(t_k) < \tilde{Y}_1(t_k) + \tilde{Y}_2(t_k)$, assuming the existence of such k.

In Scenario (i), $Y_1(t_{k-1}) \wedge Y_2(t_{k-1}) = \tilde{Y}_1(t_{k-1}) \wedge \tilde{Y}_2(t_{k-1})$ and $Y_1(t_{k-1}) + Y_2(t_{k-1}) \geq \tilde{Y}_1(t_{k-1}) + \tilde{Y}_2(t_{k-1})$. If there is an arrival event at time t_k, then based on our coupling construction, this is an arrival to both systems, and the arrival is either to the smaller queue in both systems or to the large queue in both. If $Y_1(t_{k-1}) = Y_2(t_{k-1})$ so that $Y_1 \wedge Y_2$ does not increase at t_k, then as $Y_1(t_{k-1}) + Y_2(t_{k-1}) \geq \tilde{Y}_1(t_{k-1}) + \tilde{Y}_2(t_{k-1})$, $\tilde{Y}_1(t_{k-1}) = \tilde{Y}_2(t_{k-1})$, and so $\tilde{Y}_1 \wedge \tilde{Y}_2$ does not increase either. Hence, in this case $Y_1(t_k) \wedge Y_2(t_k) = \tilde{Y}_1(t_k) \wedge \tilde{Y}_2(t_k)$. Suppose instead there is a departure event at t_k. There must be a departure from the smaller component in $\tilde{Y}$. However, by P2 and our coupling construction, there must be a departure from the smaller component in Y as well. Hence, we again have that $Y_1(t_k) \wedge Y_2(t_k) = \tilde{Y}_1(t_k) \wedge \tilde{Y}_2(t_k)$. Thus, Scenario (i) is not feasible.

In Scenario (ii), $Y_1(t_{k-1}) \wedge Y_2(t_{k-1}) \geq \tilde{Y}_1(t_{k-1}) \wedge \tilde{Y}_2(t_{k-1})$ and $Y_1(t_{k-1}) + Y_2(t_{k-1}) = \tilde{Y}_1(t_{k-1}) + \tilde{Y}_2(t_{k-1})$. Since arrivals coincide in both systems, there must be a departure from Y at t_k. However, by P1 and our coupling construction, there must be a departure from $\tilde{Y}$ as well. This rules out Scenario (ii).

Combining the analysis for the two scenarios, there is a contradiction. Thus, (B.2) holds, which further implies the stochastic dominance results for the stationary distributions. □

B.2 Application of the stochastic dominance results

B.2.1 Proofs of Lemma 2 and Lemma 4

In this section, we apply Lemma 11 to compare two system configurations. Lemmas 2 and 4 then follow as corollaries to this comparison.

Fix policy $\nu^{\lambda,*}$ for X^λ, which has n^λ servers in each dedicated server pool and n_F^λ flexible servers. Consider two auxiliary queueing systems $\tilde{X}^\lambda$ and $\check{X}^\lambda$ based on X^λ. $\tilde{X}^\lambda$ has no flexible servers. Each dedicated pool of $\tilde{X}^\lambda$ has n^λ servers that can work at rate μ and $n_F^\lambda/2$ servers that can work at rate μ_F. When assigning customers to servers, the rate-μ servers are prioritized. On the other hand, $\check{X}^\lambda$ does not have any dedicated servers. Instead, it has $2n^\lambda + n_F^\lambda$ flexible servers, among which $2n^\lambda$ servers can work at rate μ and n_F^λ servers can work at rate μ_F. When assigning

customers to servers, we again prioritize the faster servers.

Lemma 13. *Suppose $\theta \le \mu_F$. For $\tilde{X}^\lambda$, $\check{X}^\lambda$ and X^λ, if $(n^\lambda, n_F^\lambda) \in \Omega^\lambda(\theta)$,*

$$\check{X}_1^\lambda(\infty) + \check{X}_2^\lambda(\infty) \le_{st} X_1^\lambda(\infty) + X_2^\lambda(\infty) \le_{st} \tilde{X}_1^\lambda(\infty) + \tilde{X}_2^\lambda(\infty);$$

$$\check{X}_1^\lambda(\infty) \vee \check{X}_2^\lambda(\infty) \le_{st} X_1^\lambda(\infty) \vee X_2^\lambda(\infty) \le_{st} \tilde{X}_1^\lambda(\infty) \vee \tilde{X}_2^\lambda(\infty);$$

$$\left(\check{X}_1^\lambda(\infty) + \check{X}_2^\lambda(\infty) - 2n^\lambda - n_F^\lambda\right)^+ \le_{st} Q_\Sigma^\lambda(\infty) \le_{st} \left(\tilde{X}_1^\lambda(\infty) - n^\lambda - n_F^\lambda/2\right)^+ + \left(\tilde{X}_2^\lambda(\infty) - n^\lambda - n_F^\lambda/2\right)^+.$$

Proof. Because all three processes are two-dimensional birth-and-death processes with common arrival rate λ, we can apply Lemma 11. To simplify the notation, we omit the superscript λ. Set $Y = X$ and $\tilde{Y} = \tilde{X}$. Then the death rates take the form:

$$\zeta_1(y_1, y_2) = \begin{cases} \mu(y_1 \wedge n) + \mu_F((y_1 - n)^+ \wedge n_F) + \theta(y_1 - n - n_F)^+ & y_1 \ge y_2 \\ \mu(y_1 \wedge n) + \mu_F((y_1 - n)^+ \wedge (n_F - (y_2 - n)^+)^+) + \theta((y_1 - n)^+ - (n_F - (y_2 - n)^+)^+) & y_1 < y_2 \end{cases}$$

$$\zeta_2(y_1, y_2) = \begin{cases} \mu(y_2 \wedge n) + \mu_F((y_2 - n)^+ \wedge (n_F - (y_1 - n)^+)^+) + \theta((y_2 - n)^+ - (n_F - (y_1 - n)^+)^+) & y_1 \ge y_2 \\ \mu(y_2 \wedge n) + \mu_F((y_2 - n)^+ \wedge n_F) + \theta(y_2 - n - n_F)^+ & y_1 < y_2 \end{cases}$$

$$\tilde{\zeta}_1(y_1, y_2) = \mu(y_1 \wedge n) + \mu_F(n_F/2 \wedge (y_1 - n)^+) + \theta(y_1 - n - n_F/2)^+$$

$$\tilde{\zeta}_2(y_1, y_2) = \mu(y_2 \wedge n) + \mu_F(n_F/2 \wedge (y_2 - n)^+) + \theta(y_2 - n - n_F/2)^+$$

Since $\mu \ge \mu_F \ge \theta$, it is straightforward to verify that P1 and P2 in Lemma 11 hold. Thus, from the proof of Lemma 11, we can construct a coupling such that

$$Y_1(t) + Y_2(t) \le \tilde{Y}_1(t) + \tilde{Y}_2(t) \text{ and } Y_1(t) \vee Y_2(t) \le \tilde{Y}_1(t) \vee \tilde{Y}_2(t)$$

for $t \geq 0$ path-by-path. In addition,

$$\begin{aligned} Q_\Sigma(t) &= \left((X_1(t)-n)^+ + (X_2(t)-n)^+ - n_F\right)^+ \\ &\leq (X_1(t) - (n+n_F/2))^+ + (X_2(t) - (n+n_F/2))^+ \\ &\leq (\tilde{X}_1(t) - (n+n_F/2))^+ + (\tilde{X}_2(t) - (n+n_F/2))^+ \end{aligned}$$

As $\tilde{Y}$ is positive recurrent for $(n, n_F) \in \Omega^\lambda(\theta)$, so is Y. Sending t to infinity for the coupled processes, we have the stochastic dominance results in stationarity. The stochastic dominance results for X over $\check{X}$ follow similarly. □

For Lemma 2, we note that under the policy $\nu^{\lambda,*}$, for $(n^\lambda, n_F^\lambda) \in \Omega^\lambda(0)$,

$$X_1^\lambda(\infty) + X_2^\lambda(\infty) \leq_{st} \tilde{X}_1^\lambda(\infty) + \tilde{X}_2^\lambda(\infty)$$

by Lemma 13. Then, the stability of $\tilde{X}_1^\lambda$ and $\tilde{X}_2^\lambda$ implies the stability of $(X_1^\lambda, X_2^\lambda)$.

For Lemma 4, we have $\mu = \mu_F$. In this case,

$$Q_\Sigma^\lambda(\infty; 0, 2n^\lambda + n_F^\lambda) \stackrel{d}{=} \left(\check{X}_1^\lambda(\infty) + \check{X}_2^\lambda(\infty) - 2n^\lambda - n_F^\lambda\right)^+.$$

Then, by Lemma 13, under the policy $\nu^{\lambda,*}$, we have

$$Q_\Sigma^\lambda(\infty; 0, 2n^\lambda + n_F^\lambda) \leq_{st} Q_\Sigma^\lambda(\infty; n^\lambda, n_F^\lambda).$$

B.2.2 Proof of Theorem 5

Proof. We apply Lemma 11 to prove Theorem 5. To simplify notation, we omit the superscript λ. Consider $Y(t) = X(t; n, n_F; \nu^*)$ and $\tilde{Y}(t) = X(t; n, n_F; \nu)$. We will first verify that P1 and P2 in Lemma 11 hold.

For P1, $y_1 + y_2 = \tilde{y}_1 + \tilde{y}_2$ and $y_1 \vee y_2 \leq \tilde{y}_1 \vee \tilde{y}_2$. Since $\mu \geq \mu_F \geq \theta$,

$$\tilde{\zeta}_\Sigma(\tilde{y}) \leq \zeta_\Sigma(\tilde{y}) \leq \zeta_\Sigma(y).$$

For P2, $y_1 \vee y_2 = \tilde{y}_1 \vee \tilde{y}_2$ and $y_1 + y_2 \leq \tilde{y}_1 + \tilde{y_2}$. Without loss of generality, suppose $y_1 \geq y_2$ and $y_1 = \tilde{y}_1 \geq \tilde{y}_2$. Then,

$$\tilde{\zeta}_M(\tilde{y}) \leq \zeta_M(\tilde{y}) = \mu(y_1 \wedge n) + \mu_F(n_F \wedge (y_1 - n)^+) + \theta(y_1 - n - n_F)^+ = \zeta_M(y).$$

The positive recurrence of Y is established in Lemma 2. For $\tilde{Y}$, if it is not positive recurrent, we define $\tilde{Y}_i(\infty) = \infty$. Then by Lemma 11,

$$X_1^\lambda(\infty; n, n_F; \nu^*) + X_2^\lambda(\infty; n, n_F; \nu^*) \leq_{st} X_1^\lambda(\infty; n, n_F; \nu) + X_2^\lambda(\infty; n, n_F; \nu)$$

and

$$X_1^\lambda(\infty; n, n_F; \nu^*) \vee X_2^\lambda(\infty; n, n_F; \nu^*) \leq_{st} X_1^\lambda(\infty; n, n_F; \nu) \vee X_2^\lambda(\infty; n, n_F; \nu).$$

Lastly, for the queue length, consider the function $f : \mathbb{N}_0^2 \to \mathbb{N}_0$ defined by $f(y_1, y_2) = ((y_1 - n)^+ + (y_2 - n)^+ - n_F)^+$. Note that if $y_1 + y_2 \leq \tilde{y}_1 + \tilde{y}_2$ and $y_1 \vee y_2 \leq \tilde{y}_1 \vee \tilde{y}_2$, then $f(y) \leq f(\tilde{y})$. Therefore,

$$\begin{aligned} Q_\Sigma^\lambda(\infty; n, n_F, \nu^*) &= \left((X_1^\lambda(\infty; n, n_F; \nu^*) - n)^+ + (X_2^\lambda(\infty; n, n_F; \nu^*) - n)^+ - n_F\right)^+ \\ &\leq_{st} \left((X_1^\lambda(\infty; n, n_F; \nu) - n)^+ + (X_2^\lambda(\infty; n, n_F; \nu) - n)^+ - n_F\right)^+ \\ &\leq_{st} Q_\Sigma^\lambda(\infty; n, n_F; \nu). \end{aligned}$$

□

B.2.3 Optimal scheduling rule when $\theta \geq \mu_F = \mu$

Define $\phi^{\lambda,*}$ by

$$Z_i^\lambda(t) = \min\{n^\lambda, X_i^\lambda(t)\} \text{ for } i = 1, 2; \tag{B.3}$$

and if $X_1^\lambda(t) \leq X_2^\lambda(t)$,

$$Z_{F1}^\lambda(t) = \min\{n_F^\lambda, (X_1^\lambda(t) - n^\lambda)^+\}, \quad Z_{F2}^\lambda(t) = \min\{n_F^\lambda - Z_{F1}^\lambda(t), (X_2^\lambda(t) - n^\lambda)^+\}; \tag{B.4}$$

otherwise,

$$Z_{F1}^\lambda(t) = \min\{n_F^\lambda - Z_{F2}^\lambda(t), (X_1^\lambda(t) - n^\lambda)^+\}, \quad Z_{F2}^\lambda(t) = \min\{n_F^\lambda, (X_2^\lambda(t) - n^\lambda)^+\}. \tag{B.5}$$

That is, the flexible pool gives priority to the class with fewer customers in the system. The next theorem show that $\phi^{\lambda,*}$ is optimal when $\theta \geq \mu_F = \mu$.

Theorem 14. *Suppose $\theta \geq \mu = \mu_F$. For any deterministic Markovian scheduling policy ν^λ,*

$$\mathbb{E}[Q_\Sigma^\lambda(\infty; n^\lambda, n_F^\lambda; \nu^\lambda)] \geq \mathbb{E}[Q_\Sigma^\lambda(\infty; n^\lambda, n_F^\lambda; \phi^{\lambda,*})],$$

which implies that $\Pi^\lambda(n^\lambda, n_F^\lambda; \nu^\lambda) \geq \Pi^\lambda(n^\lambda, n_F^\lambda; \phi^{\lambda,*})$.

Proof. The proof of Theorem 14 uses a coupling construction similar to that of Theorem 5, but does so by considering a 'dual' problem where we maximize the number of busy servers. In particular, the key observation is that

$$\theta\mathbb{E}[Q_\Sigma^\lambda(\infty)] = 2\lambda - \mu\mathbb{E}[Z_1^\lambda(\infty) + Z_2^\lambda(\infty)] - \mu_F\mathbb{E}[Z_F^\lambda(\infty)] \tag{B.6}$$

so that minimizing $\mathbb{E}[Q_\Sigma^\lambda(\infty)]$ is equivalent to maximizing

$$\mu\mathbb{E}[Z_1^\lambda(\infty) + Z_2^\lambda(\infty)] + \mu_F\mathbb{E}[Z_F^\lambda(\infty)].$$

This may be accomplished by keeping $X_1^\lambda + X_2^\lambda$ and $X_1^\lambda \wedge X_2^\lambda$ both large. Based on this observation, we shall prove Theorem 14 using Lemma 12.

To simplify the notation, we drop the superscript λ. Let $Y(t) = X(t; n, n_F; \phi^*)$ and $\tilde{Y}(t) = X(t; n, n_F; \nu)$. We next verify P1 and P2 of Lemma 12.

For P1, $y_1 + y_2 = \tilde{y}_1 + \tilde{y}_2$ and $y_1 \wedge y_2 \geq \tilde{y}_1 \wedge \tilde{y}_2$. In this case, we have

$$\zeta_\Sigma(y) \leq \zeta_\Sigma(\tilde{y}) \leq \tilde{\zeta}_\Sigma(\tilde{y}).$$

For P2, $y_1 \wedge y_2 = \tilde{y}_1 \wedge \tilde{y}_2$ and $y_1 + y_2 \geq \tilde{y}_1 + \tilde{y_2}$. Without loss of generality, suppose $y_1 \leq y_2$ and $y_1 = \tilde{y}_1 \leq \tilde{y}_2$. Then,

$$\tilde{\zeta}_m(\tilde{y}) = \tilde{\zeta}_1(\tilde{y}) \geq \mu(y_1 \wedge (n + n_F)) + \theta(y_1 - n - n_F)^+ = \zeta_m(y)$$

From Lemma 12, we can construct a coupling, under which

$$Y_1(t) + Y_2(t) \geq \tilde{Y}_1(t) + \tilde{Y}_2(t) \text{ and } Y_1(t) \wedge Y_2(t) \geq \tilde{Y}_1(t) \wedge \tilde{Y}_2(t).$$

This further implies that

$$\mu(Z_1(t) + Z_2(t)) + \mu_F Z_F(t) \geq \mu(\tilde{Z}_1(t) + \tilde{Z}_2(t)) + \mu_F \tilde{Z}_F(t).$$

As $\theta > 0$, both Y and $\tilde{Y}$ are positive recurrent. Thus,

$$\mu(Z_1(\infty) + Z_2(\infty)) + \mu_F Z_F(\infty) \geq_{st} \mu(\tilde{Z}_1(\infty) + \tilde{Z}_2(\infty)) + \mu_F \tilde{Z}_F(\infty),$$

This completes the proof due to (B.6). □

Remark 3. *It is hard to extend the results in Theorem 14 to the case where $\mu > \mu_F$. This is because when $\mu > \mu_F$, P1 in Lemma 12 no longer holds. For example, consider $n = n_F = 1$, $y = (1, 1)$ and $\tilde{y} = (0, 2)$. In this case, $\zeta_\Sigma(y) = 2\mu > \mu + \mu_F \geq \tilde{\zeta}_\Sigma(\tilde{y})$.*

B.3 Proofs of the Results in Section 3.3.2

B.3.1 Proof of Lemma 3.

Proof. Note that $\mathbb{E}[Q_\Sigma^\lambda(\infty; \lfloor R^\lambda + \sqrt{R^\lambda} \rfloor, 0)] = O(\sqrt{\lambda})$ [59]. Thus,

$$\Pi^{\lambda,*} \leq \Pi^\lambda(\lfloor R^\lambda + \sqrt{R^\lambda} \rfloor, 0) = 2cR^\lambda + O(\sqrt{\lambda}).$$

To prove $\Pi^{\lambda,*} = 2cR^\lambda + O(\sqrt{\lambda})$, it suffices to prove that $n^{\lambda,*} = R^\lambda + O(\sqrt{\lambda})$ and $n_F^{\lambda,*} = O(\sqrt{\lambda})$.

We first prove $\limsup_{\lambda\to\infty} \frac{n^{\lambda,*}-R^\lambda}{\sqrt{\lambda}} < \infty$. Suppose by contradiction that there exists a subsequence $\{\lambda_k\}_{k\in\mathbb{N}}$ such that $\lim_{k\to\infty} \lambda_k = \infty$ and $\lim_{k\to\infty}(n^{\lambda_k,*} - R_k)/\sqrt{\lambda_k} = \infty$, where $R_k = \lambda_k/\mu$. Then,

$$\frac{\Pi^{\lambda_k}(n^{\lambda_k,*}, n_F^{\lambda_k,*}) - 2cR_k}{\sqrt{\lambda_k}} \geq \frac{c(2n^{\lambda_k,*} + n_F^{\lambda_k,*} - 2R_k)}{\sqrt{\lambda_k}} \geq \frac{2c(n^{\lambda_k,*} - R_k)}{\sqrt{\lambda_k}} \to \infty,$$

contradicting that $\Pi^{\lambda,*} \leq 2cR^\lambda + O(\sqrt{\lambda})$.

We next prove that $\liminf_{\lambda\to\infty} \frac{n^{\lambda,*}-R^\lambda}{\sqrt{\lambda}} > -\infty$ and $\limsup_{\lambda\to\infty} \frac{n_F^{\lambda,*}}{\sqrt{\lambda}} < \infty$.

Consider the case where $\theta = 0$. Note that for stability, $2n^{\lambda,*}\mu + n_F^{\lambda,*}\mu_F > 2\lambda$. To prove $\limsup_{\lambda\to\infty} \frac{n_F^{\lambda,*}}{\sqrt{\lambda}} < \infty$, we suppose for contradiction that there exists a subsequence $\{\lambda_k\}_{k\in\mathbb{N}}$ such that $\lim_{k\to\infty} \lambda_k = \infty$ and $\lim_{k\to\infty} n_F^{\lambda_k,*}/\sqrt{\lambda_k} = \infty$. Note that $2n^{\lambda_k,*} > 2\lambda_k/\mu - n_F^{\lambda_k,*}\mu_F/\mu$. Then,

$$\frac{\Pi^{\lambda_k}(n^{\lambda_k,*}, n_F^{\lambda_k,*}) - 2cR_k}{\sqrt{\lambda_k}} \geq \frac{c(2n^{\lambda_k,*} - 2R_k) + c_F n_F^{\lambda_k,*}}{\sqrt{\lambda_k}} \geq \frac{n_F^{\lambda_k,*}(c_F - c\mu_F/\mu)}{\sqrt{\lambda_k}} \to \infty,$$

contradicting that $\Pi^{\lambda,*} \leq 2cR^\lambda + O(\sqrt{\lambda})$. Since $2n^{\lambda_k,*} > 2\lambda_k/\mu - n_F^{\lambda_k,*}\mu_F/\mu$, this also shows that $\liminf_{\lambda\to\infty} \frac{n^{\lambda,*}-R^\lambda}{\sqrt{\lambda}} > -\infty$.

We now turn to the case where $\theta > 0$. We first note that $\theta\mathbb{E}[Q_\Sigma^\lambda(\infty; n^\lambda, n_F^\lambda)] \geq 2\lambda - 2n^\lambda\mu - n_F^\lambda\mu_F$. To prove $\limsup_{\lambda\to\infty} \frac{n_F^{\lambda,*}}{\sqrt{\lambda}} < \infty$, suppose for contradiction that there exists a subsequence $\{\lambda_k\}_{k\in\mathbb{N}}$

such that $\lim_{k\to\infty}\lambda_k = \infty$ and $\lim_{k\to\infty} n_F^{\lambda_k,*}/\sqrt{\lambda_k} = \infty$. Note that

$$\frac{\Pi^{\lambda_k}(n^{\lambda_k,*}, n_F^{\lambda_k,*}) - 2cR_k}{\sqrt{\lambda_k}} \geq \frac{2c(n^{\lambda_k,*} - R_k) + c_F n_F^{\lambda_k,*}}{\sqrt{\lambda_k}}. \tag{B.7}$$

Since the LHS of (B.7) must be bounded, say by $2cC$ for some constant $C > 0$, we have

$$n^{\lambda_k,*} - R_k \leq C\sqrt{\lambda_k} - \frac{c_F}{2c} n_F^{\lambda_k,*}.$$

Therefore,

$$\lambda_k - n^{\lambda_k,*}\mu \geq \frac{c_F\mu}{2c} n_F^{\lambda_k,*} - C\mu\sqrt{\lambda_k} \geq \frac{1}{2} n_F^{\lambda_k,*}\mu_F - C\mu\sqrt{\lambda_k}.$$

Next, for $h/\theta + a = c_F/\mu_F + \delta = c/\mu + \epsilon$ satisfying $0 < \delta < \epsilon$,

$$\begin{aligned}
\frac{\Pi^{\lambda_k}(n^{\lambda_k,*}, n_F^{\lambda_k,*}) - 2cR_k}{\sqrt{\lambda_k}} &= \frac{(h/\theta + a)\theta\mathbb{E}[Q_\Sigma^\lambda(\infty; n^{\lambda_k,*}, n_F^{\lambda_k,*})] + 2c(n^{\lambda_k,*} - R_k) + c_F n_F^{\lambda_k,*}}{\sqrt{\lambda_k}} \\
&\geq \frac{(h/\theta + a)(2\lambda_k - 2n^{\lambda_k,*}\mu - n_F^{\lambda_k,*}\mu_F) + 2c(n^{\lambda_k,*} - R_k) + c_F n_F^{\lambda_k,*}}{\sqrt{\lambda_k}} \\
&= \frac{2\epsilon(\lambda_k - n^{\lambda_k,*}\mu) - \delta n_F^{\lambda_k,*}\mu_F}{\sqrt{\lambda_k}} \\
&\geq \frac{(\epsilon - \delta) n_F^{\lambda_k,*}\mu_F - 2\epsilon C\mu\sqrt{\lambda_k}}{\sqrt{\lambda_k}} \\
&\to \infty
\end{aligned}$$

as $k \to \infty$. This contradicts that $\Pi^{\lambda,*} \leq 2cR^\lambda + O(\sqrt{\lambda})$, and so $n_F^\lambda = O(\sqrt{\lambda})$.

To prove that $\liminf_{\lambda\to\infty} \frac{n^{\lambda,*} - R^\lambda}{\sqrt{\lambda}} > -\infty$, assume for contradiction that there exists a subsequence $\{\lambda_k\}_{k\in\mathbb{N}}$ such that $\lim_{k\to\infty}\lambda_k = \infty$ and $\lim_{k\to\infty}(n^{\lambda_k,*} - R_k)/\sqrt{\lambda_k} = -\infty$. Then, for

$n_F^{\lambda_k,*} = O(\sqrt{\lambda_k})$, we have

$$\begin{aligned}\frac{\Pi^{\lambda_k}(n^{\lambda_k,*}, n_F^{\lambda_k,*}) - 2cR_k}{\sqrt{\lambda_k}} &\geq \frac{2\epsilon(\lambda_k - n^{\lambda_k,*}\mu) - \delta n_F^{\lambda_k,*}\mu_F}{\sqrt{\lambda_k}} \\ &= \frac{-2\epsilon\mu(n^{\lambda_k,*} - R_k) - \delta n_F^{\lambda_k,*}\mu_F}{\sqrt{\lambda_k}} \\ &\to \infty.\end{aligned}$$

This is a contradiction. □

B.3.2 Some auxiliary lemmas

Before we prove Theorem 6, we first present three auxiliary lemmas.

Lemma 14. *Let $M^\lambda = \{M^\lambda(t) : t \geq 0\}$ be a sequence of ergodic Markov chains taking values in $\mathbb{R}^m$, and $h : \mathbb{R}^m \to \mathbb{R}^n$ be a measurable function. Suppose*

1. *$h(M^\lambda(t)) \Rightarrow R(t)$ in D^n if $h(M^\lambda(0)) \Rightarrow R(0)$ as $\lambda \to \infty$, where R is a continuous ergodic process with a unique stationary distribution $R(\infty)$;*

2. *$\{h(M^\lambda(\infty)) : \lambda \geq 1\}$ is tight.*

Then, $h(M^\lambda(\infty)) \Rightarrow R(\infty)$ as $\lambda \to \infty$.

Proof. The proof follows similar lines of argument as [82]. As $\{(h(M^\lambda(\infty)) : \lambda \geq 1\}$ is tight, every subsequence has a convergent further subsequence. Let Y be a weak limit of $\{(h(M^\lambda(\infty)) : \lambda \geq 1\}$, i.e., there exists a sequence $\{\lambda_k : k \in \mathbb{N}\}$, such that $h(M^{\lambda_k}(\infty)) \Rightarrow Y$ as $k \to \infty$.

Now for each k, set $M^{\lambda_k}(0) \stackrel{d}{=} M^{\lambda_k}(\infty)$. Then, we have $M^{\lambda_k}(t) \stackrel{d}{=} M^{\lambda_k}(\infty)$ for any $t \geq 0$. This implies that $h(M^{\lambda_k}(0)) \Rightarrow Y$, which further implies that $h(M^{\lambda_k}(t)) \Rightarrow R(t)$ in D^n as $k \to \infty$. As $R(0) \stackrel{d}{=} Y$, $R(t) \stackrel{d}{=} Y$. Furthermore, as $R(t) \Rightarrow R(\infty)$ as $t \to \infty$, $Y \stackrel{d}{=} R(t) \stackrel{d}{=} R(\infty)$. Therefore, every weak limit of $\{(h(M^\lambda(\infty)) : \lambda \geq 1\}$ follows the same distribution as $R(\infty)$. This indicates that $h(M^\lambda(\infty)) \Rightarrow R(\infty)$ as $\lambda \to \infty$. □

Let $\tilde{X}_1^\lambda(\cdot)$ denote the number of customers in a system with arrival rate λ, n^λ rate-μ servers and $n_F^\lambda/2$ rate-μ_F servers.

Lemma 15. *If either (i) $\theta = 0$ and $\lambda < n^\lambda\mu + \frac{n_F^\lambda}{2}\mu_F = \lambda + \Theta(\sqrt{\lambda})$ or (ii) $\theta > 0$, $n^\lambda\mu = \lambda + O(\sqrt{\lambda})$, and $n_F^\lambda = O(\sqrt{\lambda})$,*

$$\sup_{\lambda>1} \mathbb{E}\left[\left(\frac{(\tilde{X}_1^\lambda(\infty) - n^\lambda - n_F^\lambda/2)^+}{\sqrt{\lambda}}\right)^2\right] < \infty.$$

Proof. Let $C^\lambda = n^\lambda + n_F^\lambda/2$. We first note that $\tilde{X}_1^\lambda(\cdot)$ is a positive-recurrent birth-death process. Let π^λ denote its stationary distribution. In Case (i), for $k \geq C^\lambda$, we have

$$\pi^\lambda(k) = \pi^\lambda(C^\lambda)\left(\frac{\lambda}{n^\lambda\mu + \frac{n_F^\lambda}{2}\mu_F}\right)^{k-C^\lambda}.$$

This implies that $(\tilde{X}_1^\lambda(\infty) - C^\lambda)^+$ is stochastically bounded by a geometric random variable with probability of success

$$1 - \frac{\lambda}{n^\lambda\mu + n_F^\lambda\mu_F/2} = \Theta(1/\sqrt{\lambda}).$$

Thus, $\mathbb{E}\left[\left(\tilde{X}_1^\lambda(\infty) - C^\lambda)^+\right)^2\right] = O(\lambda)$.

In Case (ii), choose $l^\lambda \geq 0$ such that $l^\lambda = O(\sqrt{\lambda})$ and $n^\lambda\mu + \frac{n_F^\lambda}{2}\mu_F + l^\lambda\theta = \lambda + \Theta(\sqrt{\lambda})$, and note that it suffices to prove that

$$\sup_{\lambda>1} \mathbb{E}\left[\left(\frac{(\tilde{X}_1^\lambda(\infty) - n^\lambda - n_F^\lambda/2 - l^\lambda)^+}{\sqrt{\lambda}}\right)^2\right] < \infty.$$

Let $D^\lambda = n^\lambda + n_F^\lambda/2 + l^\lambda$. For $k \geq D^\lambda$, we have

$$\pi^\lambda(k) = \pi^\lambda(D^\lambda)\prod_{j=1}^{k-D^\lambda}\left(\frac{\lambda}{n^\lambda\mu + \frac{n_F^\lambda}{2}\mu_F + l^\lambda\theta + j\theta}\right) \leq \pi^\lambda(D^\lambda)\left(\frac{\lambda}{n^\lambda\mu + \frac{n_F^\lambda}{2}\mu_F + l^\lambda\theta}\right)^{k-D^\lambda}.$$

Thus $(\tilde{X}_1^\lambda(\infty) - D^\lambda)^+$ is stochastically bounded by a geometric random variable with probability

of success

$$1 - \frac{\lambda}{n^\lambda \mu + n_F^\lambda \mu_F/2 + l^\lambda \theta} = \Theta(1/\sqrt{\lambda}),$$

Thus, $\mathbb{E}\left[\left(\tilde{X}_1^\lambda(\infty) - D^\lambda)^+\right)^2\right] = O(\lambda)$. □

Lemma 16. *For $(n^\lambda, n_F^\lambda) \in \Omega^\lambda(\theta)$, $2n^\lambda + n_F^\lambda = 2R^\lambda + \gamma\sqrt{R^\lambda} + o(\sqrt{R^\lambda})$, and $n_F^\lambda = O(\sqrt{\lambda})$, we have*

$$\sup_{\lambda > 1} \mathbb{E}\left[\frac{(X_i^\lambda(\infty) - n^\lambda)^-}{\sqrt{\lambda}}\right] < \infty.$$

Proof. We prove the lemma for $i = 1$ only; the case $i = 2$ is similar. Let $\pi_1^\lambda(k) = \mathbb{P}(X_1^\lambda(\infty) = k)$. Then for $0 \le k < n^\lambda$, we have $\lambda\pi_1^\lambda(k) = (k+1)\mu\pi_1^\lambda(k+1)$. This implies that

$$\begin{aligned}
\mathbb{E}[(X_1^\lambda(\infty) - n^\lambda)^-] &= \pi_1^\lambda(n^\lambda) \sum_{k=0}^{n^\lambda} (n^\lambda - k) \frac{\mu^{n^\lambda - k} n^\lambda!}{\lambda^{n^\lambda - k} k!} \\
&\le \frac{1}{\sum_{k=0}^{n^\lambda} \frac{\mu^{n^\lambda - k} n^\lambda!}{\lambda^{n^\lambda - k} k!}} \cdot \sum_{k=0}^{n^\lambda} (n^\lambda - k) \frac{\mu^{n^\lambda - k} n^\lambda!}{\lambda^{n^\lambda - k} k!} \\
&= \frac{1}{\sum_{k=0}^{n^\lambda} \frac{\mu^{n^\lambda - k} n^\lambda!}{\lambda^{n^\lambda - k} k!}} \cdot \frac{1}{\pi_c^\lambda(n^\lambda)} \cdot \pi_c^\lambda(n^\lambda) \sum_{k=0}^{n^\lambda} (n^\lambda - k) \frac{\mu^{n^\lambda - k} n^\lambda!}{\lambda^{n^\lambda - k} k!} \\
&= \frac{1}{\sum_{k=0}^{n^\lambda} \pi_c^\lambda(k)} \cdot \mathbb{E}[(X_c^\lambda(\infty) - n^\lambda)^-]
\end{aligned}$$

where X_c^λ denotes the number-in-system process of an $M/M/(n^\lambda + n_F^\lambda) + M$ queue with arrival rate λ and service rate μ, and abandonment rate $\theta \ge 0$, and π_c^λ denotes the stationary distribution of X_c^λ. As $\mathbb{E}[(X_c^\lambda(\infty) - n^\lambda - n_F^\lambda)^-] = O(\sqrt{\lambda})$ [59] and $n_F^\lambda = O(\sqrt{\lambda})$, $\mathbb{E}[(X_c^\lambda(\infty) - n^\lambda)^-] = O(\sqrt{\lambda})$. We also note that $\sum_{k=0}^{n^\lambda} \pi_c^\lambda(k) = \mathbb{P}(X_c^\lambda \le n^\lambda) = \Theta(1)$. Thus, $\mathbb{E}[(X_1^\lambda(\infty) - n^\lambda)^-] = O(\sqrt{\lambda})$. □

B.3.3 Proof of Theorem 6

Define, for $i = 1, 2$, the fluid-scale processes

$$\bar{Z}_i^\lambda(t) = \frac{Z_i^\lambda(t)}{n^\lambda}, \quad \bar{A}_i^\lambda(t) = \frac{A_i(\lambda t)}{n^\lambda}, \quad \text{and} \quad \bar{S}_i^\lambda(t) = \frac{S_i(n^\lambda \mu t)}{n^\lambda}.$$

We also define, for $i = 1, 2$,

$$\bar{G}_i^\lambda(t) = \frac{G_i(\theta\sqrt{\lambda}t)}{\sqrt{\lambda}} \text{ and } \bar{S}_{F_i}^\lambda(t) = \frac{S_{F_i}(\mu_F\sqrt{\lambda}t)}{\sqrt{\lambda}}.$$

Define, for $i = 1, 2$, the diffusion-scale processes

$$\hat{A}_i^\lambda(t) = \frac{A_i(\lambda t) - \lambda t}{\sqrt{\lambda}} \text{ and } \hat{S}_i^\lambda(t) = \frac{S_i(n^\lambda\mu t) - n^\lambda\mu t}{\sqrt{\lambda}}.$$

We first note that because

$$X_i^\lambda(t) = X_i^\lambda(0) + A_i(\lambda t) - G_i\left(\theta\int_0^t Q_i^\lambda(s)\,ds\right) - S_i\left(\mu\int_0^t Z_i^\lambda(s)\,ds\right) - S_{Fi}\left(\mu_F\int_0^t Z_{Fi}^\lambda(s)\,ds\right),$$

we have that

$$\hat{X}_i^\lambda(t) = \hat{X}_i^\lambda(0) + \hat{Y}_i^\lambda(t) + F_i(\hat{X}^\lambda)(t),$$

where

$$\hat{Y}_i^\lambda(t) = \hat{A}_i^\lambda(t) - \hat{S}_i^\lambda\left(\int_0^t \bar{Z}_i^\lambda(s)\,ds\right) - \left(\frac{S_{Fi}\left(\mu_F\int_0^t Z_{Fi}^\lambda(s)\,ds\right)}{\sqrt{\lambda}} - \mu_F\int_0^t f_i(\hat{X}^\lambda(s))\,ds\right)$$
$$-\left(\frac{G_i\left(\theta\int_0^t Q_i^\lambda(s)\,ds\right)}{\sqrt{\lambda}} - \theta\int_0^t (\hat{X}^\lambda(s)^+ - f_i(\hat{X}^\lambda(s))\,ds\right) + \frac{\lambda - n^\lambda\mu}{\sqrt{\lambda}}t$$

and

$$F_i(\hat{X}^\lambda)(t) = \mu\int_0^t \hat{X}_i^\lambda(s)^-\,ds - (\mu_F - \theta)\int_0^t f_i(\hat{X}^\lambda(s))\,ds - \theta\int_0^t \hat{X}_i^\lambda(s)^+\,ds. \tag{B.8}$$

The proof of Theorem 6 is then divided into six steps.

Step 1. Establish the convergence of the fluid-scale number-in-service processes $\bar{Z}_i^\lambda$.

Lemma 17. *For $(n^\lambda, n_F^\lambda) \in \Omega^\lambda(\theta)$, suppose $n^\lambda = R^\lambda + \beta\sqrt{R^\lambda} + o(\sqrt{R^\lambda})$ and $n_F^\lambda = \beta_F\sqrt{R^\lambda} + o(\sqrt{R^\lambda})$. If $\bar{Z}_i^\lambda(0) \to 1$, $i = 1, 2$,, then $\bar{Z}_i^\lambda \Rightarrow I$ in D as $\lambda \to \infty$.*

Proof. For any fixed $\epsilon > 0$ and $T > 0$, we shall prove that

$$\lim_{\lambda\to\infty} \mathbb{P}\left(\inf_{0\le t\le T} \bar{Z}_1^\lambda(t) < 1-\epsilon\right) \to 0.$$

Define $\bar{\tau}_1^\lambda = \inf\{0 \le t \le T : \bar{Z}_1^\lambda(t) < 1-\epsilon\}$ and $\bar{\tau}_2^\lambda = \sup\{0 \le t < \bar{\tau}_1^\lambda : \bar{Z}_1^\lambda(t) > 1-\epsilon/2\}$. Let $\bar{E}^\lambda$ be the event that $\bar{\tau}_1^\lambda$ and $\bar{\tau}_2^\lambda$ are well-defined, i.e. $\bar{\tau}_i^\lambda \le T$. The initial condition $\bar{Z}_i^\lambda(0) \to 1$ implies that $\left\{\inf_{0\le t\le T} \bar{Z}_1^\lambda(t) < 1-\epsilon\right\} \subseteq \bar{E}^\lambda$ for λ sufficiently large.

As $\bar{Z}_1^\lambda(t) < 1$ for $t \in [\bar{\tau}_2^\lambda, \bar{\tau}_1^\lambda]$, all Class 1 arrivals on $[\bar{\tau}_2^\lambda, \bar{\tau}_1^\lambda]$ join the dedicated server pool immediately on arrival. Moreover there are no abandonments from Class 1. Thus,

$$(\bar{A}_1^\lambda(\bar{\tau}_1^\lambda) - \bar{A}_1^\lambda(\bar{\tau}_2^\lambda-)) - \left(\bar{S}_1^\lambda\left(\int_0^{\bar{\tau}_1^\lambda} \bar{Z}_1(s)ds\right) - \bar{S}_1^\lambda\left(\int_0^{\bar{\tau}_2^\lambda} \bar{Z}_1(s)ds\right)\right) = \bar{Z}_1^\lambda(\bar{\tau}_1^\lambda) - \bar{Z}_1^\lambda(\bar{\tau}_2^\lambda-) \le -\epsilon/2.$$

This further implies that

$$\mathbb{P}(\bar{E}^\lambda) \le \mathbb{P}\left(\inf_{\substack{0\le s\le t\le T\\ 0\le u\le s}} (\bar{A}_1^\lambda(t) - \bar{A}_1^\lambda(s)) - (\bar{S}_1^\lambda(u+t-s) - \bar{S}_1^\lambda(u)) \le -\epsilon/2\right) \to 0,$$

where the convergence follows from the fact that, by the functional strong law of large numbers (FSLLN) for Poisson processes, $(\bar{A}_1^\lambda, \bar{S}_1^\lambda) \Rightarrow (\mu\chi, \mu\chi)$ as $\lambda \to \infty$. The analysis for $\bar{Z}_2^\lambda$ follows similarly. □

We note from Lemma 17 that in the fluid scale, the dedicated servers are busy all the time.

Step 2. Establish proper limits for the diffusion-scale processes $\hat{Y}_i^\lambda$.

Lemma 18. *For $(n^\lambda, n_F^\lambda) \in \Omega^\lambda(\theta)$, suppose $n^\lambda = R^\lambda + \beta\sqrt{R^\lambda} + o(\sqrt{R^\lambda})$ and $n_F^\lambda = \beta_F\sqrt{R^\lambda} + o(\sqrt{R^\lambda})$. If $\bar{Z}_i^\lambda(0) \to 1$, $i = 1, 2$, then*

$$\left(\hat{Y}_1^\lambda, \hat{Y}_2^\lambda\right) \Rightarrow (\sqrt{2}B_1 - \beta\sqrt{\mu}\chi, \sqrt{2}B_2 - \beta\sqrt{\mu}\chi) \text{ in } D^2 \text{ as } \lambda \to \infty,$$

where B_1 and B_2 are independent Brownian motions.

Proof. Recall that

$$\hat{Y}_i^\lambda(t) = \hat{A}_i^\lambda(t) - \hat{S}_i^\lambda\left(\int_0^t \bar{Z}_i^\lambda(s)\,ds\right) - \left(\frac{S_{Fi}\left(\mu_F\int_0^t Z_{Fi}^\lambda(s)\,ds\right)}{\sqrt{\lambda}} - \mu_F\int_0^t f_i(\hat{X}^\lambda(s))\,ds\right)$$
$$-\left(\frac{G_i\left(\theta\int_0^t Q_i^\lambda(s)\,ds\right)}{\sqrt{\lambda}} - \theta\int_0^t (\hat{X}^\lambda(s)^+ - f_i(\hat{X}^\lambda(s))\,ds\right) + \frac{\lambda - n^\lambda\mu}{\sqrt{\lambda}}t.$$

We shall analyze the five components of $\hat{Y}_i$ in sequence.

First, by the functional central limit theorem (FCLT) for Poisson processes, $\hat{A}_i^\lambda \Rightarrow B_i$ in D as $\lambda \to \infty$, where B_i is a Brownian motion.

Second, as $\int_0^\cdot \bar{Z}_i^\lambda(s)\,ds \Rightarrow \chi$ (Lemma 17), by a random time change, the FCLT for Poisson processes, and the continuous mapping theorem (Chapter 13 of [83]), we have $\hat{S}_i^\lambda\left(\int_0^\cdot \bar{Z}_i^\lambda(s)\,ds\right) \Rightarrow \tilde{B}_i$ in D as $\lambda \to \infty$, where $\tilde{B}_i$ is a Brownian motion and is independent of B_i.

Third, by the FSLLN for Poisson processes, $\bar{S}_{Fi}^\lambda \Rightarrow \mu_F\chi$ as $\lambda \to \infty$. Next, we rewrite

$$\frac{\int_0^t Z_{Fi}^\lambda(s)\,ds}{\sqrt{\lambda}} = \int_0^t f_i^\lambda(\hat{X}_1^\lambda(s), \hat{X}_2^\lambda(s))\,ds,$$

where

$$f_1^\lambda(x_1,x_2) = \begin{cases} x_1^+ \wedge \frac{n_F^\lambda}{\sqrt{\lambda}}, & x_1 \geq x_2, \\ x_1^+ \wedge \left(\frac{n_F^\lambda}{\sqrt{\lambda}} - x_2^+\right)^+, & x_1 < x_2; \end{cases} \quad \text{and} \quad f_2^\lambda(x_1,x_2) = \begin{cases} x_2^+ \wedge \left(\frac{n_F^\lambda}{\sqrt{\lambda}} - x_1^+\right)^+, & x_1 \geq x_2, \\ x_2^+ \wedge \frac{n_F^\lambda}{\sqrt{\lambda}}, & x_1 < x_2. \end{cases}$$

Then, as $f^\lambda \to f$ as $\lambda \to \infty$,

$$\begin{aligned} &\frac{S_{Fi}\left(\mu_F\int_0^t Z_{Fi}^\lambda(s)\,ds\right)}{\sqrt{\lambda}} - \mu_F\int_0^t f_i(\hat{X}^\lambda(s))\,ds \\ =&\bar{S}_{Fi}^\lambda\left(\frac{1}{\sqrt{\lambda}}\int_0^t Z_{Fi}^\lambda(s)\,ds\right) - \mu_F\int_0^t f_i^\lambda(\hat{X}^\lambda(s))\,ds + \mu_F\int_0^t f_i^\lambda(\hat{X}^\lambda(s)) - f_i(\hat{X}^\lambda(s))\,ds \\ \Rightarrow&0 \text{ as } \lambda \to \infty. \end{aligned}$$

Fourth, by the FSLLN for Poisson processes, $\bar{G}_i^\lambda \to \theta\chi$ as $\lambda \to \infty$. Then, because

$$\frac{\theta \int_0^t Q_i^\lambda(s)\,ds}{\sqrt{\lambda}} = \frac{\theta \int_0^t ((X_i^\lambda(s) - n^\lambda)^+ - Z_{Fi}^\lambda(s))\,ds}{\sqrt{\lambda}} = \theta \int_0^t (\hat{X}_i^\lambda(s)^+ - f_i^\lambda(\hat{X}_1^\lambda(s), \hat{X}_2^\lambda(s))\,ds,$$

$$\begin{aligned}
&\frac{G_i\left(\theta \int_0^t Q_i^\lambda(s)\,ds\right)}{\sqrt{\lambda}} - \theta \int_0^t (\hat{X}^\lambda(s)^+ - f_i(\hat{X}^\lambda(s))\,ds \\
=&\bar{G}_i^\lambda\left(\frac{1}{\sqrt{\lambda}} \int_0^t Q_i^\lambda(s)\,ds\right) - \frac{\theta}{\sqrt{\lambda}} \int_0^t Q_i^\lambda(s)\,ds + \theta \int_0^t (f_i(\hat{X}^\lambda(s) - f_i^\lambda(\hat{X}^\lambda(s))\,ds \\
\Rightarrow&0 \text{ as } \lambda \to \infty.
\end{aligned}$$

Fifth, under the assumption of the lemma, $(\lambda - n^\lambda\mu)/\sqrt{\lambda} \to -\beta\sqrt{\mu}$ as $\lambda \to \infty$.

Finally, putting the five parts together, we have the result. □

Step 3. Establish the C-tightness of the $\{\hat{X}^\lambda : \lambda \geq 1\}$.

Lemma 19. *For $(n^\lambda, n_F^\lambda) \in \Omega^\lambda(\theta)$, suppose $n^\lambda = R^\lambda + \beta\sqrt{R^\lambda} + o(\sqrt{R^\lambda})$ and $n_F^\lambda = \beta_F\sqrt{R^\lambda} + o(\sqrt{R^\lambda})$. If $\hat{X}^\lambda(0) \Rightarrow \hat{X}(0)$ as $\lambda \to \infty$, $\left\{\hat{X}^\lambda : \lambda \geq 1\right\}$ is C-tight in $[0,T]$ for all $T > 0$.*

Proof. Following the C-tightness definition in [84], we will prove that for any fixed $\epsilon, \gamma > 0$, there exist $\delta > 0$ and $\lambda_0 > 0$ such that for all $\lambda \geq \lambda_0$,

$$\mathbb{P}\left(\sup_{\substack{0 \leq s < t \leq T \\ |s-t| < \delta}} |\hat{X}_i^\lambda(t) - \hat{X}_i^\lambda(s)| \geq \epsilon\right) \leq \gamma$$

for $i = 1, 2$. Consider the representation

$$\begin{aligned}
\hat{X}_i^\lambda(t) =&\hat{X}_i^\lambda(0) + \hat{A}_i^\lambda(t) - \hat{S}_i^\lambda\left(\int_0^t \bar{Z}_i^\lambda(s)\,ds\right) + \frac{\lambda - n^\lambda\mu}{\sqrt{\lambda}}t + \mu \int_0^t \hat{X}_i^\lambda(s)^-\,ds \\
&- \frac{S_{Fi}\left(\mu_F \int_0^t Z_{Fi}^\lambda(s)\,ds\right)}{\sqrt{\lambda}} - \frac{G_i(\theta \int_0^t Q_i^\lambda(s)\,ds)}{\sqrt{\lambda}}.
\end{aligned}$$

First, as

$$\hat{X}_i^\lambda(0) + \hat{A}_i^\lambda(t) - \hat{S}_i^\lambda\left(\int_0^t \bar{Z}_i^\lambda(s)\,ds\right) + \frac{\lambda - n^\lambda\mu}{\sqrt{\lambda}}t \Rightarrow \hat{X}_i(0) + \sqrt{2}B_i(t) - \beta\sqrt{\mu}t \text{ in } D \text{ as } \lambda \to \infty$$

and $\sqrt{2}B_i(t) - \beta\sqrt{\mu}t$ is continuous, $\{\hat{X}_i^\lambda(0) + \hat{A}_i^\lambda(t) - \hat{S}_i^\lambda\left(\int_0^t \bar{Z}_i^\lambda(s)\,ds\right) + (\lambda - n^\lambda\mu)/\sqrt{\lambda} : \lambda \geq 1\}$ is C-tight (Lemma 4.2 of [84]).

Second, for $0 \leq s \leq t \leq T$,

$$\begin{aligned}&\frac{1}{\sqrt{\lambda}}\left(S_{Fi}\left(\mu_F\int_0^t Z_{Fi}^\lambda(u)\,du\right) - S_{Fi}\left(\mu_F\int_0^s Z_{Fi}^\lambda(u)\,du\right)\right)\\ \leq&\bar{S}_{Fi}^\lambda\left(\frac{\int_0^s Z_{Fi}^\lambda(u)\,du}{\sqrt{\lambda}} + \frac{n_F^\lambda(t-s)}{\sqrt{\lambda}}\right) - \bar{S}_{Fi}^\lambda\left(\frac{\int_0^s Z_{Fi}^\lambda(u)\,du}{\sqrt{\lambda}}\right).\end{aligned}$$

Then, the C-tightness of $\{\frac{1}{\sqrt{\lambda}}S_{Fi}\left(\mu_F\int_0^\cdot Z_{Fi}^\lambda(s)\,ds\right)\}$ follows from the fact that $n_F^\lambda/\sqrt{\lambda} \to \beta_F/\sqrt{\mu} < \infty$ and $\bar{S}_{Fi}^\lambda \Rightarrow \mu_F\chi$ in D as $\lambda \to \infty$.

Third, for $0 \leq s \leq t \leq T$, we note that

$$\begin{aligned}&\frac{1}{\sqrt{\lambda}}\left(G_i\left(\theta\int_0^t Q_i^\lambda(u)\,du\right) - G_i\left(\theta\int_0^s Q_i^\lambda(u)\,du\right)\right)\\ \leq&\bar{G}_i^\lambda\left(\frac{\int_0^s Q_i^\lambda(u)\,du + (t-s)\sup_{0\leq v\leq T}Q_i^\lambda(v)}{\sqrt{\lambda}}\right) - \bar{G}_i^\lambda\left(\frac{\int_0^s Q_i^\lambda(u)\,du}{\sqrt{\lambda}}\right).\end{aligned}$$

Then, to prove that $\{\frac{1}{\sqrt{\lambda}}G_i(\theta\int_0^\cdot Q_i^\lambda(s)\,ds)\}$ is C-tight, it suffices to prove that for any $\gamma > 0$, there exists $K, \lambda_0 > 0$, such that $\mathbb{P}\left(\sup_{0\leq v\leq T}Q_i^\lambda(v)/\sqrt{\lambda} \geq K\right) \leq \gamma/2$ for every $\lambda > \lambda_0$. Furthermore, since $n_F^\lambda = O(\sqrt{\lambda})$, it is sufficient to prove that $\mathbb{P}\left(\sup_{0\leq v\leq T}\hat{X}_i^\lambda(v) \geq K\right) \leq \gamma/2$, which follows from Lemma 13.

Fourth, we prove that $\{\mu\int_0^\cdot \hat{X}_i^\lambda(s)^-\,ds : \lambda \geq 1\}$ is C-tight. For $0 \leq s \leq t \leq T$, we first note that

$$\mu\int_0^t \hat{X}_i^\lambda(u)^-\,du - \mu\int_0^s \hat{X}_i^\lambda(u)^-\,du \leq \mu(t-s)\sup_{0\leq u\leq T}\hat{X}_i^\lambda(u)^-.$$

Next from Lemma 16 we have that for any $\gamma > 0$, there exists $K > 0$ and $\lambda_0 > 0$ such that for all

$\lambda > \lambda_0$,

$$\mathbb{P}\left(\sup_{0\le u\le T} \hat{X}_i^\lambda(u)^- > K\right) \le \gamma.$$

Thus, $\{\mu \int_0^\cdot \hat{X}_i^\lambda(s)^- \, ds : \lambda \ge 1\}$ is C-tight.

Putting the four parts together, we have the C-tightness of $\left\{\hat{X}^\lambda : \lambda \ge 1\right\}$. □

Lemma 19 implies that any subsequence of $\hat{X}^\lambda$ has a weakly convergent further subsequence and the limit is continuous almost surely (Proposition 4.1 in [84]).

Step 4. Establish that F is continuous at almost all limit points of $\hat{X}^\lambda$.

Lemma 20. *For $(n^\lambda, n_F^\lambda) \in \Omega^\lambda(\theta)$, suppose $n^\lambda = R^\lambda + \beta\sqrt{R^\lambda} + o(\sqrt{R^\lambda})$ and $n_F^\lambda = \beta_F\sqrt{R^\lambda} + o(\sqrt{R^\lambda})$. The mapping $F : D^2 \to D^2$ defined in* (B.8) *is continuous at almost all limit points of $\hat{X}^\lambda$.*

Proof. From the C-tightness of $\{\hat{X}^\lambda : \lambda \ge 1\}$, almost all sub-sequential limits of $\hat{X}^\lambda$ are continuous. Thus, it suffices to prove continuity of F under the uniform topology. We denote by $\hat{X}$ a generic sub-sequential limit of $\hat{X}^\lambda$.

Fix $T > 0$. For $X \in D^2$, define $\|X\|_T = \sup_{0\le t\le T} \max(|X_1(t)|, |X_2(t)|)$. Now, fix $\epsilon > 0$. Consider $X, Y \in D^2$ with X continuous and $\|X - Y\|_T < \epsilon/2$.

For $0 \le t \le T$,

$$\left|\int_0^t X_i(s)^- \, ds - \int_0^t Y_i(s)^- \, ds\right| < \epsilon t/2 \le \epsilon T/2.$$

Similarly,

$$\left|\int_0^t X_i(s)^+ \, ds - \int_0^t Y_i(s)^+ \, ds\right| < \epsilon t/2 \le \epsilon T/2.$$

Next, for f_i, when $|X_1(t) - X_2(t)| \ge \epsilon$, if $X_1(t) > X_2(t)$, then $Y_1(t) > Y_2(t)$, and if $X_1(t) < X_2(t)$, then $Y_1(t) < Y_2(t)$. In this case, we have $|f_i(X(t)) - f_i(Y(t))| \le \epsilon/2$. If, instead, $|X_1(t) - X_2(t)| < \epsilon$,

$|f_i(X(t)) - f_i(Y(t))| \le \beta_F/\sqrt{\mu}$. Putting the two cases together, we have

$$\begin{aligned}
&\left|\int_0^t f_i(X(s))\,ds - \int_0^t f_i(Y(s))\,ds\right| \\
\le&\frac{\epsilon}{2}\int_0^t 1\{|X_1(s) - X_2(s)| \ge \epsilon\}\,ds + \frac{\beta_F}{\sqrt{\mu}}\int_0^t 1\{|X_1(s) - X_2(s)| < \epsilon\}\,ds \\
\le&\frac{\epsilon T}{2} + \frac{\beta_F}{\sqrt{\mu}}\int_0^T 1\{|X_1(s) - X_2(s)| < \epsilon\}\,ds.
\end{aligned}$$

Above all,

$$\begin{aligned}
|F_i(X)(t) - F_i(Y)(t)| \le&\mu\left|\int_0^t X_i(s)^-\,ds - \int_0^t Y_i(s)^-\,ds\right| + \theta\left|\int_0^t X_i(s)^+\,ds - \int_0^t Y_i(s)^+\,ds\right| \\
&+ (\mu_F - \theta)\left|\int_0^t f_i(X(s))\,ds - \int_0^t f_i(Y(s))\,ds\right| \\
\le&\frac{\epsilon\mu T}{2} + \frac{\epsilon\theta T}{2} + \frac{\epsilon(\mu_F - \theta)T}{2} + \frac{\beta_F}{\sqrt{\mu}}(\mu_F - \theta)\int_0^T 1\{|X_1(t) - X_2(t)| < \epsilon\}\,dt \\
\to&\frac{\beta_F}{\sqrt{\mu}}(\mu_F - \theta)\int_0^T 1\{X_1(t) = X_2(t)\}\,dt \text{ as } \epsilon \downarrow 0.
\end{aligned}$$

This implies that to prove continuity of F at $\hat{X}$, it suffices to prove that

$$\mathbb{P}\left(\int_0^T 1\{\hat{X}_1(t) = \hat{X}_2(t)\}\,dt = 0\right) = 1.$$

Note that $\hat{X}^\lambda \Rightarrow \hat{X}$ implies that $\hat{X}$ takes the form

$$\hat{X}_i(t) = \hat{X}_i(0) + \sqrt{2}B_i(t) - \beta\sqrt{\mu}t + \mu\int_0^t \hat{X}_i(s)^-\,ds + \theta\int_0^t \hat{X}_i(s)^+\,ds - L_i(t),$$

where $L_i(t)$ is a weak limit of $\{(\mu_F - \theta)\int_0^t f_i(\hat{X}^\lambda(s))\,ds\}$. We also note that $L_i(t)$ is monotone increasing and bounded by $(\mu_F - \theta)\beta_F t/\sqrt{\mu}$. Thus, L_i has finite total variation. Meanwhile, since $\hat{X}$ is continuous, $\left\|\hat{X}\right\|_T < \infty$. As $\int_0^t \hat{X}_i(s)^-\,ds \le \int_0^T \hat{X}_i(s)^-\,ds < \infty$, $\mu\int_0^t \hat{X}_i(s)^-\,ds$ has finite total variation as well. Similarly, $\theta\int_0^t \hat{X}_i(s)^+\,ds$ has finite total variation as well. It then follows that $\hat{X}(t)$ is the sum of a Brownian motion and other terms of finite total variation. Therefore $\hat{X}$ spends almost surely zero time on $\{\hat{X}_1(s) = \hat{X}_2(s)\}$ [85]. □

Step 5. Establish that $\hat{X}$ is suitably well-posed.

The following lemma follows directly from Proposition 5.3.10 in [86].

Lemma 21. *The diffusion equation*

$$\hat{X}_i(t) = \hat{X}_i(0) + \sqrt{2}B_i(t) - \beta\sqrt{\mu}t + \mu\int_0^t \hat{X}_i(s)^- \, ds - (\mu_F - \theta)\int_0^t f_i(\hat{X}(s)) \, ds - \theta\int_0^t \hat{X}_i(s)^+ \, ds$$

has a unique (weak) solution.

Steps 1-5 together establish the process level convergence of $\hat{X}^\lambda$, i.e.,

$$\hat{X}^\lambda \Rightarrow \hat{X} \text{ in } D^2 \text{ as } \lambda \to \infty.$$

We also note that

$$\hat{Q}^\lambda_\Sigma(t) = \left(\hat{X}^\lambda_1(t)^+ + \hat{X}^\lambda_2(t)^+ - n^\lambda_F/\sqrt{\lambda}\right)^+ = \left(\hat{X}^\lambda_1(t)^+ + \hat{X}^\lambda_2(t)^+ - \beta_F/\sqrt{\mu}\right)^+ + g^\lambda(\hat{X}^\lambda_1(t), \hat{X}^\lambda_2(t))$$

where

$$\begin{aligned}\left|g^\lambda(\hat{X}^\lambda_1(t), \hat{X}^\lambda_2(t))\right| &= \left|\left(\hat{X}^\lambda_1(t)^+ + \hat{X}^\lambda_2(t)^+ - n^\lambda_F/\sqrt{\lambda}\right)^+ - \left(\hat{X}^\lambda_1(t)^+ + \hat{X}^\lambda_2(t)^+ - \beta_F/\sqrt{\mu}\right)^+\right| \\ &\leq |n^\lambda_F/\sqrt{\lambda} - \beta_F/\sqrt{\mu}| \to 0 \text{ as } \lambda \to \infty.\end{aligned}$$

This implies that $\hat{Q}^\lambda_\Sigma \Rightarrow \left(\hat{X}^+_1 + \hat{X}^+_2 - \beta_F/\sqrt{\mu}\right)^+$ in D as $\lambda \to \infty$.

Step 6. Establish the appropriate interchange of limits and uniform integrability results.

Lemma 22. *For $(\beta, \beta_F) \in \hat{\Omega}(\theta)$, the diffusion process $\hat{X}$ is positive recurrent.*

Proof. We will show that the function $V(x_1, x_2) = \frac{1}{2}(x_1^2 + x_2^2)$ is a Lyapunov function. The generator

G of $\hat{X}$ applied to V is given by

$$GV(x) = \sum_{i=1}^{2} x_i \left(-\beta\sqrt{\mu} + \mu x_i^- - \theta x_i^+ - (\mu_F - \theta) f_i(x)\right)$$

for $x \in \mathbb{R}^2$.

We first consider the case $\theta > 0$. Because f_i is bounded (by $\beta_F/\sqrt{\mu}$), we have that $-\beta\sqrt{\mu}+\mu x_i^- - \theta x_i^+ - (\mu_F - \theta) f_i(x) \leq -1$ for all $x_i > 0$ large enough, and $-\beta\sqrt{\mu} + \mu x_i^- - \theta x_i^+ - (\mu_F - \theta) f_i(x) \geq 1$ for all $-x_i > 0$ large enough. It follows that $GV(x) \leq -1$ for all $|x|$ large enough.

Suppose instead $\theta = 0$. If $\beta > 0$, $-\beta\sqrt{\mu} + \mu x_i^- - \mu_F f_i(x) \leq -\beta\sqrt{\mu} < 0$ for all $x_i > 0$, and $-\beta\sqrt{\mu} + \mu x_i^- - \mu_F f_i(x) \geq 1$ for all $-x_i > 0$ large enough. Thus we may suppose $\beta \leq 0$.

Suppose first both x_i are non-negative, with $x_1 \geq x_2 \geq 0$ (the case $x_2 > x_1 \geq 0$ is similar). Then, if $x_1 \geq \beta_F/\sqrt{\mu}$,

$$GV(x) = x_1(-\beta\sqrt{\mu} - \mu_F\beta_F/\sqrt{\mu}) - x_2\beta\sqrt{\mu} \leq \frac{-x_1}{\sqrt{\mu}}(2\beta\mu + \beta_F\mu_F) \leq -1$$

for x_1 large enough, since $2\beta\mu + \beta_F\mu_F > 0$.

Next, suppose exactly one x_i is non-negative, with $x_1 \geq 0 > x_2$ (the case $x_2 \geq 0 > x_1$ is similar). We have, if $x_1 \geq \beta_F/\sqrt{\mu}$,

$$\begin{aligned} GV(x) &= x_1(-\beta\sqrt{\mu} - \mu_F f_i(x)) - \mu x_2^2 - \beta\sqrt{\mu}x_2 \\ &\leq -\frac{x_1}{\sqrt{\mu}}(\beta\mu + \beta_F\mu_F) - \mu x_2^2 \leq -\frac{x_1}{\sqrt{\mu}}(2\beta\mu + \beta_F\mu_F) - \mu x_2^2 \leq -1 \end{aligned}$$

for $|x|$ large enough, since $2\beta\mu + \beta_F\mu_F > 0$. If instead $0 \leq x_1 < \beta_F/\sqrt{\mu}$, we have that $x_1(-\beta\sqrt{\mu} - \mu_F f_i(x))$ is bounded, so again $GV(x) \leq -1$ for $|x|$ large enough.

Finally, suppose $x_i < 0$ for $i = 1, 2$. We have

$$GV(x) = \sum_{i=1}^{2} x_i(-\beta\sqrt{\mu} - \mu x_i) \leq -1$$

for $|x|$ large enough. This completes the proof. □

Lemma 22 implies that $\hat{X}(\infty)$ is well defined.

Lemma 23. *Suppose* $n^\lambda = R^\lambda + \beta\sqrt{R^\lambda} + o(\sqrt{R^\lambda})$ *and* $n_F^\lambda = \beta_F\sqrt{R^\lambda} + o(\sqrt{R^\lambda})$, *with* $(n^\lambda, n_F^\lambda) \in \Omega^\lambda(\theta)$ *and* $(\beta, \beta_F) \in \hat{\Omega}(\theta)$. *Then,*

$$\hat{Q}_\Sigma^\lambda(\infty) \Rightarrow \left(\hat{X}_1(\infty)^+ + \hat{X}_2(\infty)^+ - \beta_F/\sqrt{\mu}\right)^+ \text{ as } \lambda \to \infty$$

and

$$\mathbb{E}[\hat{Q}_\Sigma^\lambda(\infty)] \to \mathbb{E}\left[\left(\hat{X}_1(\infty)^+ + \hat{X}_2(\infty)^+ - \beta_F/\sqrt{\mu}\right)^+\right] \text{ as } \lambda \to \infty.$$

Proof. Note that

$$\begin{aligned}
&\sup_{\lambda>1} \mathbb{E}[(\hat{X}_i^\lambda(\infty)^+)^2] \\
&= \sup_{\lambda>1} \mathbb{E}\left[\left(\frac{(X_i^\lambda(\infty) - n^\lambda)^+}{\sqrt{\lambda}}\right)^2\right] \\
&\le \sup_{\lambda>1} \mathbb{E}\left[\left(\frac{\sum_{j=1}^2 (X_j^\lambda(\infty) - n^\lambda)^+}{\sqrt{\lambda}}\right)^2\right] \\
&\le \sup_{\lambda>1} \mathbb{E}\left[\left(\frac{\sum_{j=1}^2 \left((X_j^\lambda(\infty) - n^\lambda - n_F^\lambda/2)^+ + n_F^\lambda/2\right)}{\sqrt{\lambda}}\right)^2\right] \\
&\le \sup_{\lambda>1} 4\mathbb{E}\left[\left(\frac{(\tilde{X}_1^\lambda(\infty) - n^\lambda - n_F^\lambda/2)^+ + n_F^\lambda/2}{\sqrt{\lambda}}\right)^2\right] \text{ by Lemma 13 and Cauchy-Schwarz Inequality} \\
&< \infty \text{ by Lemma 15.}
\end{aligned} \tag{B.9}$$

In addition,

$$\sup_{\lambda>1} \mathbb{E}[\hat{X}_i^\lambda(\infty)^-] = \sup_{\lambda>1} \mathbb{E}\left[\frac{(X_i^\lambda(\infty) - n^\lambda)^-}{\sqrt{\lambda}}\right] < \infty$$

by Lemma 16. Then we have $\sup_{\lambda>1} \mathbb{E}[|\hat{X}_i^\lambda(\infty)|] < \infty$, i.e., $\{\hat{X}^\lambda(\infty) : \lambda > 1\}$ is tight. Thus, $\hat{X}^\lambda(\infty) \Rightarrow \hat{X}(\infty)$ as $\lambda \to \infty$ by Lemma 14. By the continuous mapping and converging together theorems, we have $\hat{Q}_\Sigma^\lambda(\infty) \Rightarrow \left(\hat{X}_1(\infty)^+ + \hat{X}_2(\infty)^+ - \beta_F/\sqrt{\mu}\right)^+$ as $\lambda \to \infty$.

Next, the bound in (B.9) also implies that $\{\hat{X}_i^\lambda(\infty)^+ : \lambda > 1\}$ is uniformly integrable. As $\hat{Q}_\Sigma^\lambda(\infty) \le \hat{X}_1^\lambda(\infty)^+ + \hat{X}_2^\lambda(\infty)^+$, $\{\hat{Q}_\Sigma^\lambda(\infty) : \lambda > 1\}$ is also uniformly integrable. Thus,

$$\mathbb{E}[\hat{Q}_\Sigma^\lambda(\infty)] \to \mathbb{E}\left[\left(\hat{X}_1(\infty)^+ + \hat{X}_2(\infty)^+ - \beta_F/\sqrt{\mu}\right)^+\right] \text{ as } \lambda \to \infty.$$

□

This concludes the proof of Theorem 6.

B.3.4 Proof of Theorem 7.

Proof. We first prove the 'only if' part. Let (n^λ, n_F^λ) be asymptotically optimal, and suppose for contradiction that it is not of the form stated in the theorem. That is, there exists $\epsilon > 0$ and a subsequence, which we index again by λ for convenience, satisfying

$$\min_{(a,b)\in \arg\min_{\beta,\beta_F} \hat{V}_p(\beta,\beta_F)} \frac{\left|n^\lambda - R^\lambda - a\sqrt{R^\lambda}\right| + \left|n_F^\lambda - b\sqrt{R^\lambda}\right|}{\sqrt{R^\lambda}} > \epsilon$$

for each λ. This subsequence is asymptotically optimal, and so it follows from the proof of Lemma 3 that

$$n_F^\lambda = b_\lambda\sqrt{R^\lambda} + o(\sqrt{R^\lambda}) \text{ and } n^\lambda = R^\lambda + a_\lambda\sqrt{R^\lambda} + o(\sqrt{R^\lambda})$$

for some bounded sequences $\{a_\lambda\}$ and $\{b_\lambda\}$. Then, there exist finite constants (a, b) and a subsequence indexed by λ', such that $(a, b) \notin \arg\min_{\beta,\beta_F} \hat{V}_p(\beta, \beta_F)$ and

$$a_{\lambda'} \to a \text{ and } b_{\lambda'} \to b \text{ as } \lambda' \to \infty.$$

For the ease of notation, we re-index this subsequence by λ. As $(a,b) \notin \arg\min_{\beta,\beta_F} \hat{V}_p(\beta,\beta_F)$, there exists (β,β_F) such that $\hat{V}_p(\beta,\beta_F) < \hat{V}_p(a,b)$. Define

$$\bar{n}_F^\lambda = \beta_F\sqrt{R^\lambda} + o(\sqrt{R^\lambda}) \text{ and } \bar{n}^\lambda = R^\lambda + \beta\sqrt{R^\lambda} + o(\sqrt{R^\lambda}).$$

Then,

$$\begin{aligned}
&\limsup_{\lambda\to\infty} \frac{\Pi^\lambda(n^\lambda, n_F^\lambda) - \Pi^{\lambda,*}}{\sqrt{\lambda}} \\
\geq &\limsup_{\lambda\to\infty} \frac{\Pi^\lambda(n^\lambda, n_F^\lambda) - \Pi^\lambda(\bar{n}^\lambda, \bar{n}_F^\lambda)}{\sqrt{\lambda}} \\
= &\limsup_{\lambda\to\infty} \left\{ \frac{2c(n^\lambda - R^\lambda) + c_F n_F^\lambda + (h+a\theta)\mathbb{E}[Q_\Sigma^\lambda(\infty; n^\lambda, n_F^\lambda)]}{\sqrt{\lambda}} \right. \\
&\left. - \frac{2c(\bar{n}^\lambda - R^\lambda) + c_F \bar{n}_F^\lambda + (h+a\theta)\mathbb{E}[Q_\Sigma^\lambda(\infty; \bar{n}^\lambda, \bar{n}_F^\lambda)])}{\sqrt{\lambda}} \right\} \\
= &\hat{V}_p(a,b) - \hat{V}_p(\beta,\beta_F) > 0
\end{aligned}$$

where the last equality follows from Theorem 6, contradicting asymptotic optimality.

It remains to prove the 'if' part. From the proof of the 'only if' part, the sequence of optimal staffing levels $(n^{\lambda,*}, n_F^{\lambda,*})$ satisfy

$$n_F^{\lambda,*} = d_\lambda\sqrt{R^\lambda} + o(\sqrt{R^\lambda}) \text{ and } n^{\lambda,*} = R^\lambda + c_\lambda\sqrt{R^\lambda} + o(\sqrt{R^\lambda})$$

for some $(c_\lambda, d_\lambda) \in \arg\min_{\beta,\beta_F} \hat{V}_p(\beta,\beta_F)$. Next, consider any sequence

$$n_F^\lambda = b_\lambda\sqrt{R^\lambda} + o(\sqrt{R^\lambda}) \text{ and } n^\lambda = R^\lambda + a_\lambda\sqrt{R^\lambda} + o(\sqrt{R^\lambda})$$

where $(a_\lambda, b_\lambda) \in \arg\min_{\beta,\beta_F} \hat{V}_p(\beta,\beta_F)$. Then,

$$
\begin{aligned}
&\limsup_{\lambda\to\infty} \frac{\Pi^{\lambda}(n^{\lambda}, n_F^{\lambda}) - \Pi^{\lambda,*}}{\sqrt{\lambda}} \\
=&\limsup_{\lambda\to\infty} \frac{\Pi^{\lambda}(n^{\lambda}, n_F^{\lambda}) - \Pi^{\lambda}(n^{\lambda,*}, n_F^{\lambda,*})}{\sqrt{\lambda}} \\
=&\limsup_{\lambda\to\infty} \left\{ \frac{2c(n^{\lambda} - R^{\lambda}) + c_F n_F^{\lambda} + (h + a\theta)\mathbb{E}[Q_{\Sigma}^{\lambda}(\infty; n^{\lambda}, n_F^{\lambda})]}{\sqrt{\lambda}} \right. \\
&\qquad \left. - \frac{2c(n^{\lambda,*} - R^{\lambda}) + c_F n_F^{\lambda,*} + (h + a\theta)\mathbb{E}[Q_{\Sigma}^{\lambda}(\infty; n^{\lambda,*}, n_F^{\lambda,*})]}{\sqrt{\lambda}} \right\} \qquad \text{(B.10)} \\
=&\hat{V}_p^* - \hat{V}_p^* = 0,
\end{aligned}
$$

where $\hat{V}_p^* = \min_{\beta,\beta_F} \hat{V}_p(\beta, \beta_F)$. To see (B.10), note that by Theorem 6, we have that for any $(a, b) \in \arg\min_{\beta,\beta_F} \hat{V}_p(\beta, \beta_F)$,

$$
\frac{2ca\sqrt{R^{\lambda}} + c_F b\sqrt{R^{\lambda}} + o(\sqrt{R^{\lambda}})}{\sqrt{\lambda}} + (h + a\theta)\mathbb{E}[\hat{Q}_{\Sigma}^{\lambda}(\infty; a, b)] = \hat{V}_p^* + o(1).
$$

Then, (B.10) follows because $\arg\min_{\beta,\beta_F} \hat{V}_p(\beta, \beta_F)$ is finite under Assumption 4. □

B.4 Proofs of the Results in Section 3.4

For $x, y \in \mathbb{R}$ and $z \geq 0$, define

$$
K_{\lambda}(x, y, z) = \tilde{\Pi}^{\lambda}\left(\frac{p_1\lambda + x\lambda^{\alpha_1}}{\mu}, \frac{p_2\lambda + y\lambda^{\alpha_2}}{\mu}, \frac{z\lambda^{\alpha_2}}{\mu_F}\right).
$$

B.4.1 Proof of Lemma 5.

Proof. In this case,

$$
K_{\lambda}(x, y, z) = c(p_1 + p_2)R^{\lambda} + \lambda^{\alpha}\left(\frac{c}{\mu}x + \frac{c}{\mu}y + \frac{c_F}{\mu_F}z\right) + c_P\lambda^{\alpha}\mathbb{E}\left[\left((Y_1 - x)^+ + (Y_2 - y)^+ - z\right)^+\right].
$$

In the first case, note that $K_\lambda(x, y, z)$ is convex and

$$\nabla K_\lambda(q_1, q_2, 0) = \lambda^\alpha \left(0, 0, \frac{c_F}{\mu_F} - c_P \mathbb{P}(Y_1 > q_1 \text{ or } Y_2 > q_2)\right).$$

As $\frac{c_F}{\mu_F} - c_P \mathbb{P}(Y_1 > q_1 \text{ or } Y_2 > q_2) \geq 0$, $(q_1, q_2, 0)$ is optimal.

In the second case, we have

$$\nabla K_\lambda(r_1, r_2, r_F) = (0, 0, 0).$$

The optimality of (r_1, r_2, r_F) follows due to the convexity of $K_\lambda(x, y, z)$. □

B.4.2 Proof of Lemma 6.

Proof. In this case,

$$\begin{aligned} K_\lambda(x, y, z) =& c(p_1 + p_2)R^\lambda + \lambda^{\alpha_1} \frac{c}{\mu} x + \lambda^{\alpha_2} \left(\frac{c}{\mu} y + \frac{c_F}{\mu_F} z\right) \\ &+ c_P \lambda^{\alpha_1} \mathbb{E}\left[\left((Y_1 - x)^+ + \lambda^{\alpha_2 - \alpha_1}(Y_2 - y)^+ - \lambda^{\alpha_2 - \alpha_1} z\right)^+\right]. \end{aligned}$$

Let $(x^*_\lambda, y^*_\lambda, z^*_\lambda)$ be the minimizer of K_λ.

We first show that $x^*_\lambda = q_1 + o(1)$. Note that $K^*_\lambda := \tilde{\Pi}^{\lambda,*} \leq K_\lambda(0, 0, 0) = c(p_1 + p_2)R^\lambda + O(\lambda^{\alpha_1})$. Since $K_\lambda(x, y, z) \geq c(p_1 + p_2)R^\lambda + \lambda^{\alpha_1} \frac{c}{\mu} x$, we have that $x^{*+}_\lambda = O(1)$.

Now suppose for contradiction that there exists a subsequence, indexed by λ_k, such that either i) $x^*_{\lambda_k} \to -\infty$ or ii) $x^*_{\lambda_k} \to C \in \mathbb{R} \backslash \{q_1\}$. Note that

$$\begin{aligned} K_\lambda(x, y, z) &\geq c(p_1 + p_2)R^\lambda + \lambda^{\alpha_1} \frac{c}{\mu} x + \lambda^{\alpha_2} \frac{c_F}{\mu_F} z + c_P \lambda^{\alpha_1} \mathbb{E}[(Y_1 - x)^+ - \lambda^{\alpha_2 - \alpha_1} z)^+] \\ &= c(p_1 + p_2)R^\lambda + \lambda^{\alpha_1} \frac{c}{\mu} x + \lambda^{\alpha_2} \frac{c_F}{\mu_F} z + c_P \lambda^{\alpha_1} \mathbb{E}[(Y_1 - x - \lambda^{\alpha_2 - \alpha_1} z)^+] \\ &= c(p_1 + p_2)R^\lambda + \lambda^{\alpha_1} \frac{c}{\mu}(x + \lambda^{\alpha_2 - \alpha_1} z) + c_P \lambda^{\alpha_1} \mathbb{E}[(Y_1 - x - \lambda^{\alpha_2 - \alpha_1} z)^+] + \lambda^{\alpha_2} \left(\frac{c_F}{\mu_F} - \frac{c}{\mu}\right) z. \end{aligned} \tag{B.11}$$

First suppose that $x^*_{\lambda_k} \to -\infty$. Since $c_P > c_F/\mu_F > c/\mu$ and $K^*_\lambda \le c(p_1+p_2)R^\lambda + O(\lambda^{\alpha_1})$, it follows that $\left(x^*_\lambda + \lambda^{\alpha_2-\alpha_1} z^*_\lambda\right)^- = O(1)$. This in turn implies that $\lambda^{\alpha_2-\alpha_1} z^*_\lambda \to \infty$, so that $\lambda^{\alpha_2}\left(\frac{c_F}{\mu_F} - \frac{c}{\mu}\right) z^*_\lambda$ grows to infinity faster than $O(\lambda^{\alpha_1})$. Then, the second and third terms of the last equation (B.11) will be $O(\lambda^{\alpha_1})$, while the last is of a larger order. This contradicts that $(x^*_\lambda, y^*_\lambda, z^*_\lambda)$ is optimal.

Consider the second case $x^*_\lambda \to C \in \mathbb{R}\backslash\{q_1\}$. Note that

$$K_\lambda(q_1, 0, 0) = c(p_1 + p_2)R^\lambda + \lambda^{\alpha_1} f(q_1)$$

where $f(x) = \frac{c}{\mu}x + c_P \mathbb{E}[(Y_1 - x)^+]$. From (B.11), we have that

$$K_\lambda(x^*_\lambda, y^*_\lambda, z^*_\lambda) \ge c(p_1 + p_2)R^\lambda + \lambda^{\alpha_1} f(x^*_\lambda + \lambda^{\alpha_2-\alpha_1} z^*_\lambda) + \lambda^{\alpha_2}\left(\frac{c_F}{\mu_F} - \frac{c}{\mu}\right) z^*_\lambda.$$

Since q_1 is uniquely optimal for f and $x^*_\lambda \to C$, we must have $\lambda^{\alpha_2-\alpha_1} z^*_\lambda \to q_1 - C > 0$. But then $\lambda^{\alpha_2}\left(\frac{c_F}{\mu_F} - \frac{c}{\mu}\right) z^*_\lambda \ne o(\lambda^{\alpha_1})$, contradicting optimality.

This completes the proof that $x^*_\lambda = q_1 + o(1)$. Next, we prove that y^*_λ and z^*_λ are of the appropriate form. We first show they are $O(1)$. Consider the partial derivatives

$$\frac{\partial K_\lambda}{\partial y} = \lambda^{\alpha_2}\left(\frac{c}{\mu} - c_P \mathbb{P}[Y_2 > y, \lambda^{\alpha_1-\alpha_2}(Y_1 - x)^+ + Y_2 - y > z]\right)$$

and

$$\frac{\partial K_\lambda}{\partial z} = \lambda^{\alpha_2}\left(\frac{c_F}{\mu_F} - c_P \mathbb{P}[\lambda^{\alpha_1-\alpha_2}(Y_1 - x)^+ + (Y_2 - y)^+ > z]\right).$$

By optimality, we have

$$0 < \frac{c}{c_P \mu} = \mathbb{P}[Y_2 > y^*_\lambda, \lambda^{\alpha_1-\alpha_2}(Y_1 - x^*_\lambda)^+ + Y_2 - y^*_\lambda > z^*_\lambda] \le \mathbb{P}[Y_2 > y^*_\lambda]$$

which implies that $y^{*+}_\lambda = O(1)$. If $y^{*-}_\lambda \ne O(1)$, then there is a subsequence (re-indexed by λ) such

that $y_\lambda^* \to -\infty$, which implies that

$$\begin{aligned} 1 > \frac{c}{c_P \mu} &= \mathbb{P}[Y_2 > y_\lambda^*, \lambda^{\alpha_1-\alpha_2}(Y_1 - x_\lambda^*)^+ + Y_2 - y_\lambda^* > z_\lambda^*] \\ &= \mathbb{P}[\lambda^{\alpha_1-\alpha_2}(Y_1 - x_\lambda^*)^+ + Y_2 - y_\lambda^* > z_\lambda^*] + o(1) \\ &\geq \mathbb{P}[Y_2 > y_\lambda^* + z_\lambda^*] + o(1). \end{aligned}$$

This in turn implies $z_\lambda^* \to \infty$, and in particular $z_\lambda^* > 0$. But then

$$\begin{aligned} \mathbb{P}[\lambda^{\alpha_1-\alpha_2}(Y_1 - x_\lambda^*)^+ + (Y_2 - y_\lambda^*)^+ > z_\lambda^*] &= \mathbb{P}[Y_2 > y_\lambda^*, \lambda^{\alpha_1-\alpha_2}(Y_1 - x_\lambda^*)^+ + Y_2 - y_\lambda^* > z_\lambda^*] + o(1) \\ &= \frac{c}{c_P \mu} + o(1) < \frac{c_F}{c_P \mu_F} \end{aligned}$$

so that $\frac{\partial K_\lambda}{\partial z}(x_\lambda^*, y_\lambda^*, z_\lambda^*) > 0$, contradicting optimality. Hence, $y_\lambda^* = O(1)$.

Next, we show $z_\lambda^* = O(1)$. If not, we can obtain a subsequence indexed again by λ such that $z_\lambda^* \to \infty$, and in particular $z_\lambda^* > 0$. Since $y_\lambda^* = O(1)$ and $x_\lambda^* = q_1 + o(1)$, we have

$$\begin{aligned} \mathbb{P}[\lambda^{\alpha_1-\alpha_2}(Y_1 - x_\lambda^*)^+ + (Y_2 - y_\lambda^*)^+ > z_\lambda^*] &= \mathbb{P}[\lambda^{\alpha_1-\alpha_2}(Y_1 - q_1)^+ > z_\lambda^*] + o(1) \\ &\leq \mathbb{P}[Y_1 > q_1] + o(1) = \frac{c}{c_P \mu} + o(1) \end{aligned}$$

so that $\frac{\partial K_\lambda}{\partial z}(x_\lambda^*, y_\lambda^*, z_\lambda^*) > 0$ contradicting optimality. Thus $z_\lambda^* = O(1)$.

Finally, we show that y_λ^* and z_λ^* have the right asymptotics. Suppose $\tilde{n}_2^*$ and $\tilde{n}_F^*$ are not of the specified form. First suppose $\mathbb{P}(Y_1 > q_1 \text{ or } Y_2 > q_2) > \frac{c_F}{c_P \mu_F}$. Then, there is a subsequence re-indexed by λ such that $x_\lambda^* \to q_1$, $y_\lambda^* \to D \in \mathbb{R}$ and $z_\lambda^* \to E \geq 0$, where either $D \neq l$ or $E \neq l_F$. Note that

$$\begin{aligned} \mathbb{P}[Y_2 > y_\lambda^*, \lambda^{\alpha_1-\alpha_2}(Y_1 - x_\lambda^*)^+ + Y_2 - y_\lambda^* > z_\lambda^*] &= \mathbb{P}[Y_2 > D, Y_1 > q_1 \text{ or } Y_2 > D + E] + o(1) \\ &= \mathbb{P}[Y_2 > D + E \text{ or } (Y_2 > D, Y_1 > q_1)] + o(1) \end{aligned}$$

and

$$\mathbb{P}[\lambda^{\alpha_1-\alpha_2}(Y_1 - x_\lambda^*)^+ + (Y_2 - y_\lambda^*)^+ > z_\lambda^*] = \mathbb{P}[Y_1 > q_1 \text{ or } Y_2 > D + E] + o(1).$$

By optimality, we must have either (i) the first probability is $\frac{c}{c_P\mu}$ and the second is $\frac{c_F}{c_P\mu_F}$ or (ii) the first probability is $\frac{c}{c_P\mu}$, the second is $\leq \frac{c_F}{c_P\mu_F}$ and $E = 0$. Case (i) is ruled out by the uniqueness of l and l_F. If (ii), then $D > q_2$ in order for the second probability to be $\leq \frac{c_F}{c_P\mu_F}$, but $D = q_2$ is necessary for the first probability to be $\frac{c}{c_P\mu}$. This is a contradiction and thus $y_\lambda^* = l + o(1)$ and $z_\lambda^* = l_F + o(1)$.

We now turn to the other case $\mathbb{P}(Y_1 > q_1 \text{ or } Y_2 > q_2) < \frac{c_F}{c_P\mu_F}$. Using the previous notation, we can again obtain a subsequence such that either $D \neq q_2$ or $E > 0$. The case $E = 0$ and $D \neq q_2$ can be ruled out by our previous discussion, so suppose $E > 0$. By optimality, we must have that

$$\mathbb{P}[Y_2 > D + E \text{ or } (Y_2 > D, Y_1 > q_1)] = \frac{c}{c_P\mu}$$

and

$$\mathbb{P}[Y_1 > q_1 \text{ or } Y_2 > D + E] = \frac{c_F}{c_P\mu_F}.$$

The second equation ensures $D + E < q_2$. Then, the first probability is at least $\mathbb{P}(Y_2 > D + E) > \mathbb{P}(Y_2 > q_2) = \frac{c}{c_P\mu}$, a contradiction. This completes the proof. □

B.4.3 Two auxiliary lemmas

Note that for any given arrival rate realization, under the scheduling policy $\tilde{v}^\lambda$, the two-class queue can be decomposed into two independent single-class queues with two types of servers: the high-priority rate-μ servers and the low-priority rate-μ_F servers. In this section, we study the single-class queue with arrival rate γ, n high-priority rate-μ servers, n_F low-priority rate-μ_F servers, and abandonment rate $\theta = \mu_F$. To simplify the notation, we denote by X the steady-state number of customers in the system, Q the steady-state number of customers waiting in queue, Z the

steady-state number of customers in service with rate-μ servers, and Z_F the steady-state number of customers in service with rate-μ_F servers.

Lemma 24. *For the single-class queue with two types of servers and $\theta = \mu_F$, there are universal constants $K_1, K_2, K_3 > 0$ (i.e. not depending on n, n_F or γ), such that*

$$\theta\mathbb{E}[Q] \leq (\gamma - n\mu - n_F\mu_F)^+ + K_1\sqrt{\gamma}\exp(-\frac{K_2}{\gamma}(\gamma - n\mu - n_F\mu_F)^2) + K_3.$$

Proof. We start by showing that

$$\mathbb{E}[X] = \gamma/\theta - n\mu/\theta + n + (\mu - \theta)\mathbb{E}[(n - X)^+]/\theta. \tag{B.12}$$

Let $\xi(x)$ denote the death rate at state x. When $x > n$, $\xi(x) = x\theta + n(\mu - \theta)$; when $x \leq n$, $\xi(x) = x\mu = x\theta + x(\mu - \theta)$. Equating the birth rate and the death rate in stationarity, we have

$$\gamma = \mathbb{E}[X\theta + n(\mu - \theta) - (n - X)^+(\mu - \theta)],$$

which implies (B.12).

First, consider the case where $\gamma \geq n\mu + n_F\mu_F$. We have

$$\begin{aligned}
\mathbb{E}[Q] &= \mathbb{E}[(X - n - n_F)^+] \\
&= \mathbb{E}[X] - n - n_F + \sum_{x=0}^{n+n_F-1} \mathbb{P}[X \leq x] \\
&= \gamma/\theta - n\mu/\theta + n + (\mu - \theta)\mathbb{E}[(n - X)^+]/\theta - n - n_F + \sum_{x=0}^{n+n_F-1} \mathbb{P}[X \leq x] \text{ by (B.12)} \\
&= (\gamma/\theta - n\mu/\theta - n_F) + \frac{\mu - \theta}{\theta}\sum_{x=0}^{n-1}\mathbb{P}[X \leq x] + \sum_{x=0}^{n+n_F-1} \mathbb{P}[X \leq x].
\end{aligned}$$

It suffices to bound $\sum_{x=0}^{n+n_F-1} \mathbb{P}[X \leq x]$. Let $M_1 = \lfloor \gamma/\theta - n\mu/\theta + n \rfloor$, which is the mode of X (to see this, note that the death rate $\xi(x)$ is increasing in x, and that the birth rate γ is larger than the

death rate for $x < M_1$, and is smaller than the death rate for $x > M_1$). We then have $M_1 \geq n + n_F$, and $\xi(M_1) = n\mu + \lfloor \gamma/\theta - n\mu/\theta \rfloor \theta \leq \gamma$. For $0 < k \leq M_1 - n + 1$, note that

$$\begin{aligned}
\mathbb{P}[X = M_1 - k] &= \mathbb{P}[X = M_1] \cdot \frac{\xi(M_1)(\xi(M_1) - \theta) \cdots (\xi(M_1) - (k-1)\theta)}{\gamma^k} \\
&\leq \mathbb{P}[X = M_1] \cdot \frac{\gamma(\gamma - \theta) \cdots (\gamma - (k-1)\theta)}{\gamma^k} \\
&= \mathbb{P}[X = M_1] \cdot 1(1 - \theta/\gamma) \cdots (1 - (k-1)\theta/\gamma) \\
&\leq \mathbb{P}[X = M_1] \cdot (1 - (k-1)\theta/2\gamma)^k \\
&\leq \mathbb{P}[X = M_1] \cdot \exp(-(k-1)k\theta/2\gamma) \text{ as } 1 - x \leq \exp(-x).
\end{aligned}$$

Similarly, for $M_1 - n + 1 < k \leq M_1$,

$$\begin{aligned}
&\mathbb{P}[X = M_1 - k] \\
=&\mathbb{P}[X = M_1] \cdot \frac{1}{\gamma^k} \xi(M_1)(\xi(M_1) - \theta) \cdots (\xi(M_1) - (M_1 - n)\theta) \cdot (\xi(M_1) - (M_1 - n)\theta - \mu) \\
&\cdots (\xi(M_1) - (M_1 - n)\theta - (k - M_1 + n - 1)\mu) \\
\leq&\mathbb{P}[X = M_1] \times \\
&\frac{\xi(M_1)(\xi(M_1) - \theta) \cdots (\xi(M_1) - (M_1 - n)\theta) \cdot (\xi(M_1) - (M_1 - n + 1)\theta) \cdots (\xi(M_1) - (k-1)\theta)}{\gamma^k} \\
\leq&\mathbb{P}[X = M_1] \cdot (1 - (k-1)\theta/2\gamma)^k \\
\leq&\mathbb{P}[X = M_1] \cdot \exp(-(k-1)k\theta/2\gamma).
\end{aligned}$$

Choose $A_1 > 0$ such that $A_1 k^2 \leq k(k-1)\theta/2$ for all $k \geq 2$, then for any $1 < k \leq M_1$

$$\mathbb{P}[X = M_1 - k] \leq \mathbb{P}[X = M_1] \cdot \exp(-A_1 k^2/\gamma).$$

Next, we bound $\mathbb{P}[X = M_1]$. Note that for $k > 0$,

$$\mathbb{P}[X = M_1 + k] = \mathbb{P}[X = M_1] \cdot \frac{\gamma^k}{(\xi(M_1) + \theta) \cdots (\xi(M_1) + k\theta)}.$$

Then, because

$$\begin{aligned}\frac{\gamma^{\lfloor\sqrt{\gamma}\rfloor}}{(\xi(M_1)+\theta)\cdots(\xi(M_1)+\lfloor\sqrt{\gamma}\rfloor\theta)} &\geq \frac{\gamma^{\lfloor\sqrt{\gamma}\rfloor}}{(\gamma+\theta)\cdots(\gamma+\lfloor\sqrt{\gamma}\rfloor\theta)}\\ &\geq \frac{\gamma^{\lfloor\sqrt{\gamma}\rfloor}}{(\gamma+\lfloor\sqrt{\gamma}\rfloor\theta)^{\lfloor\sqrt{\gamma}\rfloor}}\\ &= \left(1-\frac{\lfloor\sqrt{\gamma}\rfloor\theta}{\gamma+\lfloor\sqrt{\gamma}\rfloor\theta}\right)^{\lfloor\sqrt{\gamma}\rfloor}\\ &\geq \left(1-\frac{\lfloor\sqrt{\gamma}\rfloor\theta}{\gamma}\right)^{\lfloor\sqrt{\gamma}\rfloor} \to \exp(-\theta) \text{ as } \gamma\to\infty,\end{aligned}$$

for γ large enough and $1 \leq k \leq \lfloor\sqrt{\gamma}\rfloor$,

$$\mathbb{P}[X = M_1 + k] \geq \mathbb{P}[X = M_1]\cdot\exp(-\theta)/2.$$

Thus, for γ large enough,

$$1 \geq \sum_{k=M_1}^{M_1+\lfloor\sqrt{\gamma}\rfloor} \mathbb{P}[X=k] \geq \mathbb{P}[X=M_1]\cdot(1+\lfloor\sqrt{\gamma}\rfloor\exp(-\theta)/2)$$

which implies that

$$\mathbb{P}[X=M_1] \leq \frac{1}{1+\lfloor\sqrt{\gamma}\rfloor\exp(-\theta)/2} \leq \frac{B_1}{\sqrt{\gamma}},$$

where $B_1 = 2\exp(\theta)$.

Above all, we have proven that for γ large enough, when $1 < k \leq M_1$,

$$\mathbb{P}[X = M_1 - k] \leq B_1\cdot\exp(-A_1k^2/\gamma)/\sqrt{\gamma}.$$

Then, for $1 < k \leq n + n_F$,

$$\begin{aligned}
\mathbb{P}[X \leq n + n_F - k] &= \mathbb{P}[X \leq M_1 - (M_1 - n - n_F + k)] \\
&= \sum_{j=M_1-n-n_F+k}^{M_1} \mathbb{P}[X = M_1 - j] \\
&\leq \sum_{j=M_1-n-n_F+k}^{M_1} B_1 \exp(-A_1 j^2/\gamma)/\sqrt{\gamma} \\
&\leq \int_{M_1-n-n_F+k-1}^{\infty} B_1 \exp(-A_1 j^2/\gamma)/\sqrt{\gamma}\, dj \\
&\leq \frac{B_1\sqrt{2\pi}}{2\sqrt{A_1}} \exp(-A_1(M_1 - n - n_F + k - 1)^2/(2\gamma)) \text{ by Chernoff-Cramer bound.}
\end{aligned}$$

Choose $D_1 > 0$ such that $D_1 x^2 \leq A_1(x-1)^2/2$ for all $x \geq 2$. Then, for γ large enough, and $2 \leq k \leq n + n_F$,

$$\mathbb{P}[X \leq n + n_F - k] \leq C_1 \exp(-D_1(M - n - n_F + k)^2/\gamma)$$

where $C_1 = \frac{B_1\sqrt{2\pi}}{2\sqrt{A_1}}$.

Finally, we have for γ large enough

$$\begin{aligned}
\sum_{x=0}^{n+n_F-1} \mathbb{P}[X \leq x] &\leq \int_0^{n+n_F} \mathbb{P}[X \leq n + n_F - x]\, dx \\
&\leq \int_2^{n+n_F} \mathbb{P}[X \leq n + n_F - x]\, dx + 2 \\
&\leq \int_0^{n+n_F} C_1 \exp(-D_1(M_1 - n - n_F + x)^2/\gamma)\, dx + 2 \\
&= \int_{M_1-n-n_F}^{M_1} C_1 \exp(-D_1 x^2/\gamma)\, dx + 2 \\
&\leq \frac{C_1\sqrt{2\pi\gamma}}{2\sqrt{D_1}} \exp(-D_1(M_1 - n - n_F)^2/(2\gamma)) + 2.
\end{aligned}$$

Since $M_1 - n - n_F = \lfloor \frac{\gamma - n\mu - n_F\mu_F}{\theta} \rfloor$, this completes the proof for $\gamma > n\mu + n_F\mu_F$.

Next, consider the case where $n\mu \leq \gamma < n\mu + n_F\mu_F$. Let $M_2 = n + \lfloor \frac{\gamma - n\mu}{\theta} \rfloor$ be the mode of X

in this case. Note that

$$n + n_F - M_2 \geq \frac{n\mu + n_F\mu_F - \gamma}{\theta}.$$

Thus, for any $C > 0$,

$$\exp(-C\theta^2(n + n_F - M_2)^2) \leq \exp(-C(n\mu + n_F\mu_F - \gamma)^2).$$

Next, note that

$$\mathbb{E}[Q] = \mathbb{E}[(X - n - n_F)^+] = \sum_{x=n+n_F}^{\infty} \mathbb{P}[X > x].$$

As $\xi(M_2) = n\mu + \lfloor \frac{\gamma - n\mu}{\theta} \rfloor \theta$, we have for $k > 0$,

$$\begin{aligned}
\mathbb{P}[X = M_2 + k] &= \mathbb{P}[X = M_2] \cdot \frac{\gamma^k}{(\xi(M_2) + \theta) \cdots (\xi(M_2) + k\theta)} \\
&\leq \mathbb{P}[X = M_2] \cdot \frac{\gamma^k}{\gamma(\gamma + \theta) \cdots (\gamma + (k-1)\theta)} \\
&\leq \mathbb{P}[X = M_2] \cdot \left(\frac{\gamma}{\gamma + (k-1)\theta/2} \right)^{k/2} \\
&= \mathbb{P}[X = M_2] \cdot \left(1 - \frac{(k-1)\theta/2}{\gamma + (k-1)\theta/2} \right)^{k/2} \\
&\leq \mathbb{P}[X = M_2] \cdot \exp\left(-\frac{k(k-1)\theta/4}{\gamma + (k-1)\theta/2} \right) \\
&\leq \mathbb{P}[X = M_2] \cdot \left(\exp\left(-\frac{k(k-1)\theta/4}{2\gamma} \right) + \exp\left(-\frac{k(k-1)\theta/4}{(k-1)\theta} \right) \right) \\
&= \mathbb{P}[X = M_2] \cdot \left(\exp\left(-\frac{k(k-1)\theta}{8\gamma} \right) + \exp\left(-k/4 \right) \right).
\end{aligned}$$

The last inequality comes from the fact that $\frac{k(k-1)\theta/4}{\gamma + (k-1)\theta/2} \geq \frac{k(k-1)\theta/4}{2\gamma}$ if $\gamma \geq (k-1)\theta/2$ and

$\frac{k(k-1)\theta/4}{\gamma+(k-1)\theta/2} \geq \frac{k(k-1)\theta/4}{(k-1)\theta}$ otherwise. Thus, for all $k > 0$,

$$\begin{aligned}
\mathbb{P}[X \geq M_2 + k] &= \sum_{j=k}^{\infty} \mathbb{P}[X = M_2 + j] \\
&\leq \sum_{j=k}^{\infty} \mathbb{P}[X = M_2] \cdot \left(\exp\left(-\frac{j(j-1)\theta}{8\gamma} \right) + \exp\left(-j/4\right) \right) \\
&\leq \mathbb{P}[X = M_2] \cdot \left(\int_{k-1}^{\infty} \exp\left(-\frac{j(j-1)\theta}{8\gamma} \right) \, dj + \frac{\exp(-k/4)}{1 - \exp(-1/4)} \right).
\end{aligned}$$

Choose $B_2 > 0$ such that $B_2 x^2 \leq x(x-1)\theta/8$ for all $x \geq 2$, and choose $F_2 > 0$ such that $F_2 x^2 \leq B_2(x-1)^2/2$ for all $x \geq 3$. Also, choose $A_2 > 0$ such that $\mathbb{P}[X = M_2] \leq A_2/\sqrt{\gamma}$ (the existence of A_2 follows similarly to before.) Then, for all $k \geq 3$,

$$\begin{aligned}
\mathbb{P}[X \geq M_2 + k] &\leq \mathbb{P}[X = M_2] \cdot \left(\int_{k-1}^{\infty} \exp\left(-\frac{j(j-1)\theta}{8\gamma} \right) \, dj + \frac{\exp(-k/4)}{1 - \exp(-1/4)} \right) \\
&\leq \frac{A_2}{\sqrt{\gamma}} \left(\int_{k-1}^{\infty} \exp\left(-B_2 j^2/\gamma \right) \, dj + \frac{\exp(-k/4)}{1 - \exp(-1/4)} \right) \\
&\leq \frac{A_2\sqrt{2\pi}}{2\sqrt{B_2}} \cdot \exp(-B_2(k-1)^2/2\gamma) + C_2 \exp(-D_2 k)/\sqrt{\gamma} \\
&\leq E_2 \exp(-F_2 k^2/\gamma) + C_2 \exp(-D_2 k)/\sqrt{\gamma}
\end{aligned}$$

for some universal $C_2, D_2, E_2 > 0$. Finally,

$$\begin{aligned}
\sum_{x=n+n_F}^{\infty} \mathbb{P}[X > x] &= \int_{n+n_F-M_2}^{\infty} \mathbb{P}[X > M_2 + x] \, dx \\
&\leq 3 + \int_{n+n_F-M_2+3}^{\infty} E_2 \exp(-F_2 x^2/\gamma) + C_2 \exp(-D_2 x)/\sqrt{\gamma} \, dx \\
&\leq 3 + \int_{n+n_F-M_2}^{\infty} E_2 \exp(-F_2 x^2/\gamma) + C_2 \exp(-D_2 x)/\sqrt{\gamma} \, dx \\
&\leq 3 + L_2\sqrt{\gamma} \exp(-F_2(n + n_F - M_2)^2/2\gamma) + G_2 \exp(-D_2(n + n_F - M_2))/\sqrt{\gamma} \\
&\leq L_2\sqrt{\gamma} \exp(-F_2(n + n_F - M_2)^2/2\gamma) + 3 + G_2
\end{aligned}$$

for some universal $L_2, G_2 > 0$.

Lastly, consider the case where $0 < \gamma < n\mu$. This is very similar to the proof of the case $n\mu \le \gamma < n\mu + n_F\mu_F$, but we include it here for completeness. Let $M_3 = \lfloor \gamma/\mu \rfloor$ be the mode of X in this case. Note that

$$n + n_F - M_3 \ge \frac{n\mu + n_F\mu - \gamma}{\mu} \ge \frac{n\mu + n_F\mu_F - \gamma}{\mu}.$$

Thus, for any $C > 0$,

$$\exp(-C\mu^2(n + n_F - M_3)^2) \le \exp(-C(n\mu + n_F\mu_F - \gamma)^2).$$

Next, note that

$$\mathbb{E}[Q] = \mathbb{E}[(X - n - n_F)^+] = \sum_{x=n+n_F}^{\infty} \mathbb{P}[X > x].$$

For $\xi(M_3) = \lfloor \gamma/\mu \rfloor \mu$, we have for $0 < k \le n - M_3$,

$$\mathbb{P}[X = M_3 + k] = \mathbb{P}[X = M_3] \cdot \frac{\gamma^k}{(\xi(M_3) + \mu) \cdots (\xi(M_3) + k\mu)},$$

and for $k > n - M_3$,

$$\mathbb{P}[X = M_3 + k] \le \mathbb{P}[X = M_3] \cdot \frac{\gamma^k}{(\xi(M_3) + \mu) \cdots (\xi(M_3) + k\mu)}.$$

Thus, for all $k > 0$, we have

$$\begin{aligned}
\mathbb{P}[X = M_3 + k] &\le \mathbb{P}[X = M_3] \cdot \frac{\gamma^k}{(\xi(M_3) + \mu) \cdots (\xi(M_3) + k\mu)} \\
&\le \mathbb{P}[X = M_3] \cdot \frac{\gamma^k}{\gamma(\gamma + \mu) \cdots (\gamma + (k-1)\mu)} \\
&\le \mathbb{P}[X = M_3] \cdot \left(\frac{\gamma}{\gamma + (k-1)\mu/2}\right)^{k/2} \\
&= \mathbb{P}[X = M_3] \cdot \left(1 - \frac{(k-1)\mu/2}{\gamma + (k-1)\mu/2}\right)^{k/2} \\
&\le \mathbb{P}[X = M_3] \cdot \exp\left(-\frac{k(k-1)\mu/4}{\gamma + (k-1)\mu/2}\right) \\
&\le \mathbb{P}[X = M_3] \cdot \left(\exp\left(-\frac{k(k-1)\mu/4}{2\gamma}\right) + \exp\left(-\frac{k(k-1)\mu/4}{(k-1)\mu}\right)\right) \\
&= \mathbb{P}[X = M_3] \cdot \left(\exp\left(-\frac{k(k-1)\mu}{8\gamma}\right) + \exp\left(-k/4\right)\right).
\end{aligned}$$

The last inequality comes from the fact that $\frac{k(k-1)\mu/4}{\gamma+(k-1)\mu/2} \ge \frac{k(k-1)\mu/4}{2\gamma}$ if $\gamma \ge (k-1)\mu/2$ and $\frac{k(k-1)\mu/4}{\gamma+(k-1)\mu/2} \ge \frac{k(k-1)\mu/4}{(k-1)\mu}$ otherwise. Thus, for all $k > 0$,

$$\begin{aligned}
\mathbb{P}[X \ge M_3 + k] &= \sum_{j=k}^{\infty} \mathbb{P}[X = M_3 + j] \\
&\le \sum_{j=k}^{\infty} \mathbb{P}[X = M_3] \cdot \left(\exp\left(-\frac{j(j-1)\mu}{8\gamma}\right) + \exp\left(-j/4\right)\right) \\
&\le \mathbb{P}[X = M_3] \cdot \left(\int_{k-1}^{\infty} \exp\left(-\frac{j(j-1)\mu}{8\gamma}\right) dj + \frac{\exp(-k/4)}{1 - \exp(-1/4)}\right).
\end{aligned}$$

Choose $B_3 > 0$ such that $B_3 x^2 \le x(x-1)\mu/8$ for all $x \ge 2$, and choose $F_3 > 0$ such that $F_3 x^2 \le B_3(x-1)^2/2$ for all $x \ge 3$. Also, choose $A_3 > 0$ such that $\mathbb{P}[X = M_3] \le A_3/\sqrt{\gamma}$. Then,

for all $k \geq 3$,

$$\begin{aligned}
\mathbb{P}[X \geq M_3 + k] &\leq \mathbb{P}[X = M_3] \cdot \left(\int_{k-1}^{\infty} \exp\left(-\frac{j(j-1)\mu}{8\gamma} \right) dj + \frac{\exp(-k/4)}{1 - \exp(-1/4)} \right) \\
&\leq \frac{A_3}{\sqrt{\gamma}} \left(\int_{k-1}^{\infty} \exp\left(-B_3 j^2/\gamma \right) dj + \frac{\exp(-k/4)}{1 - \exp(-1/4)} \right) \\
&\leq \frac{A_3\sqrt{2\pi}}{2\sqrt{B_3}} \cdot \exp(-B_3(k-1)^2/2\gamma) + C_3 \exp(-D_3 k)/\sqrt{\gamma} \\
&\leq E_3 \exp(-F_3 k^2/\gamma) + C_3 \exp(-D_3 k)/\sqrt{\gamma}
\end{aligned}$$

for some universal $C_3, D_3, E_3 > 0$. Finally,

$$\begin{aligned}
\sum_{x=n+n_F}^{\infty} \mathbb{P}[X > x] &= \int_{n+n_F-M_3}^{\infty} \mathbb{P}[X > M_3 + x]\, dx \\
&\leq 3 + \int_{n+n_F-M_3+3}^{\infty} E_3 \exp(-F_3 x^2/\gamma) + C_3 \exp(-D_3 x)/\sqrt{\gamma}\, dx \\
&\leq 3 + \int_{n+n_F-M_3}^{\infty} E_3 \exp(-F_3 x^2/\gamma) + C_3 \exp(-D_3 x)/\sqrt{\gamma}\, dx \\
&\leq 3 + L_3\sqrt{\gamma} \exp(-F_3(n + n_F - M_3)^2/2\gamma) + G_3 \exp(-D_3(n + n_F - M_3))/\sqrt{\gamma} \\
&\leq L_3\sqrt{\gamma} \exp(-F_3(n + n_F - M_3)^2/2\gamma) + 3 + G_3
\end{aligned}$$

for some universal $L_3, G_3 > 0$. This completes the proof. □

Next, consider two single-class queues, A and B, with common arrival rate λ and abandonment rate θ. System A has m high priority rate-μ servers and m_F low priority rate-μ_F servers, while system B has m/r high priority rate-$r\mu$ servers and m_F/r low priority rate-$r\mu_F$ servers, for some $r > 1$. Let $Q^A(\infty)$ and $Q^B(\infty)$ denote the stationary queue lengths of the two systems respectively.

Lemma 25. *For systems A and B,*

$$Q^A(\infty) \leq_{st} Q^B(\infty).$$

Proof. The proof follows the same lines of argument as the proof of Lemma 2 in [7], and is only

included for completeness. Let $X^A(t)$ and $X^B(t)$ denote the headcount processes, which are birth-death processes. Let $\alpha = m + m_F$ and $\beta = (m + m_F)/r$ be the number of servers in the two systems, and let $\xi^A(\cdot)$ and $\xi^B(\cdot)$ denote the death rates.

We first note that the two systems have the same birth rates. For $x \geq 0$, $\xi^A(\alpha + x) = \xi^B(\beta + x)$, and for $0 \leq x \leq \beta$, $\xi^A(\alpha - x) \geq \xi^B(\beta - x)$. Initialize $X^A(0) = \alpha$, $X^B(0) = \beta$. Couple the two systems such that (i) the arrivals to both systems coincide, and (ii) the departures in system B is a subset of the departures in system A. Then, for all $t \geq 0$, $X^A(t) - \alpha \leq X^B(t) - \beta$ and

$$Q^A(t) = (X^A(t) - \alpha)^+ \leq (X^B(t) - \beta)^+ = Q^B(t).$$

As the stationary distribution is well-defined, we have the stochastic dominance of the stationary distribution. □

B.4.4 Proof of Lemma 7

Proof. To simplify the notation, we drop the superscript λ and the '(∞)'. In particular, let X_i denote the stationary number of Class i customers in the system, Q_Σ denote the stationary total queue length, Z_i denote the stationary number of Class i customers served by the dedicated servers, and Z_{Fi} denote the stationary number of Class i customers served by the flexible servers.

We first prove the lower bound. *Consider the case where* $\theta \leq \mu_F$. Note that $Q_\Sigma \geq f(X_1, X_2) := ((X_1 - n_1)^+ + (X_2 - n_2)^+ - n_F)^+$. As f is convex, by Jensen's inequality

$$\mathbb{E}[Q_\Sigma | \Lambda = \gamma; \nu] \geq \mathbb{E}[f(X_1, X_2) | \Lambda = \gamma; \nu] \geq f(\mathbb{E}[X_1 | \Lambda = \gamma; \nu], \mathbb{E}[X_2 | \Lambda = \gamma; \nu]).$$

Thus,

$$\theta\mathbb{E}[Q_\Sigma | \Lambda = \gamma; \nu] \geq \left(\theta\mathbb{E}[X_1 | \Lambda = \gamma; \nu] - \theta n_1)^+ + (\theta\mathbb{E}[X_2 | \Lambda = \gamma; \nu] - \theta n_2)^+ - \theta n_F\right)^+.$$

Equating the arrival and departure rates in stationarity, we have

$$\gamma_i = \theta\mathbb{E}[X_i|\Lambda = \gamma; \nu] + (\mu - \theta)\mathbb{E}[Z_i|\Lambda = \gamma; \nu] + (\mu_F - \theta)\mathbb{E}[Z_{Fi}|\Lambda = \gamma; \nu].$$

Because $\mathbb{E}[Z_i|\Lambda = \gamma; \nu] \le n_i$ and $\mathbb{E}[Z_{F1}|\Lambda = \gamma; \nu] + \mathbb{E}[Z_{F2}|\Lambda = \gamma; \nu] \le n_F$, for some $\alpha = \alpha(\gamma) \in [0, 1]$,

$$\gamma_1 \le \theta\mathbb{E}[X_1|\Lambda = \gamma; \nu] + (\mu - \theta)n_1 + \alpha(\mu_F - \theta)n_F$$

and

$$\gamma_2 \le \theta\mathbb{E}[X_2|\Lambda = \gamma; \nu] + (\mu - \theta)n_2 + (1 - \alpha)(\mu_F - \theta)n_F.$$

Then,

$$\begin{aligned}
\theta\mathbb{E}[Q_\Sigma|\Lambda = \gamma; \nu] &\ge \left(\theta\mathbb{E}[X_1|\Lambda = \gamma; \nu] - \theta n_1\right)^+ + \left(\theta\mathbb{E}[X_2|\Lambda = \gamma; \nu] - \theta n_2\right)^+ - \theta n_F\Big)^+ \\
&\ge \left((\gamma_1 - \mu n_1 - \alpha(\mu_F - \theta)n_F)^+ + (\gamma_2 - \mu n_2 - (1 - \alpha)(\mu_F - \theta)n_F)^+ - \theta n_F\right)^+ \\
&\ge \left((\gamma_1 - \mu n_1)^+ + (\gamma_2 - \mu n_2)^+ - \mu_F n_F\right)^+,
\end{aligned}$$

where the last inequality follows from the fact that $((a-c)^+ + (b-d)^+ - e)^+ \ge (a^+ + b^+ - (c+d+e))^+$ for any $c, d, e \ge 0$.

Next, consider the case where $\theta > \mu_F$. Note that

$$\theta\mathbb{E}[Q_\Sigma|\Lambda = \gamma; \nu] = \gamma_1 + \gamma_2 - \mu\mathbb{E}[Z_1 + Z_2|\Lambda = \gamma; \nu] - \mu_F\mathbb{E}[Z_{F1} + Z_{F1}|\Lambda = \gamma; \nu].$$

Consider an auxiliary system, $\tilde{X}$, with all parameters the same except that its abandonment rate is $\tilde{\theta} = \mu_F$. We next construct a scheduling policy ν' such that

$$\mathbb{E}[Z_1 + Z_2|\Lambda = \gamma; \nu] = \mathbb{E}[\tilde{Z}_1 + \tilde{Z}_2|\Lambda = \gamma; \nu'] \text{ and } \mathbb{E}[Z_{F1} + Z_{F2}|\Lambda = \gamma; \nu] = \mathbb{E}[\tilde{Z}_{F1} + \tilde{Z}_{F2}|\Lambda = \gamma; \nu']. \tag{B.13}$$

The policy for ν' is constructed through a coupling that keeps $Z_i = \tilde{Z}_i$ and $Z_{Fi} = \tilde{Z}_{Fi}$ at all times. This can be achieved by assuming that arrivals and service completions in both systems coincide, and that the abandonments in the auxiliary system is a subset of the abandonments in the original system since $\tilde{\theta} < \theta$.

From (B.13), we have

$$\theta\mathbb{E}[Q_\Sigma|\Lambda = \gamma; \nu] = \tilde{\theta}\mathbb{E}[\tilde{Q}_\Sigma|\Lambda = \gamma; \nu'] \geq \left((\gamma_1 - \mu n_1)^+ + (\gamma_2 - \mu n_2)^+ - \mu_F n_F\right)^+,$$

where the last inequality follows from our analysis of the case where $\theta \leq \mu_F$.

We next prove the upper bound. *We first consider the case when $\theta = \mu_F$.* By Lemma 24, we have

$$\begin{aligned}
&\theta\mathbb{E}[Q_\Sigma|\Lambda = \gamma; \tilde{\nu}] \\
=&\theta\mathbb{E}[Q_1|\Lambda = \gamma; \tilde{\nu}] + \theta\mathbb{E}[Q_2|\Lambda = \gamma; \tilde{\nu}] \\
\leq&(\gamma_1 - \mu n_1 - \lfloor \delta n_F \rfloor \mu_F)^+ + K_1\sqrt{\gamma}_1 \exp(-\frac{K_2}{\gamma_1}(\gamma_1 - \mu n_1 - \lfloor \delta n_F \rfloor \mu_F)^2) + K_3 \\
&+ (\gamma_2 - \mu n_2 - \lceil (1-\delta) n_F \rceil \mu_F)^+ + K_1\sqrt{\gamma}_2 \exp(-\frac{K_2}{\gamma_2}(\gamma_2 - \mu n_2 - \lceil (1-\delta) n_F \rceil \mu_F)^2) + K_3 \\
\leq&(\gamma_1 - \mu n_1 - \delta n_F \mu_F)^+ + K_1\sqrt{\gamma}_1 \exp(-\frac{K_2}{\gamma_1}(\gamma_1 - \mu n_1 - \lfloor \delta n_F \rfloor \mu_F)^2) + K_3 + \mu_F \\
&+ (\gamma_2 - \mu n_2 - (1-\delta) n_F \mu_F)^+ + K_1\sqrt{\gamma}_2 \exp(-\frac{K_2}{\gamma_2}(\gamma_2 - \mu n_2 - \lceil (1-\delta) n_F \rceil \mu_F)^2) + K_3 + \mu_F \\
=&((\gamma_1 - \mu n_1)^+ + (\gamma_2 - \mu n_2)^+ - \mu_F n_F)^+ + 2(K_3 + \mu_F) \\
&+ K_1\sqrt{\gamma}_1 \exp(-\frac{K_2}{\gamma_1}(\gamma_1 - \mu n_1 - \lfloor \delta n_F \rfloor \mu_F)^2) + K_1\sqrt{\gamma}_2 \exp(-\frac{K_2}{\gamma_2}(\gamma_2 - \mu n_2 - \lceil (1-\delta) n_F \rceil \mu_F)^2).
\end{aligned}$$

The result then follows using the proof of Lemma 1 in [7].

Next, consider the case when $\theta < \mu_F$. Let the original system be labeled I. We form an auxiliary system II with the same parameters, except that the abandonment rate is $\theta^{II} = \mu_F$ and the holding

cost is $h^{II} = h\mu_F/\theta$. We write

$$\Pi^I(n_1, n_2, n_F; \tilde{\nu}) = c(n_1 + n_2) + c_F n_F + (a + h/\theta)A^I(n_1, n_2, n_F; \tilde{\nu})$$

where $A^I(n_1, n_2, n_F; \tilde{\nu}) = \theta\mathbb{E}[Q^I_\Sigma(n_1, n_2, n_F; \tilde{\nu})]$ is the stationary abandonment rate in system I. Similarly,

$$\Pi^{II}(n_1, n_2, n_F; \tilde{\nu}) = c(n_1 + n_2) + c_F n_F + (a + h/\theta)A^{II}(n_1, n_2, n_F; \tilde{\nu}).$$

We next show that

$$A^I(n_1, n_2, n_F; \tilde{\nu}) \leq A^{II}(n_1, n_2, n_F; \tilde{\nu}). \tag{B.14}$$

Note that

$$A^i = \mathbb{E}_\Lambda[\Lambda - \mu Z^i - \mu_F Z^i_F]$$

where Z^i and Z^i_F are the stationary number of busy rate-μ servers and rate-μ_F servers respectively, in system i, $i = I, II$. Thus, we only need to show that $Z^I \geq_{st} Z^{II}$ and $Z^I_F \geq_{st} Z^{II}_F$. Based on the scheduling policy $\tilde{\nu}$, it suffices to verify the following: If X^I is the stationary headcount in a single-class queue with m high priority rate-μ servers, m_F low priority rate-μ_F servers, and abandonment rate θ, and X^{II} is the same but with abandonment rate μ_F, then $X^I \geq_{st} X^{II}$. This is true because the birth rates of the two corresponding processes are the same, while the death rate in II is higher than in I. This proves (B.14), which further implies that

$$\Pi^{\lambda,I}(n^\lambda_1, n^\lambda_2, n^\lambda_F; \tilde{\nu}^\lambda) \leq \Pi^{\lambda,II}(n^\lambda_1, n^\lambda_2, n^\lambda_F; \tilde{\nu}^\lambda) \leq \tilde{\Pi}^\lambda(n^\lambda_1, n^\lambda_2, n^\lambda_F) + O(\lambda^{1-\alpha_2}) \text{ as } \theta^{II} = \mu_F.$$

Lastly, consider the case where $\theta > \mu_F$. We form a new auxiliary system III with the same parameters as X, except that $\mu_F^{III} = \theta$, $\mu^{III} = \mu\theta/\mu_F$, $c^{III} = c\theta/\mu_F$, and $c_F^{III} = c_F\theta/\mu_F$. Then,

$$\begin{aligned}\Pi^{\lambda,I}(n_1^\lambda, n_2^\lambda, n_F^\lambda; \tilde{\nu}^\lambda) &\leq \Pi^{\lambda,III}\left(\frac{\mu_F}{\theta}n_1^\lambda, \frac{\mu_F}{\theta}n_2^\lambda, \frac{\mu_F}{\theta}n_F^\lambda; \tilde{\nu}^\lambda\right) \text{ by Lemma 25} \\ &\leq \tilde{\Pi}^\lambda(n_1^\lambda, n_2^\lambda, n_F^\lambda) + O(\lambda^{1-\alpha_2}) \text{ as } \mu_F^{III} = \theta.\end{aligned}$$

□

B.4.5 Proof of Theorem 8.

Proof. Let $(n_1^{\lambda,*}, n_2^{\lambda,*}, n_F^{\lambda,*}; \nu^{\lambda,*})$ be optimal for (3.2). We have

$$\begin{aligned}&\Pi^\lambda(\lceil\tilde{n}_1^{\lambda,*}\rceil, \lceil\tilde{n}_2^{\lambda,*}\rceil, \lfloor\tilde{n}_F^{\lambda,*}\rfloor; \tilde{\nu}^\lambda) \\ \leq&\tilde{\Pi}^\lambda(\lceil\tilde{n}_1^{\lambda,*}\rceil, \lceil\tilde{n}_2^{\lambda,*}\rceil, \lfloor\tilde{n}_F^{\lambda,*}\rfloor) + O(\lambda^{1-\alpha_2}) \text{ by the upper bound in Lemma 7} \\ \leq&\tilde{\Pi}^\lambda(\tilde{n}_1^{\lambda,*}, \tilde{n}_2^{\lambda,*}, \tilde{n}_F^{\lambda,*}) + 2c + c_P\mu_F + O(\lambda^{1-\alpha_2}) \\ \leq&\tilde{\Pi}^\lambda(n_1^{\lambda,*}, n_2^{\lambda,*}, n_F^{\lambda,*}) + O(\lambda^{1-\alpha_2}) \\ \leq&\Pi^\lambda(n_1^{\lambda,*}, n_2^{\lambda,*}, n_F^{\lambda,*}; \nu^{\lambda,*}) + O(\lambda^{1-\alpha_2}) \text{ by the lower bound in Lemma 7.}\end{aligned}$$

□