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

Jinsheng Chen
Singapore Institute of Manufacturing Technology, Agency for Science, Technology and Research, Singapore 636732
chen_jinsheng@simtech.a-star.edu.sg

Jing Dong
Graduate School of Business, Columbia University, New York, NY 10027, jing.dong@gsb.columbia.edu

Pengyi Shi
Krannert School of Management, Purdue University, West Lafayette, IN 47907, shi178@purdue.edu

Motivated by the growing availability of advanced demand forecast tools, we study how to utilize future demand information in designing routing strategies in queueing systems under demand surges. We consider a parallel server system operating in a nonstationary environment with general time-varying arrival rates. Servers are cross-trained to help non-primary customer classes during demand surges. However, such flexibility comes with various operational costs, such as a loss of efficiency and inconvenience in coordination. We characterize how to incorporate the future arrival information into the routing policy to balance the tradeoff between various costs, and quantify the benefit of doing so. Based on transient fluid control analysis, we develop a two-stage index-based look-ahead policy that explicitly takes the overflow costs and future arrival rates into account. The policy has an interpretable structure, is easy to implement, and is adaptive when the future arrival information is inaccurate. In the special case of the N-model, we prove that this policy is asymptotically optimal even in the presence of certain prediction errors in the demand forecast. We substantiate our theoretical analysis with extensive numerical experiments, showing that our policy achieves superior performance compared to other benchmark policies (i) in complicated parallel server systems and (ii) when the demand forecast is imperfect with various forms of prediction errors.

Key words: skill-based routing, demand surge, managing flexibility, transient queue, optimal control theory, asymptotic analysis

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 in a time non-stationary environment when certain classes of customers experience demand surges, and yet their dedicated capacity cannot be scaled up quickly. Meanwhile, with the recent advancement of statistical learning tools and the growing availability of data, many advanced forecasting models have been developed to accurately predict future demand patterns and surges. Take the COVID-19 pandemic as an example: researchers from different fields have worked together

to develop prediction models of demand surges for different types of hospital resources (e.g., ICU beds and ventilators). Figure 1 shows the prediction made on one index date ("today"), including the realized demand (before today, with multiple demand surges) and the projected demand (after today) for Intensive Care Unit (ICU) beds of COVID-19 patients in the US from Institute for Health Metrics and Evaluation (2022). Yet, the majority of these prediction models do not provide prescriptive solutions for the effective allocation of resources. Hospital management needs more concrete decision support on how to translate the demand forecasts into determining whether they have enough beds to accommodate the COVID-19 patients, and, if not, whether they should use beds from other specialties, e.g., by postponing elective surgeries. More generally, there is a pressing need to integrate demand forecasts in implementable resource allocation decision support for various service operations applications. In this paper, we address this need by studying how to utilize future demand information to design optimal routing strategies to deal with demand surges in parallel server systems. We explicitly characterize how to incorporate future demand into the routing decision – matching customers with proper resources – and quantify its benefit.

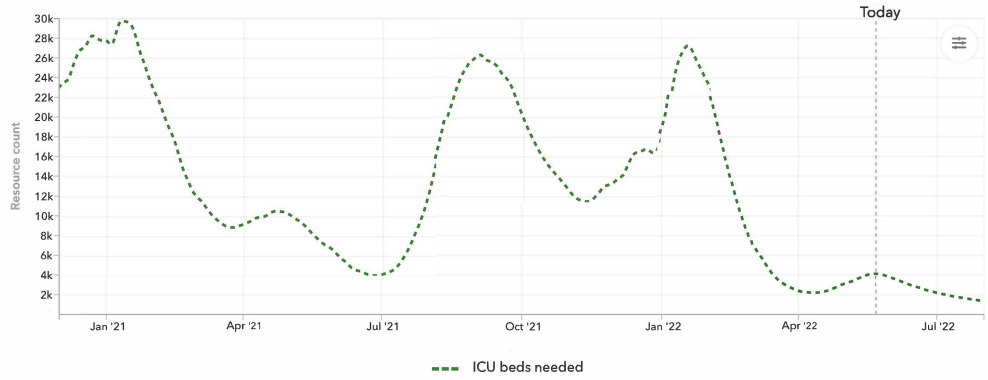


Figure 1 **Demand for ICU beds for COVID-19 patients in the US, based on IMHE projection (Institute for Health Metrics and Evaluation 2022).**

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 this paper, we use servers to refer to the critical resources in service systems, such as staffed 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 main reason is the presence of variability in demand. In particular, while demand can change significantly from time to time, 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 types 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 a demand surge. To deal with the normal range of stochastic fluctuations in demand, a certain degree of slackness is usually added in the capacity – 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 (Borst et al. 2004).

Demand surges happen infrequently, but can lead to severe congestion 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 machine learning and predictive analytics have significantly increased our ability to forecast some of these surges. With the future demand information, system managers may proactively leverage partial flexibility in customer routing, that is, temporarily assigning customers experiencing demand surges to non-primary servers, to effectively mitigate the effect of demand surges.

However, concrete decision rules are, in general, lacking to support the routing decision under time-varying demand, especially when we need to carefully balance the benefits and costs associated with using flexible resources. For example, in the hospital inpatient flow setting, 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 a longer length of stay and higher readmission rate (Song et al. 2019). In addition, it may generate a greater workload for nurses in the off-service ward due to multi-tasking (Best et al. 2015), and can create a sense of unfairness among staff (Armony and Ward 2010). Similar tradeoffs between the benefits and the costs of flexibility are also pervasive in other service systems. Examples include call centers (Aksin et al. 2007), bike-sharing (Shu et al. 2013), and emergency departments (Song et al. 2015).

In this paper, we take an important first step to study how to leverage *future demand information* 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 server 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 certain non-primary pools. Such routing is referred to as "overflow." Decisions need to be made between routing a customer to the primary pool or to the non-primary pool. The obvious benefit of overflow is to reduce congestion, captured via the holding cost. Meanwhile, overflow is associated with the following costs:

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 that 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 compromised service quality. We model these costs through an overflow cost. 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.

To model demand surges, we allow general time-varying arrival rates whereby one or more customer classes may experience one or multiple demand surges within a certain time period. 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 future arrival rates, which may be subject to prediction errors.

Our objective for optimal routing is to minimize the cumulative holding (linear waiting) costs and overflow costs until the demand surge is fully absorbed. The problem is challenging due to the salient features in our model. Specifically,

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

2. Because of service slowdown, a very aggressive overflow policy generates more workload, which may lead to a higher holding cost than having no overflow.

3. In the presence of overflow costs, the optimal routing policy may not be work-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 those non-primary servers idle to avoid the overflow cost.

To sum up, the overflow cost and service slowdown, compounded with the general time-varying arrival rates, make the routing decisions highly nontrivial. Unnecessary overflow can result in both higher holding costs and higher overflow costs. These complications can cause existing well-known

policies, such as the $c\mu$-rule (Buyukkoc et al. 1985) or the maximum pressure policy (Dai and Lin 2005), to perform highly suboptimally. The non-stationarity and high dimensionality of the problem also make many existing analytical and numerical tools inapplicable. To derive structural insights into the routing problem, we take the fluid approximation approach. We formulate a transient fluid control problem and derive closed-form index-based optimal policies using its dual. Our main results and contributions are as follows:

I. Prescriptive framework. We are one of the first papers to prescribe how to incorporate demand forecasts 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 and have some periodic patterns (Liu and Whitt 2011), or they focus on capacity planning to hedge against the uncertain arrival rates (Bassamboo and Zeevi 2009). On the other hand, most demand forecasting works focus on the prediction side only; see, for example, Ibrahim and L'Ecuyer (2013) for call center arrivals and Baas et al. (2021) for hospital occupancy. Our paper 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 transient fluid control problem. We then translate the fluid-based policy to the stochastic system and provide the performance guarantee by proving its asymptotical optimality for a sequence of properly scaled stochastic systems even when there are prediction errors, as long as the errors are of a smaller order than the actual arrival rates. To the best of our knowledge, the asymptotic optimality result in the presence of prediction errors in a time-nonstationary environment is novel.

II. Two-stage index-based policy using future demand information. Our main development focuses on 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 scheduling policy derived from the fluid control 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 3.

Interpreting our two-stage index-based look-ahead policy provides insights into the value of future demand information and how to proactively 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 helping this class in anticipation of the upcoming drop in demand. This proactive nature of our policy is distinct from other well-known policies that are agnostic to the arrival rate, such as the $c\mu$ rule and the maximum pressure policy. In our numerical experiments, we demonstrate that the $c\mu$ rule or the maximum pressure policy, even after adjustments to account for the overflow cost, can sway from performing reasonably well to significantly worse depending on the arrival rate settings: in certain scenarios, the cost gap between these policies and our proposed policy can exceed 100%, whereas our policy consistently performs well across a large combination of parameter settings tested.

We stress that although it is somewhat expected that having future demand information is beneficial, 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 and account for the externality of congestion (Naor 1969).

III. Pontryagin's minimum principle. Due to the time nonstationarity and high dimensionality, the corresponding scheduling problem has not been well-studied in the literature. We derive structural insights by studying the corresponding fluid transient control problem and its dual. In particular, our index-based policy is derived using Pontryagin's minimum principle, and the derivation involves non-trivial applications of the principle due to state constraints. The main difficulty lies in coming up with the right dual variables. The problem is further complicated by the fact that there are multiple customer classes and general time-varying arrival rates. By explicitly characterizing the dual variables, we derive a two-stage index structure of the optimal policy. These developments could shed light on other transient queueing control problems.

IV. Practical applicability to complex systems. For a scheduling policy to be useful in practice, it needs to be adaptable to (i) complicated network structures and (ii) imperfect demand forecasts. For (i), we extend the fluid-control analysis beyond the N-model, and explicitly characterize the optimal fluid-control policy for the X-model and some multi-class extensions of the N-model. Based on the structure of optimal control in these models, we propose 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 also 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 parameters in different network settings, including 5-by-5 systems that are parameterized based on the setting of a hospital inpatient ward network.

For (ii), it is worth noting that our policy is adaptive by nature. In particular, the estimation of the time to empty the queue can be easily updated based on the demand information available and the observed queue length at each decision epoch. We discuss these extensions in Sections 3.2

and 3.3. Furthermore, we substantiate our asymptotic optimality results, which require the prediction errors to be of a smaller order than the arrival rates, with numerical evaluation in scenarios where (1) the demand forecast has errors that increase over time and can correlate across different time periods, (2) the forecast is only available up to a limited time window, and (3) the forecast has a delay. Numerical results suggest that our proposed policy continues to perform well even when the prediction errors are of a similar magnitude as the arrival rates. The index structure of our proposed policy has built-in resilience to estimation errors. In particular, we note that a) our index depends on the time to empty the queue, which in turn depends on the *aggregated* demand over time – this quantity is more robust to errors than demand prediction at individual time epochs, e.g., days, since the daily errors may cancel out when aggregating over multiple days; b) the index is dynamically updated as the queue builds up or more information becomes available; c) the order of the index determines the overflow action, so as long as the prediction errors do not change the orders, our policy is not affected. 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 systems when demand surges are present. The comparison with other benchmark policies also provides useful insights into managing systems under demand surges.

1.1. Paper Organization

The rest of the paper is organized as follows. We conclude this section with a brief review of the literature. In Section 2, we introduce our main stochastic model, the N-model, and the associated optimal routing problem. We study a deterministic fluid control problem in Section 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 4. We study the optimal fluid control problem for several extended models and propose a two-stage index-based look-ahead policy for general parallel-server systems in Section 5. We substantiate our theoretical analysis with extensive numerical experiments in Section 6. All proofs are left to the Appendix and E-companion.

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 information.

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 (Akşin and Karaesmen 2007, Graves and Tomlin 2003, Simchi-Levi and Wei 2012,

Smith and Whitt 1981, van Mieghem 1998). Bassamboo et al. (2012) and Tsitsiklis and Xu (2012) 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 (Mandelbaum and Reiman 1998), different priorities among different classes of jobs (Ata and Van Mieghem 2009), efficiency loss due to multi-tasking (Pinker and Shumsky 2000), and agent incentives (Song et al. 2015), 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 (Garnett and Mandelbaum 2000). 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, van Mieghem (1995) 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, Harrison (1998) shows that naive $c\mu$-rule can lead to instability. A discrete-review policy is proposed for the setting without overflow cost and shown to achieve asymptotically optimal performance in heavy traffic. Bell and Williams (2001) show that a threshold-based priority rule is asymptotically optimal. Mandelbaum and Stolyar (2004) 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. Dai and Lin (2005) and Bramson et al. (2021) show that the maximum pressure policy is throughput optimal. Dai and Lin (2008) further prove that with quadratic holding cost, the maximum pressure policy is asymptotically optimal under the conventional heavy-traffic scaling for some models. Stolyar (2004) establishes the asymptotic optimality of a general class of MaxWeight policies with strongly convex holding costs. There is also rich literature on SBR in the many-server asymptotic regimes; see Chen et al. (2020) 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 for stationary settings 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 captures first-order system dynamics well (Liu and Whitt 2011, Mandelbaum et al. 1998), is often used to analyze transient queueing behavior. Maglaras (2000) and Bäuerle (2002) detail how to develop effective queueing control policies based

on the optimal fluid control. Most of these "fluid-inspired" policies try to track the optimal fluid trajectory in the original stochastic system. Similar ideas have also been applied to other online resource allocation problems where probabilistic routing based on the fluid optimal solution has been proposed (see, for example, Jasin and Kumar (2012), Stein et al. (2020)). Different from this line of research, we employ optimal control theory to explicitly characterize an index structure for the optimal fluid control; see Sethi and Thompson (2000) and Grass et al. (2008) for an overview of optimal control theory. In particular, we leverage Pontryagin's Minimum Principle (Hartl et al. 1995). Hampshire and Massey (2010) review several applications of optimal control theory to dynamic rate queues. Compared to previous fluid-tracking policies, our policy has an interpretable structure and is easy to implement in practice. Recently, Hu et al. (2019) apply Pontryagin's Minimum Principle to study the optimal scheduling of proactive service in systems with customer deterioration. Ata and Peng (2020) 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 most relevant work is Chang et al. (2004), who study scheduling policies in a two-class, single-server system. However, they do not incorporate the overflow cost, and their analysis is limited to simple arrival patterns (high/low constant arrivals) without considering possible prediction errors. Our analytical framework allows us to study general time-varying arrivals and the tradeoff between holding and overflow costs.

The value of future demand information. 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 (Ata and Peng 2020, Delana et al. 2021, Xu and Chan 2016). These works require more detailed demand information, including the actual 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., Bassamboo and Zeevi (2009), Gurvich et al. (2010) for call center staffing and Hu et al. (2021) 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. Second, routing decisions are fundamentally different from staffing decisions as they happen at different time scales.

2. Problem Formulation

To demonstrate our methodology and key insights, we use the N-model as our main model; other network structures including the X-model and extensions of the N-model are studied in Section 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 $(\lambda_i(t))_{t\geq 0}$. Each class has its own queue when waiting to get served and customers within the same class are served on a first-come-first-served basis. 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 incompatibility in the N-model.

Let $X_i(t)$ denote the number of class i customers in the system and $Z_{ij}(t)$ denote the number of class i customers in service in pool j at time t. Let A_i and S_{ij} denote rate-1 Poisson processes modeling the arrival and service processes, respectively. 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 some routing 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 the realizations of underlying stochastic processes in the future, but we allow the policies to take future arrival rates (or estimated 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 within the considered policy class. 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. We use the words "routing" and "scheduling" interchangeably in the rest of the paper. Both refer to the action of matching customers to servers, i.e., which available server (if any) a customer should be routed to or which waiting customer (if any) an available server should serve next.

Focusing on planning under demand surges, we consider time-varying arrival rates that can cause one or more customer classes to experience surges in demand (arrivals). We assume that time 0 is the beginning of the demand surge and the surge will last for a finite amount of time. In addition, the 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). Figure 1 in Section 1 illustrates one example of demand surges in the COVID-19 context.

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 costs related to the routing decisions: holding cost and overflow cost. We aim to find a routing policy that minimizes the total cost of holding customers in the system and overflowing class 1 customers to pool 2 over a properly defined planning 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 routing problem is formulated as finding a policy π that minimizes

$$V^\pi(x) = \mathbb{E}\left[\int_0^T h_1 X_1^\pi(t) + h_2 X_2^\pi(t) + \phi_{12} Z_{12}^\pi(t) dt \;\middle|\; X(0) = x\right]. \tag{1}$$

Here, the planning horizon T is deterministic and is a long enough time such that the system can fully absorb the demand surge by time T, i.e., T is a time after the demand surge such that $X_1(T)$ and $X_2(T)$ are "small" with a high probability. (The definition of T and the interpretation of small with a high probability will be made precise in Section 4.)

Solving (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 (1).

3. Fluid Optimal Control

We first specify the deterministic fluid model $q(t) = (q_1(t), q_2(t))$ that resembles the stochastic system described in Section 2. The arrival rates and service rates in the fluid model are the same as those in the stochastic system. The dynamics of the fluid model are characterized via

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

where $\dot{q}_i(t) := dq_i(t)/dt$. The amount of service capacity assigned to each class in the fluid model, $z(t) = (z_{11}(t), z_{12}(t), z_{22}(t)) \in \mathbb{R}_0^{+,3}$, 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 $\mathcal{Z}(t)$. Note that the set of admissible controls has to satisfy two state constraints: $q_1(t) \geq 0, q_2(t) \geq 0$. In particular, when $q_1(t) = 0$, $\mu_{11} z_{11}(t) + \mu_{22} z_{22}(t) \leq \lambda_1(t)$, and when $q_2(t) = 0$, $\mu_{22} z_{22}(t) \leq \lambda_2(t)$. When viewed as an approximation to the stochastic system described in Section 2, the fluid model is known to be a good approximation when the system is very congested or when the variability in the arrival rates dominates the stochastic variation in the arrivals and services (see, for example, Liu and Whitt (2011), Maglaras (2000), Yom-Tov and Mandelbaum (2014)). The fluid model can also be viewed as the limit of a sequence of stochastic systems under the conventional scaling, in which we speed up the arrival and service rates while keeping the number of servers constant (Chen et al. 2020, Iglehart and Whitt 1970). We utilize this limit interpretation to establish asymptotic performance guarantees when applying the scheduling policy derived based on the fluid model to the stochastic system in Section 4. We interpret $q_i(t)$ as the amount of class i "fluid" in the system, and refer to $q(t) = (q_1(t), q_2(t))$ as the fluid queue. $z_{ij}(t)/s_j$ can be interpreted as the proportion of time pool j capacity is allocated to serve class i jobs in $[t, t + dt)$.

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 $i = 1, 2$, there exists $\kappa_i \in [0, \infty)$ such that $\lambda_i(t) \geq s_i \mu_{ii}$ when $t < \kappa_i$ and $\lambda_i(t) < s_i \mu_{ii}$ when $t \geq \kappa_i$, i.e., κ_i is the length of the demand surge for class 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*

$$
\begin{aligned}
W(x,t) = \inf_{z} \;& q_1(t) + q_2(t) \\
& s.t.\ \dot{q}_1(u) = \lambda_1(u) - \mu_{11} z_{11}(u) - \mu_{12} z_{12}(u), \quad q_1(0) = x_1 \\
& \quad \dot{q}_2(u) = \lambda_2(u) - \mu_{22} z_{22}(u), \quad q_2(0) = x_2 \\
& \quad z(u) \in \mathcal{Z}(u) \text{ for all } u \in [0, t].
\end{aligned} \tag{2}
$$

Condition 4 in Assumption 1 indicates that the demand surge is large enough such that the fluid queue cannot be emptied by any admissible control before $\kappa_1 \vee \kappa_2$. It might be violated if the demand surge for class 1 is small ($\lambda_1(t) < s_1 \mu_{11} + s_2 \mu_{12}$ for $t < \kappa_1$) and pool 2 has a lot of spare capacity ($\lambda_2(t) < s_2 \mu_{22}$). In addition, note that Assumption 1 considers a single demand surge for each class. We relax this assumption in Section 3.4 to consider multiple surges.

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 adaptations of the optimal policy to scenarios where (i) we only have access to estimated arrival rates (Section

3.2), and (ii) we only have access to a limited look-ahead time window (Section 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 4).

The fluid control problem corresponding to (1) is formulated as

$$\begin{aligned} \inf_{z} \ & \int_0^{\sigma} h_1 q_1(t) + h_2 q_2(t) + \phi_{12} z_{12}(t) dt \\ \text{s.t. } & \dot{q}_1(t) = \lambda_1(t) - \mu_{11} z_{11}(t) - \mu_{12} z_{12}(t), \ q_1(0) = x_1 \\ & \dot{q}_2(t) = \lambda_2(t) - \mu_{22} z_{22}(t), \ q_2(0) = x_2 \\ & z(t) \in \mathcal{Z}(t) \text{ for all } t \geq 0, \end{aligned} \tag{3}$$

where $\sigma = \inf\{t \geq \kappa_1 \vee \kappa_2 : q_1(t) + q_2(t) = 0\}$. Note that under Assumption 1, with a proper scheduling policy, the fluid queue will eventually hit zero and stay there. Thus, in this case, solving the optimal fluid control problem with a fixed termination state, i.e., $q_1(t) = q_2(t) = 0$ in (3), is the same as solving an optimal control with fixed planning horizon T when T is large enough.

Remark 2 (Surge scenarios). Our development applies to two types of surges. The first (main) demand surge scenario is that for a certain period of time, the arrival rate is substantially higher than the service capacity of the system, i.e, Condition 1 in Assumption 1. Examples include demand for hospital resources during a bad flu season, and phone calls to an airline customer contact center after widespread flight cancellations due to inclement weather. Note that this notion of demand surge is different from system congestion caused by stochastic fluctuations of the arrivals and services (i.e., even in a stationary system, we can have periods of time where the queue is long, especially when the system is critically loaded). The second surge scenario is when the system starts with an unusually large queue (backlog of demand) while the "future" arrival rates are at the normal level, i.e., $\kappa_i = 0$ but $\max\{q_1(0), q_2(0)\} \gg 0$. Examples include mass casualty incidents where a large number of casualties arrive almost at once or during a very short period of time (Yom-Tov and Mandelbaum 2014). Numerical examples in the main paper focus on the first scenario, and those for the second scenario can be found in E-Companion(EC) 3.2.

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

$$G_i^t(x_i) := \inf\left\{\Delta \geq 0 : \int_t^{t+\Delta} (s_i \mu_{ii} - \lambda_i(s))\, ds = x_i\right\}, \tag{4}$$

and for $x_i = 0$,

$$G_i^t(0) := \lim_{x_i \downarrow 0} G_i^t(x_i). \tag{5}$$

We can interpret $G_i^t(x_i)$ as the time it takes to empty queue i after time t using only primary resources, given that $q_i(t) = x_i$. For a fixed value of t, it is continuous and strictly increasing in x_i. Note that under (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 scheduling policy for the fluid control problem.

THEOREM 1 **(Optimal control policy in N-model)**. *Under Assumption 1, the optimal control for* (3) *takes the following form. Pool 1 serves as many class 1 customers as possible – i.e.,*

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

I. If $h_1\mu_{12} \geq h_2\mu_{22}$, 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. If $h_1\mu_{12} \leq h_2\mu_{22}$, 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.,

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

$$z_{22}^*(t) = s_2 \wedge \frac{\lambda_2(t)}{\mu_{22}}.$$

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 B. It utilizes Pontryagin's Minimum Principle. The indices are derived based on the dual functions which are also known as the adjoint vectors. The 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 prioritize class 1, or only partially help when there is spare capacity. In particular, if $h_1\mu_{12} > h_2\mu_{22}$, pool 2 may prioritize class 1; otherwise, pool 2 prioritizes its own class and may provide partial help to class 1. 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 as providing *partial help* to class 1.

3.1. Interpretation of the Two-Stage Policy

3.1.1. 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 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 $t \le 10$, 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 $t \ge 15.8$.

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 $x_i \ge 0$. The optimal scheduling policy is similar to the maximum pressure policy but takes the slack capacity $s_i\mu_{ii} - \lambda_i$ into account.

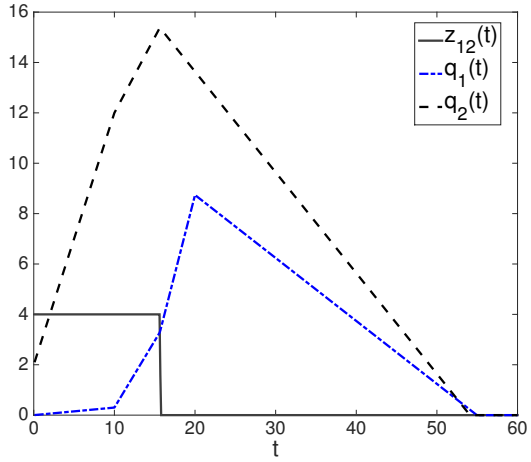


Figure 2 **Optimal trajectory of the N-model. (Parameter setting:** $s_1 = 3$, $s_2 = 4$, $\mu_{11} = \mu_{22} = 0.25$, $\mu_{12} = 0.18$, $h_1 = 1.5$, $h_2 = 1$, $\phi_{12} = 1$, $\lambda_1(t) = 1.5 \times \mathbf{1}\{0 \le t \le 10\} + 2 \times \mathbf{1}\{10 < t \le 20\} + 0.5 \times \mathbf{1}\{t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

3.1.2. 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:

$$h_1 G_1^t(q_1(t)) - \frac{\phi_{12}}{\mu_{12}} > h_2 G_2^t(q_2(t)) \frac{\mu_{22}}{\mu_{12}}. \tag{6}$$

Similarly, in Case II, we check the condition $h_1 G_1^t(q_1(t)) > \phi_{12}/\mu_{12}$. Here, ϕ_{ij}/μ_{ij} corresponds to the expected 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 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 (Ata and Peng 2020).

We note that this cost comparison is in contrast to comparing both costs (overflow and holding) at the myopic cost-rate level. For the cost rate, when using pool 2 to serve class 1 customers, the holding cost decreases at rate $h_1\mu_{12}$, while the overflow cost increases at rate $\phi_{12}\mu_{12}$. If we compare the instantaneous cost rate, 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 6. This myopic rule can result in significantly worse performance than our proposed policy in many settings, which 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.

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. The estimation error affects the performance of the policy because the look-ahead function is now calculated based on the estimated arrival rate:

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

for $x_i > 0$. In this case, if $h_1\mu_{12} \geq h_2\mu_{22}$, 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 $\tilde{G}_i^t(x_i)$ to be close to $G_i^t(x_i)$, and our policy should perform well. This intuition will be made rigorous when translating the fluid policy back to the stochastic systems. In particular, we will show in Section 4 that under suitable conditions on the estimation error $\epsilon_i(t)$, the policy based on $\tilde{G}_i^t(x_i)$ is asymptotically optimal in the stochastic systems. We substantiate this analytical result with numerical results in Section 6, where we show the robust performance of our policy under various forms of prediction errors.

3.3. Adaptivity to Limited Look-ahead Windows

In this section, 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 a 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 define the nominal arrival rate as the arrival rate after the end of the demand surge, i.e., the arrival rate in "normal" times. This nominal rate is assumed to be a constant that is less than the service capacity. Then, we can calculate $\hat{G}_i^t$'s using the nominal arrival rates outside the prediction time window. In particular,

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

where $\hat{\lambda}_i(s) = \lambda_i(s)$ for $t \leq s < t+W$ and $\hat{\lambda}_i(s) = \lambda_i^0$ for $s \geq t+W$, with λ_i^0 being the nominal arrival rate. For example, when $W = 0$, $\hat{G}_i^t(q_i(t)) = q_i(t)/(s_i\mu_{ii} - \lambda_i^0)$. In Section 6.3.2, we test the performance of our proposed policy with varying values of W in the stochastic system. We observe that our policy achieves good performance even with a relatively small look-ahead time window.

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. $(\lambda_1(t))_{0\le t\le\kappa_c}$ *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).

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 G_i^t defined in (4).

THEOREM 2 **(Optimal control under two demand surges)**. *Under Assumption 2, the optimal control for* (3) *takes the following form. Pool 1 serves as many class 1 customers as possible. Moreover:*

I. If $h_1\mu_{12}\ge h_2\mu_{22}$, *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.*

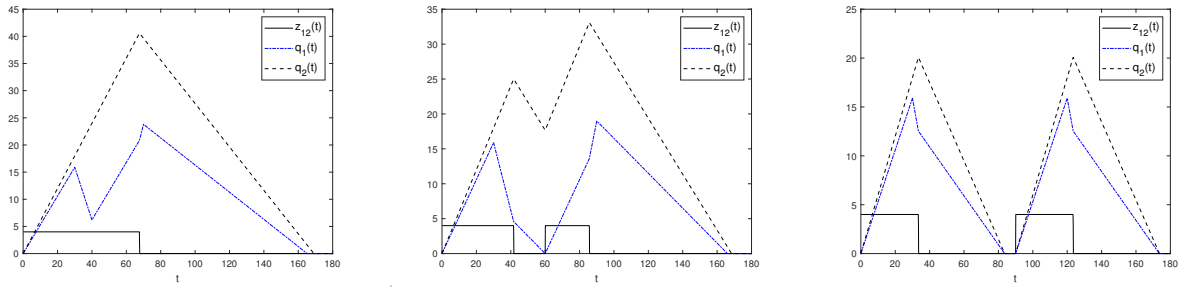
II. If $h_1\mu_{12}\le h_2\mu_{22}$, *pool 2 gives priority to class 2 and will help class 1 when* $q_2(t)=0$ *and* $h_1\mu_{12}G_1^t(q_1(t))>\phi_{12}$; *otherwise, pool 2 serves class 2 only.*

The proof of Theorem 2 is in the E-companion, and follows a similar framework to that of Theorem 1. The actual helping behavior of the policy characterized in Theorem 2 depends on the length of the interval between the two demand surges. We illustrate the idea via a numerical example: Class 2 has a constant arrival rate $\lambda_2(t)\equiv 0.6$, while the arrival rate for class 1 follows

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

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 3. 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 periods.

We conclude 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 implementation 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 surge were



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

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

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, as the G values can be easily updated in real-time. We substantiate this point with more numerical results in Section 6.3.2 when dealing with two demand surges and limited look-ahead windows.

4. Asymptotic Optimality

In this section, we translate the optimal control defined in Theorem 1 to the original stochastic system introduced in Section 2. We prove that the translated policy is asymptotically optimal along a properly scaled sequence of stochastic systems.

To specify the sequence of stochastic systems, we first elaborate on the planning horizon T in (1). 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, we adopt a deterministic planning horizon for the stochastic system. Given the initial state $X(0) = x$, we define the planning horizon $T = T(x)$ based on the fluid dynamics. 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

$$\begin{aligned} \dot{q}_1^o(t) &= (\lambda_1(t) - s_1\mu_{11} - s_2\mu_{12})1\{q_1^o(t) > 0\} + (\lambda_1(t) - s_1\mu_{11} - s_2\mu_{12})^+ 1\{q_1^o(t) = 0\}, \\ \dot{q}_2^o(t) &= \lambda_2(t) - \left(s_2 - \frac{(\lambda_1(t) - s_1\mu_{11})^+}{\mu_{12}}\right)^+ \mu_{22} 1\{q_1^o(t) = 0\}. \end{aligned} \tag{7}$$

Define

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

which corresponds to the time by which the class i queue should be emptied if pool 2 prioritizes class 1 all the time. Note that $G_1^0(x_1) \ge \tau_1^o(x)$ and $G_2^0(x_2) \le \tau_2^o(x)$. 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 help 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 is fixed along the sequence. We speed up time and scale down space by n. Specifically, for the n-th system, the arrival rate at time t is $\lambda_i^n(t) = n\lambda_i(t)$ for class i 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; $Z^n(t) = (Z_{11}^n(t), Z_{12}^n(t), Z_{22}^n(t))$ 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., $Z^n(t) = \pi_t^n(X^n(t))$. As discussed in Section 2, the admissible controls are preemptive and non-anticipative, but we have access to some estimated arrival rate $\Lambda_i^n(t)$ for each class i. The server allocation policies satisfy the following conditions:

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

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:

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

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. This is because the queue length processes scale with n (as the speed at which arrivals and departures happen scales with n) while the number of servers does not, i.e., $X_1^{n,\pi^n}(t) = O(n)$ while $Z_{12}^{n,\pi^n}(t) = O(1)$. To interpret $Z_{12}^{n,\pi^n}(t)$, we note that the number of servers does not scale with n, so $Z_{12}^{n,\pi^n}(t)/s_2$ can be interpreted as the fraction of time pool 2 servers are allocated to serve class 1 customers in $[t, t+dt)$. A similar scaling is used in Bäuerle (2000).

We next translate the optimal fluid policy to 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., $\mathbb{P}(E_i^n(\cdot)/n \to 0$ u.o.c. as $n \to \infty) = 1$. 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 on arrival rate uncertainty, see, e.g., Bassamboo et al. (2010), Maman et al. (2009).

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

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

for $x_i > 0$. The scheduling policy $\{\tilde{\nu}^n\}_{n\geq 1}$ is defined as follows. For the n-th system: pool 1 serves class 1 customers as much as possible. Consider the case $h_1\mu_{12} > h_2\mu_{22}$. If

$$h_1\mu_{12}\tilde{G}_{1,n}^t(X_1^n(t)) - \phi_{12} > h_2\mu_{22}\tilde{G}_{2,n}^t(X_2^n(t)) \tag{8}$$

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

$$h_1\mu_{12}\tilde{G}_{1,n}^t(X_1^n(t)) - \phi_{12} > 0 \tag{9}$$

at time t, pool 2 serves both classes but gives preemptive priority to class 2; otherwise, pool 2 serves class 2 only. When implementing the policy, $\tilde{G}_{i,n}^t(X_i^n(t))$ is supposed to be calculated at every decision epoch (when there is an arrival or a departure). We make two remarks about its calculation. First, $\tilde{G}_{i,n}^t(X_i^n(t))$ can be easily calculated numerically, especially when the arrival rate is piecewise linear. Second, $\tilde{G}_{i,n}^t(x_i)$ is continuous in t and x_i, and the policy only changes when the inequality (8) or (9) changes sign. Thus, we may not need to update this calculation very frequently in implementation.

The following theorem shows that $\{\tilde{\nu}^n\}_{n\geq 1}$ is asymptotically optimal. Let $\bar{V}^*(x)$ denote the optimal objective value of the corresponding fluid control problem (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 $\{\tilde{\nu}^n\}_{n\geq 1}$,

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

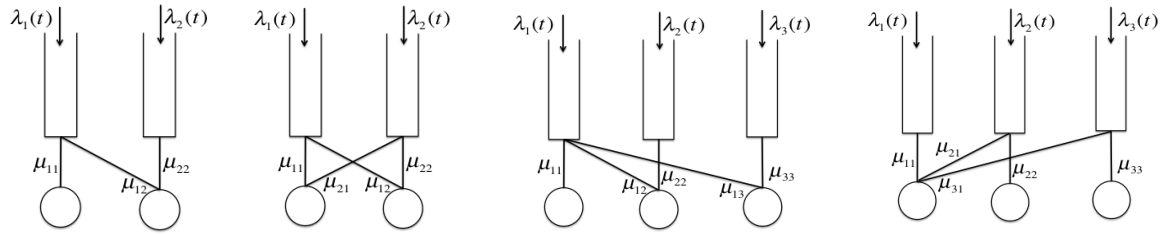
The proof of Theorem 3 is in Appendix C. To incorporate the prediction error in the asymptotic optimality result, we leverage the continuity properties of the look-ahead function G_i^t. This theorem suggests that when applied to stochastic systems, the two-stage index-based look-ahead policy achieves 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 6, we go beyond this theoretical result and numerically investigate the performance of our policy with more general forms of prediction errors. 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. This benefits from the index structure of our policy, which has some built-in resilience to perturbations. First, calculating G_i^t requires us to integrate the arrival rates over a period of time. The estimation error at individual times s may cancel out when aggregated over a period of time. Second, 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 a given time t.

5. Beyond the N-Model

In this section, we study three 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), i.e., many server-pools can serve class 1; and iii) the one-helps-many extended N-model (exN2), i.e., pool 1 can serve many classes. See Figure 4 for a pictorial illustration of these models together with the N-model. Note that the exN2-model 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. Moreover, when presenting the results, we focus on emphasizing the key difference between these models and the N-model. By comparing the X-model with the N-model, we highlight the effect of 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 or multiple pools of servers. These insights lead us to propose a heuristic policy for general multi-class multi-pool systems in Section 5.3.

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 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



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

Figure 4 Queueing models with partial flexibility

$\mu_{ij} = 0$, class i customers and pool j servers are not compatible. Let $q_i(t) \in [0, \infty)$ denote the amount of class i fluid 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 $z(t) = (z_{ij}(t) : i, j = 1, \cdots, I)$, where $z(t) \in \mathcal{Z}(t)$, and

$$\mathcal{Z}(t) = \left\{ z(t) : z_{ij}(t) \geq 0, i, j = 1, \cdots, I, \ \sum_i z_{ij}(t) \leq s_j, j = 1, \cdots, I, \ q_i(t) \geq 0, i = 1, \ldots, I \right\}.$$

Similar to the N-model, we allow the policies to use the future arrival rate information.

For arrival rates, we consider the scenario in which each class may experience a demand surge that lasts for a finite amount of time; the extension to multiple surges can be done similarly as in Section 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 4 in Appendix A for a full specification of the assumptions for general multi-class multi-pool systems), we define $\sigma = \inf \{t \geq \bar{\kappa} : \sum_i q_i(t) = 0\}$, which can be interpreted as the time to fully absorb the demand surge. Class i fluid in the system incurs a cost of h_i per unit job per unit of 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:

$$\begin{aligned} &\inf_z \int_0^\sigma \sum_{i=1}^{I} h_i q_i(t) + \sum_{i=1}^{I} \sum_{j=1}^{I} \phi_{ij} z_{ij}(t) dt \\ &\text{s.t. } q_i(0) = x_i, \ \dot{q}_i(t) = \lambda_i(t) - \sum_{j=1}^{I} \mu_{ij} z_{ij}(t), i = 1, \cdots, I \\ &\qquad z(t) \in \mathcal{Z}(t) \text{ for all } t \geq 0. \end{aligned} \tag{10}$$

We also define $G_i^t(x_i)$ as in (4), now for $i = 1, \ldots, I$.

Summary of the optimal policy structure. For the more general models studied in this section, the optimal policy follows a similar two-stage structure as in the N-model. That is, in the

first stage, we decide, based on the $h\mu$ index, whether a certain pool is going to fully prioritize some non-primary class or just provide partial help. 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.

5.1. X-Model

The X-model has a similar network structure except that help 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 $\mu_{22} > \mu_{21}$ to reflect the slowdown effect. Following the development of the N-model, we first compare the $h\mu$ index. Note that there are four possible cases in total. Without loss of generality, we consider two possible cases (the other two cases can be implied from Cases I and II by swapping the class indices):

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 that in an otherwise identical N-model.

To 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. The condition $h_2\mu_{21} G_2^{t+G_1^t(q_1)+u}(\tilde{q}_2(t + G_1^t(q_1) + u)) \leq \phi_{21}$ in the definition of $P^t(q)$ follows the same rationale as the condition in Case IIa of Theorem 1. When this condition is satisfied, pool 1 stops providing partial help to class 2. We define $\bar{G}^t_{X,2}(q(t))$ as the time to empty queue 2 when accounting for the partial help from pool 1:

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

The first term in the definition of $\bar{G}^t_{X,2}(q(t))$ corresponds to the case when queue 2 does not receive any partial help from pool 1 and it is the same as in the N-model. The second term corresponds to the case when there is partial helping. Note that $G^t_1(q_1(t))$ is the time to empty queue 1 (before any partial helping can start), $P^t(q(t))$ is the partial helping period, and $\phi_{21}/(h_2\mu_{21})$ is the remaining time to empty queue 2 using only pool 2, which can be seen from the condition in the definition of $P^t(q)$. 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 4, the optimal control for* (10) *takes the following form.*

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

ia. If $G^t_1(q_1(t)) = 0$ and $h_2\mu_{21}G^t_2(q_2(t)) > \phi_{21}$, pool 1 provides partial help to class 2.

ib. Otherwise, pool 1 serves class 1 only.

For pool 2,

iia. If

$$h_1\mu_{12}G^t_1(q_1(t)) > h_2\mu_{22}\bar{G}^t_{X,2}(q(t)) - h_2\mu_{22}\frac{\mu_{12}\mu_{21}}{\mu_{11}\mu_{22}}P^t(q(t)) + \phi_{12}, \tag{11}$$

pool 2 gives priority to class 1.

iib. Otherwise, pool 2 serves class 2 only.

II. If $h_1\mu_{12} < h_2\mu_{22}$ and $h_2\mu_{21} < h_1\mu_{11}$, each pool prioritizes its own class. For $i = 1, 2$, if

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

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 $\bar{G}^t_{X,2}(q(t)) \leq G^t_2(q_2(t))$,

$$h_2\mu_{22}\bar{G}^t_{X,2}(q(t)) - h_2\mu_{22}\frac{\mu_{12}\mu_{21}}{\mu_{11}\mu_{22}}P^t(q(t)) + \phi_{12} \leq h_2\mu_{22}G^t_2(q_2(t)) + \phi_{12}.$$

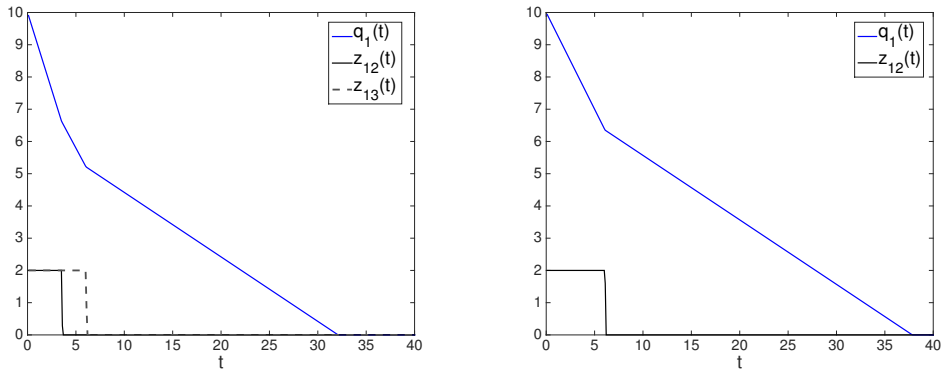
This implies that in the X-model, because pool 1 can help back class 2 later, pool 2 may provide more help to class 1 initially than in the N-model. We elaborate on this further in EC.2.

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.

For the exN1-model, the optimal policy still has a two-stage index-based look-ahead structure (see Theorem 5 in Appendix A). 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 5 compares the optimal trajectory of an exN1-model (plot a) with the optimal trajectory of a similar N-model (plot 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 5 for details of the numerical setting. 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 this 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 5 **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 6 in Appendix A). The key difference between the exN2-model and the N-model again 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.

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. Thus, 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

$$\theta h_{k(i^*)}\mu_{k(i^*),j} G^t_{k(i^*)}(q_{k(i^*)}(t)) - \phi_{k(i^*),j} > h_j \mu_{jj} G^t_j(q_j(t)). \tag{12}$$

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

$$\theta h_{k(i^*)}\mu_{k(i^*),j} G^t_{k(i^*)}(q_{k(i^*)}(t)) > \phi_{k(i),j}. \tag{13}$$

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 $G^t_i(q_i(t))$'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 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 $\theta = 1$.

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 *look-ahead policy* specified in Section 5.3, in the stochastic systems via simulation. We compare the performance of our proposed policy to that of several well-established benchmark policies in both the N-model and two more-general 5-by-5 networks. We demonstrate that, in the face of demand surges, the performance of our look-ahead policy, even using imperfect estimation of arrival rates (with time-varying and temporally-correlated prediction errors, a limited look-ahead time window, or prediction delays), is superior to that of the $c\mu$ rule or the maximum pressure policy or their adapted versions that account for the overflow costs. These numerical results suggest that the insights generated from our fluid analysis of parsimonious models are robust and useful for routing decisions in complex systems under various imperfect demand prediction scenarios.

To calibrate the simulation model, we consider settings motivated by hospital inpatient flow (Shi et al. 2016). 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 (Dong et al. 2019, Song et al. 2019). In particular, 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 what follows, we first present results for our main model (N-model with perfect information) and then results for more complicated networks and imperfect predictions.

6.1. Performance Comparison in N-model: Value of Proactive Routing

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

$$\lambda_1(t) = 8 \times \mathbf{1}\{t < 40\} + 4 \times \mathbf{1}\{t \geq 40\}$$

and $\lambda_2(t) = 3$. 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 test a large combination of settings by varying the following parameters: (i) the surge arrival rate of class 1, $\max \lambda_1(t) \in \{6, 8, 10\}$; (ii) the arrival rate of class 2, $\lambda_2(t) = \lambda_2 \in \{3, 3.5, 4, 4.5\}$; (iii) the surge duration $\kappa \in [20, 100]$; (iv) the initial states $X_1(0) \in \{40, 60, 80\}$ and $X_2(0) \in \{70, 90, 110, 130\}$. These parameter combinations lead to different levels of

system congestion. In general, the system dynamics are closer to the fluid limit when the system is more congested. Under a given congestion scenario, 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.

6.1.1. Benchmark policies We compare five scheduling policies: (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 that takes the overflow cost into account (ModCmu); and (v) the modified maximum pressure policy that takes the overflow cost into account (ModMaxP).

For the **modified $c\mu$ rule**, we prioritize different classes according to the following index (from high to low): $h_i\mu_{ij} - \phi_{ij}\mu_{ij}$. Similarly, for the **modified maximum pressure policy**, we prioritize different classes according to the following index (from high to low): $h_i X_i(t)\mu_{ij} - \phi_{ij}\mu_{ij}$, with $h_i X_i(t)\mu_{ij}$ being the index in the original maximum pressure policy. The intuition of adjusting the original indices in both policies by the term $\phi_{ij}\mu_{ij}$ is to maximize the instantaneous cost reduction rate under the preemptive service setting (with μ_{ij} being the rate of clearing a customer). As discussed in Section 3, comparing our proposed policy with the modified maximum pressure policy, the main difference is that the latter 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.

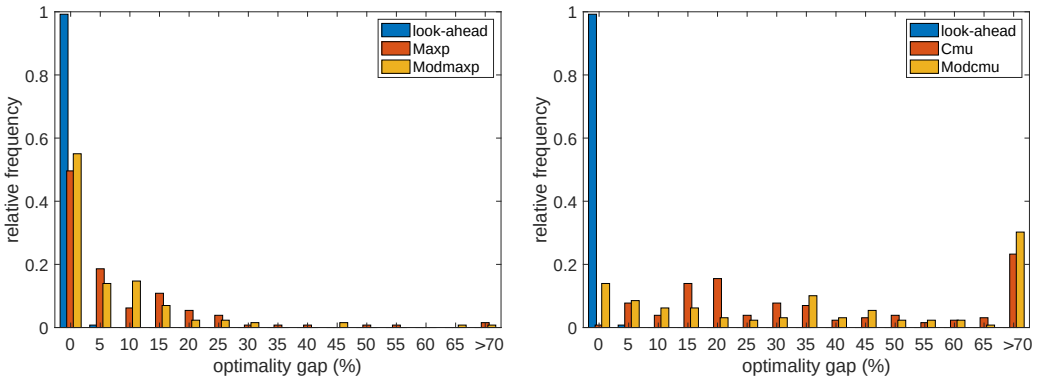
6.1.2. Robust performance Table 1 shows the cost comparison among the five policies in the baseline setting. Our proposed look-ahead policy performs significantly better than the $c\mu$ and the modified $c\mu$ rules. The maximum pressure policy and its modified version perform better than the $c\mu$ 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 6 plots a histogram of the optimality gap among all the tested combinations of arrival rates, initial states, and overflow costs, as specified earlier. The optimality gap is defined as the relative cost difference between the investigated policy and the best-performing policy in the corresponding parameter setting. It is clear from the figure that our policy always performs the best or near the 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 practice. 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 1 Expected total cost for the baseline N-model under different scheduling policies. The costs shown in the table are in units of 10^4. "SE" stands for the standard error for the average total cost (holding + overflow). Parameter setting: $h = (1.5, 1)$**,** $\phi_{12} = \phi$**,** $\lambda_2(t) = 3$**,** $\lambda_1(t) = 8 \times \mathbf{1}\{t < 40\} + 4 \times \mathbf{1}\{t \geq 40\}$**,** $X(0) = (60, 70)$**.**

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. On the other hand, the two maximum pressure policies tend to have a better performance 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 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*, for which we take a deeper dive in Section 6.1.3.



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

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

Figure 6 Histogram of the optimality gap. Parameter setting combinations are described in the main text.

6.1.3. Reactive versus proactive: the value of future information The modified maximum pressure policy is the best-performing benchmark policy in Table 1, i.e., it performs better than other benchmark policies in most scenarios tested in Figure 6. From this, it may appear that knowing the future arrival information is not as beneficial as one would expect. However, a closer investigation into different arrival rate 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:

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

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 7 we compare two sample paths when $\phi = 2$, one under our policy and the other under the modified maximum pressure policy.

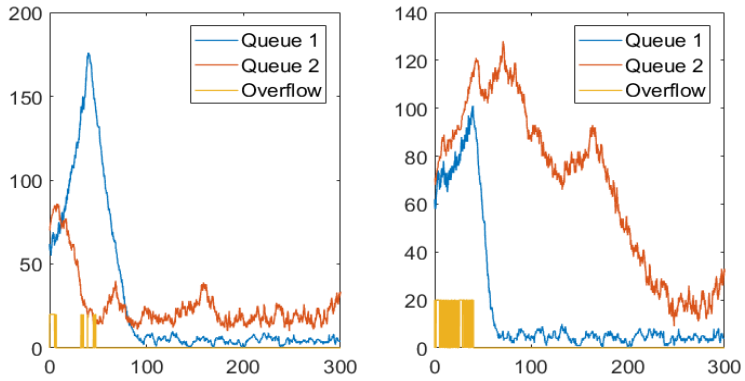


Figure 7 **Sample path comparison between our proposed policy (left) and the modified maximum pressure policy (right). ($\lambda_1(t) = 8 \times \mathbf{1}\{t < 40\} + 1 \times \mathbf{1}\{t \geq 40\}$, $\lambda_2(t) = 3 \times \mathbf{1}\{t < 40\} + 4.5 \times \mathbf{1}\{t \geq 40\}$. Other parameters are the same as the baseline with $\phi = 2$.)**

We see that under the modified maximum pressure policy, pool 2 helps class 1 throughout the demand surge period (till $t = 40$). This is because class 1 has a large queue and the policy is reacting to 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 information 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 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. For the parameter setting in Figure 7, the average costs are 1.11×10^4 under our policy versus 1.68×10^4 under the modified maximum pressure policy – 50% higher than ours. This performance gap further enlarges to over 150% as ϕ increases to 25. More generally, the maximum pressure policies can perform much worse than our policy as $\lambda_2(t)$ gets closer to 5 after the demand surge of class 1 is over, due to over-helping. This indicates that the performance of "arrival-agnostic" policies can vary a lot depending on the arrival rate patterns, which highlights the value of our look-ahead policy in a nonstationary environment.

Other variants of index adjustment. We remark that the adjustments to the $c\mu$ and maximum pressure policies are heuristic. It is possible to motivate other heuristics such as adjusting by 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 choosing seconds versus hours to be the time unit, the resulting policy can vary drastically from zero overflow to full-sharing. The main takeaway is that the optimal format to incorporate the overflow costs is highly nontrivial. The performance can be arbitrarily bad when using the wrong format, highlighting the necessity of properly comparing the overflow cost with the holding cost and rigorously deriving optimal routing policies in the presence of demand surges.

6.2. Extensions to General Multi-class Multi-pool Systems

In this section, we test our proposed heuristic policy as given in Section 5.3. In the interest of space, we focus on the results for two 5-by-5 networks. (See EC.3.2. for additional numerical experiments for the X-model, which shows that our heuristic policy has a performance comparable to that of the optimal fluid-translated policy.) 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.

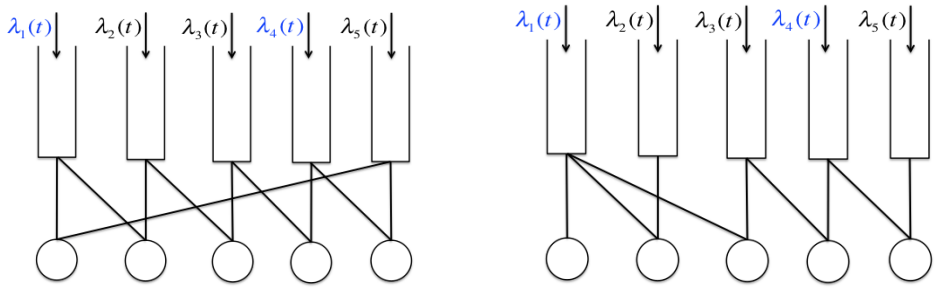
We set the holding costs to be $(1.5, 1, 1, 1.5, 1)$. The overflow costs ϕ are the same for all overflow assignments. We consider two arrival rate settings. For the first setting, the arrival rates are

$$\lambda_1(t) = \begin{cases} 12, & t < 40 \\ 4.5, & t \geq 40 \end{cases} \text{ and } \lambda_4(t) = \begin{cases} 8, & t < 40 \\ 4, & t \geq 40 \end{cases}, \tag{14}$$

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

$$\lambda_1(t) = \begin{cases} 12, & t < 40 \\ 2, & t \geq 40 \end{cases} \text{ and } \lambda_2(t) = \begin{cases} 3, & t < 40 \\ 4.5, & t \geq 40 \end{cases} \tag{15}$$

respectively, 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 8. The first network has a closed-chain structure, which is a commonly advocated flexibility architecture in supply chain and manufacturing applications (Simchi-Levi and Wei 2012, Tekin et al. 2002).



(a) Network I (b) Network II

Figure 8 Two network structures. Classes 1 and 4 have demand surge in both arrival rate settings.

Table 2 compares the cost under our policy and two of the most competitive benchmark policies for different problem instances. 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 network structures and both arrival rate settings. The modified $c\mu$ rule does not overflow, and thus performs the worst 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 rate setting. Similar to what we explained in Section 6.1.3, 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.

6.3. Impact of Prediction Error in Arrival Rates

In Section 4, we have analytically studied the effects of prediction errors in the N-model. In this section, we numerically investigate the effect of prediction errors in the stochastic system. Throughout this section, we assume there is a known constant nominal arrival rate, which is the arrival

[1] Tuning, in general, improves the cost from the untuned version by 0.5% to 4%; see EC.3.2. for the detailed results for the X-model and the 5-by-5 network.

		Look-ahead	ModMaxP	ModCmu	Look-ahead	ModMaxP	ModCmu
		Network Structure I Arrival Rate Setting I			Network Structure I Arrival Rate Setting II		
$\phi=2$	Holding	3.21	3.27	11.08	3.00	4.26	5.94
	Overflow	0.52	0.59	0.00	0.29	0.29	0.00
	Total	3.73	3.87	11.08	3.29	4.55	5.94
	SE	0.007	0.007	0.014	0.005	0.008	0.010
$\phi=10$	Holding	3.45	3.27	11.08	3.36	4.22	5.94
	Overflow	2.05	2.75	0.00	0.89	1.42	0.00
	Total	5.50	6.03	11.08	4.25	5.64	5.94
	SE	0.008	0.009	0.014	0.006	0.009	0.010
$\phi=25$	Holding	4.29	3.31	11.08	4.20	4.14	5.94
	Overflow	3.57	6.27	0.00	1.10	3.46	0.00
	Total	7.86	9.58	11.08	5.31	7.61	5.94
	SE	0.012	0.013	0.014	0.007	0.010	0.010
		Network Structure II Arrival Rate Setting I			Network Structure II Arrival Rate Setting II		
$\phi=2$	Holding	3.00	3.27	11.08	2.93	3.69	5.94
	Overflow	0.48	0.42	0.00	0.25	0.34	0.00
	Total	3.48	3.69	11.08	3.18	4.03	5.94
	SE	0.005	0.007	0.014	0.005	0.008	0.010
$\phi=10$	Holding	3.07	3.27	11.08	3.15	3.64	5.94
	Overflow	1.98	2.06	0.00	0.93	1.68	0.00
	Total	5.05	5.34	11.08	4.09	5.33	5.94
	SE	0.007	0.008	0.014	0.006	0.009	0.010
$\phi=25$	Holding	3.57	3.27	11.08	4.19	3.55	5.94
	Overflow	3.77	5.03	0.00	1.11	4.05	0.00
	Total	7.34	8.31	11.08	5.30	7.61	5.94
	SE	0.009	0.010	0.014	0.007	0.010	0.010

Table 2 Simulation costs for the 5-by-5 model. The costs shown in the table are in units of 10^4. "SE" stands for the standard error for the 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 $X(0)=(30,40,50,60,70)$. Arrival rate settings I and II are given in (14) and (15) respectively.

rate during the normal days (non-surge period), denoted as λ_i^0. We focus on larger values of ϕ, i.e., $\phi=10,25$. We compare the performance of our policy with two of the most competitive benchmarks: the modified maximum pressure policy and the modified $c\mu$ rule.

6.3.1. Error in estimation We consider two (true) arrival rate settings, one specified in Table 1 and the other in Figure 7. We assume that the arrival rate for class 2 is known, but the arrival rate for class 1 suffers from prediction error during the surge. That is, at a decision time t, the estimated arrival rate for the future is $\tilde{\lambda}_1(s|t)=\lambda_1(s)+\epsilon_1(s|t)$ for $t\le s\le\kappa_1$ and $\tilde{\lambda}_1(s|t)=\lambda_1^0$ for $s>\kappa_1$. To reflect the fact that forecasts tend to be more accurate for the near future than for the distant future, we allow $\epsilon_1(s|t)$ to grow in magnitude with time. We also allow correlations in $\epsilon_1(s|t)$'s across different time periods.

Specifically, we define a sequence of random variables $\{R_k:k\ge1\}$ with $R_1\sim N(0,1)$, and $R_k=\rho R_{k-1}+\sqrt{1-\rho^2}\nu_k$ for $k\ge2$, where ν_k's are independent and identically distributed standard

normal. The estimated rate is $\tilde{\lambda}_1(s|t) = \lambda_1(s) + \sqrt{\ell\lceil (s-t)/\ell \rceil} R_{\lceil (s-t)/\ell \rceil}$ at decision time t, where ℓ is the interval for discretization and $\lceil \cdot \rceil$ is the ceiling function. In the experiments, we set $\ell = 5$. This form of estimated arrival rates is motivated by autoregressive models and deep latent Gaussian models (Tzen and Raginsky 2019). Note that the standard deviation of the estimation error can be as large as $\sqrt{40}$, which is of the same order as the true arrival rate. In other words, the setting goes beyond the analytical setting in Section 4. When $\rho < 0$, successive R_k values are negatively correlated, in which case the errors tend to cancel out each other when integrating the arrival rates to calculate $\tilde{G}_1^t$, leading to a more accurate estimate of G_1^t. When $\rho > 0$, successive R_k values are positively correlated, and there is less of a cancellation effect when calculating $\tilde{G}_1^t$. Thus, as ρ increases, we would expect a less accurate estimate of G_1^t.

Table 3 summarizes the simulation results when using the estimated arrival rates with different values of ρ in our look-ahead policy. As expected, the performance of our look-ahead policy tends to deteriorate as ρ increases. (The differences are quite small though.) Nevertheless, our policy performs better than the modified maximum pressure policy (best benchmark policy) in the first arrival rate setting, and comparable to or better than the modified $c\mu$ rule (best benchmark policy) in the second arrival rate setting. Importantly, the robust performance of our policy is in contrast to the swaying performance of the two benchmark policies: they could perform well in one setting but poorly in another.

The good performance of our policy, even under the presence of large prediction errors, benefits from two aspects of our policy. First, taking a closer look at the look-ahead function $\tilde{G}_i^t$, we note that even though $\tilde{\lambda}_i(s)$ can be far from $\lambda_i(s)$ for some s, as long as $\int_t^{t+\Delta} \tilde{\lambda}_i(s)ds$ is close to $\int_t^{t+\Delta} \lambda_i(s)ds$, $\tilde{G}_i^t$ will be close to G_i^t. When taking the integration, some of the estimation errors may cancel out. Second, $\tilde{G}_i^t$ is dynamically updated as the queue builds up and as more information becomes available. This helps dynamically adjust the policy to respond to the load of the system.

6.3.2. Limited look-ahead time window We next study the impact of a limited look-ahead time window. Using the same notation introduced in Section 3.3, we assume that for a given time window W, at time t, only the future arrival rate up to time $t + W$ is known. We test the settings with two demand surges, where a limited look-ahead time may have a larger impact on performance than the single-surge setting since the policy might not be able to anticipate the second demand surge when planning during the first demand surge.

We consider two arrival rate settings similar to those studied in Table 3, except that we break the initial demand surge into two separate surges. Table 4 summarizes the simulation results under these two arrival rate settings, with the detailed arrival rates specified in the caption. We test three different values of W: $0, 5$, and 10. Note that these time windows are smaller than or equal to the

	ρ	-1	0	1	MMP	MC	-1	0	1	MMP	MC
		Arrival Rate Setting I					Arrival Rate Setting II				
$\phi=10$	Holding	1.104	1.105	1.107	1.108	2.754	1.111	1.110	1.120	1.543	1.282
	Overflow	0.567	0.569	0.569	0.619	0.000	0.119	0.120	0.118	0.449	0.000
	Total	1.671	1.674	1.676	1.727	2.754	1.229	1.230	1.238	1.992	1.282
	SE	0.004	0.004	0.004	0.004	0.008	0.002	0.002	0.002	0.006	0.002
$\phi=25$	Holding	1.258	1.251	1.260	1.118	2.754	1.278	1.278	1.252	1.466	1.282
	Overflow	1.034	1.054	1.054	1.504	0.000	0.007	0.007	0.050	1.061	0.000
	Total	2.292	2.305	2.314	2.622	2.754	1.285	1.284	1.302	2.527	1.282
	SE	0.005	0.005	0.005	0.005	0.008	0.002	0.002	0.002	0.006	0.002

Table 3 Simulation costs for the N-model when using our policy with ρ between -1 and 1, modified maximum pressure policy ('MMP'), and modified $c\mu$ policy ('MC'). The costs shown in the table are in units of 10^4. "SE" stands for the standard error for the total cost. Parameter setting: $h=(1.5,1)$, $\phi_{12}=\phi$, $X(0)=(60,70)$. The first arrival setting is the same as that in Table 1 and the second the same as that in Figure 7.

length of the interval between the two demand surges, so that the policy does not "know" about the second demand surge during the first one. To implement the look-ahead policy, we use the adapted policy introduced in Section 3.3, i.e., when the arrival rate information is not available, we use the nominal rate λ_i^0. As W increases, more of the arrival rate information is available for surge planning, and, hence, the value of $\hat{G}_1^t(X_1(t))$ increases. Consequently, more help is offered, which explains why the overflow cost increases while the holding cost generally decreases. $W=\infty$ refers to the full information case. As expected, the total cost generally decreases with W, but the performance change is quite small. It is worth noting that our policy, even with a small value of W, generally performs much better than the two benchmark policies. These numerical results demonstrate that our policy is robust to limited future arrival rate information in different parameter settings, which is very desirable for practical implementations.

	W	0	5	10	∞	MMP	MC	0	5	10	∞	MMP	MC
		Arrival Rate Setting I						Arrival Rate Setting II					
$\phi=10$	Holding	1.011	0.999	0.992	1.002	0.995	2.553	1.045	1.019	0.979	0.902	1.080	1.049
	Overflow	0.537	0.543	0.549	0.563	0.598	0.000	0.003	0.023	0.052	0.111	0.361	0.000
	Total	1.549	1.543	1.542	1.565	1.593	2.553	1.048	1.041	1.031	1.013	1.442	1.049
	SE	0.004	0.004	0.004	0.004	0.004	0.007	0.002	0.002	0.002	0.002	0.004	0.002
$\phi=25$	Holding	1.270	1.237	1.215	1.152	1.006	2.553	1.049	1.049	1.049	1.049	1.030	1.049
	Overflow	0.896	0.920	0.939	1.000	1.452	0.000	0.000	0.000	0.000	0.000	0.841	0.000
	Total	2.166	2.158	2.154	2.152	2.458	2.553	1.049	1.049	1.049	1.049	1.871	1.049
	SE	0.005	0.005	0.005	0.005	0.005	0.007	0.002	0.002	0.002	0.002	0.005	0.002

Table 4 Simulation costs for the N-model when using our policy with a limited prediction window $W=0,5,$ and 10. The costs shown in the table are in units of 10^4. "SE" stands for the standard error for the total cost. Parameter setting: $h=(1.5,1)$, $\phi_{12}=\phi$, $X(0)=(60,70)$. For arrival rate setting I, $\lambda_1(t)=8\times\mathbf{1}\{t<20\}+4\times\mathbf{1}\{20\le t<30\}+8\times\mathbf{1}\{30\le t<50\}+4\times\mathbf{1}\{t\ge 50\}$ and $\lambda_2(t)=3$. For arrival rate setting II, $\lambda_1(t)=8\times\mathbf{1}\{t<20\}+1\times\mathbf{1}\{20\le t<30\}+8\times\mathbf{1}\{30\le t<50\}+1\times\mathbf{1}\{t\ge 50\}$ and $\lambda_2(t)=3\times\mathbf{1}\{t<20\}+4.5\times\mathbf{1}\{20\le t<30\}+3\times\mathbf{1}\{30\le t<50\}+4.5\times\mathbf{1}\{t\ge 50\}$.

6.3.3. Prediction delay In practice, it is possible that there is a delay (time lag) in the estimated arrival rate, e.g., when the arrival rate is estimated based on the historical average. To examine the impact of a potential prediction delay on the performance of our algorithm, we assume for a given delay d, at a decision time $t < d$, the predicted future arrival rate is the constant nominal arrival rate, i.e., $\tilde{\lambda}_1(s|t) = \lambda_1^0$ for $s \geq t$. At a decision time $t \geq d$, the predicted future arrival rate is $\tilde{\lambda}_1(s|t) = \lambda_1(s-d)$ for $s \geq t$, i.e., there is a constant lag of d between the predicted arrival rate and the actual arrival rate. In this case, there is a delay of d to detect the start and the end of the demand surge.

We consider a piecewise-linear arrival rate function for class 1:

$$\lambda_1(t) = (5 + 0.25t) \times \mathbf{1}\{t < 20\} + (10 - 0.3(t-20)) \times \mathbf{1}\{20 \leq t < 40\} + 4 \times \mathbf{1}\{t \geq 40\} \tag{16}$$

and $\lambda_2(t) = 3$. All other parameters are the same as in the baseline setting. We test the performance of our look-head policy under different values of d and the results are summarized in Table 5. We observe that the performance of the look-ahead policy deteriorates as d increases. We tend to do more overflow as d increases, because the delay causes the system to continue to overflow for longer than it would otherwise. When $\phi = 10$, the look-ahead policy performs better than the modified maximum pressure policy (best benchmark) when $d \leq 5$. However, when $d \geq 10$, the modified maximum pressure policy performs slightly better. When $\phi = 25$, the look-ahead policy performs better than the modified $c\mu$ rule (best benchmark) when $d \leq 20$. Only when $d = 30$ does the look-ahead policy perform slightly worse than the modified $c\mu$ rule. Overall, when the overflow cost is large (so that balancing between the cost and benefit from overflow is critical), our policy is more robust and preferred, even if there may be a large prediction delay. When the overflow cost is small yet there is a concern for large prediction delays, the arrival-agnostic policies could be a better choice.

	d	0	1	5	10	15	20	30	MMP	MC
$\phi = 10$	Holding	0.854	0.854	0.859	0.870	0.890	0.915	0.948	0.848	2.018
	Overflow	0.462	0.467	0.493	0.526	0.557	0.586	0.615	0.530	0.000
	Total	1.316	1.321	1.352	1.396	1.448	1.501	1.563	1.378	2.018
	SE	0.003	0.003	0.003	0.003	0.003	0.003	0.004	0.003	0.007
$\phi = 25$	Holding	1.031	1.017	0.961	0.919	0.913	0.908	0.902	0.855	2.018
	Overflow	0.751	0.771	0.879	0.987	1.040	1.096	1.255	1.282	0.000
	Total	1.781	1.788	1.840	1.906	1.954	2.004	2.156	2.137	2.018
	SE	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.007

Table 5 Simulation costs for the N-model when using our policy with prediction delay d between 0 and 30. The costs shown in the table are in units of 10^4. "SE" stands for the standard error for the total cost. Parameter setting: $h = (1.5, 1)$, $\phi_{12} = \phi$, $X(0) = (60, 70)$, $\lambda_1(t)$ is as in (16), $\lambda_2(t) = 3$.

7. Conclusion

In this paper, we study how to leverage demand forecasts in designing the optimal routing policy for systems with partial flexibility under demand surges. Our model incorporates salient features of service systems, such as general time-varying arrival rates with possibly multiple surges, efficiency loss and inconvenience costs associated with overflow, and various service compatibility architectures (e.g., the N- and X-models). These features make our research problem highly nontrivial. We characterize how to properly incorporate demand forecasts that are potentially imperfect into the routing policies. We study transient fluid control for the N-model, the X-model, and two extensions of the N-model, and we explicitly characterize the optimal control. All of these fluid-based policies have a two-stage index-based structure with a look-ahead component that takes future arrival rates into account. Based on the insights from the fluid analysis, we propose a two-stage index-based look-ahead policy for general stochastic systems. Via extensive simulation results, we quantify the value of future arrival rate information and show that our policy is able to achieve superior performance to other benchmark policies in various parallel server networks. In particular, our policy is robust under various scenarios of prediction errors, due to the built-in resilience of the index structure.

Several future extensions may be considered. One is to jointly optimize the system's architectural design and the real-time routing policy, e.g. N-model versus X-model (see EC.2. for more discussions). Another is to study the effect of more general prediction errors. Our theoretical analysis assumes the prediction error is of a smaller order than the arrival rate. Our numerical experiments show that our policy performs well even with relatively large prediction errors, benefiting from the built-in resilience of the index structure. Investigating how to optimally adjust the routing policy to account for various forms of prediction errors would be an interesting future research direction.

Acknowledgments

The authors thank the anonymous associated editor, and reviewers for their constructive suggestions which help improve the exposition of the paper greatly. J. Dong was supported by National Science Foundation, Division of Civil, Mechanical, and Manufacturing Innovation (CMMI) Grant 1994209.

References

Aksin, Zeynep, Mor Armony, Vijay Mehrotra. 2007. The modern call center: A multi-disciplinary perspective on operations management research. *Production and Operations Management* **16**(6) 665–688.

Akşin, Zeynep, Fikri Karaesmen. 2007. Characterizing the performance of process flexibility structures. *Operations Research Letters* **35**(4) 477–484.

Armony, Mor, Amy R. Ward. 2010. Fair Dynamic Routing in Large-Scale Heterogeneous-Server Systems. *Operations Research* **58**(3) 624–637.

Ata, B., X. Peng. 2020. An optimal callback policy for general arrival processes: A pathwise analysis. *Operations Research* **68**(2) 1–21.

Ata, B., J.A. Van Mieghem. 2009. The value of partial resource pooling: Should a service network be integrated or product-focused? *Management Science* **55**(1) 115–131.

Baas, Stef, Sander Dijkstra, Aleida Braaksma, Plom van Rooij, Fieke J Snijders, Lars Tiemessen, Richard J Boucherie. 2021. Real-time forecasting of covid-19 bed occupancy in wards and intensive care units. *Health Care Management Science* **24**(2) 402–419.

Bassamboo, Achal, Ramandeep S Randhawa, Jan A Van Mieghem. 2012. A little flexibility is all you need: on the asymptotic value of flexible capacity in parallel queuing systems. *Operations Research* **60**(6) 1423–1435.

Bassamboo, Achal, Ramandeep S. Randhawa, Assaf Zeevi. 2010. Capacity Sizing Under Parameter Uncertainty: Safety Staffing Principles Revisited. *Management Science* **56**(10) 1668–1686.

Bassamboo, Achal, Assaf Zeevi. 2009. On a data-driven method for staffing large call centers. *Operations Research* **57**(3) 714–726.

Bäuerle, Nicole. 2000. Asymptotic optimality of tracking policies in stochastic networks. *Annals of Applied Probability* **10**(4) 1065–1083.

Bäuerle, Nicole. 2002. Optimal control of queueing networks: An approach via fluid models. *Advances in Applied Probability* 313–328.

Bell, S. L., R. J. Williams. 2001. 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* **11**(3) 608–649.

Best, Thomas J, Burhaneddin Sandıkçı, Donald D Eisenstein, David O Meltzer. 2015. Managing hospital inpatient bed capacity through partitioning care into focused wings. *Manufacturing & Service Operations Management* **17**(2) 157–176.

Borst, Sem, Avi Mandelbaum, Martin I. Reiman. 2004. Dimensioning Large Call Centers. *Operations Research* **52**(1) 17–34.

Bramson, Maury, Bernardo D'Auria, Neil Walton. 2021. Stability and instability of the maxweight policy. *Mathematics of Operations Research* **46**(4) 1611–1638.

Buyukkoc, C, P Varaiya, Jean Walrand. 1985. The cμ rule revisited. *Advances in applied probability* **17**(1) 237–238.

Chang, Junxia, Hayriye Ayhan, JG Dai, Cathy H Xia. 2004. Dynamic scheduling of a multiclass fluid model with transient overload. *Queueing Systems* **48**(3) 263–307.

Chen, Jinsheng, Jing Dong, Pengyi Shi. 2020. A survey on skill-based routing with applications to service operations management. *Queueing Systems* **96** 53–82.

Dai, J. G., J. Michael Harrison. 2020. *Processing Networks: Fluid Models and Stability*. Cambridge University Press.

Dai, J. G., Wuqin Lin. 2005. Maximum Pressure Policies in Stochastic Processing Networks. *Operations Research* **53**(2) 197–218.

Dai, J. G., Wuqin Lin. 2008. Asymptotic optimality of maximum pressure policies in stochastic processing networks. *The Annals of Applied Probability* **18**(6) 2239–2299.

Delana, Kraig, Nicos Savva, Tolga Tezcan. 2021. Proactive customer service: operational benefits and economic frictions. *Manufacturing & Service Operations Management* **23**(1) 70–87.

Dong, Jing, Pengyi Shi, Fanyin Zheng, Xin Jin. 2019. Off-service placement in inpatient ward network: Resource pooling versus service slowdown. Working paper.

Garnett, Ofer, Avishai Mandelbaum. 2000. An introduction to skills-based routing and its operational complexities. Teaching notes.

Grass, Dieter, Jonathan Caulkins, Gustav Feichtinger, Gernot Tragler, Doris Behrens. 2008. *Optimal Control of Nonlinear Processes: With Applications in Drugs, Corruption, and Terror*. Springer.

Graves, S.C., B.T. Tomlin. 2003. Process flexibility in supply chains. *Management Science* **49**(7) 907–919.

Gurvich, Itai, James Luedtke, Tolga Tezcan. 2010. Staffing call centers with uncertain demand forecasts: A chance-constrained optimization approach. *Management Science* **56**(7) 1093–1115.

Hampshire, Robert C, William A Massey. 2010. Dynamic optimization with applications to dynamic rate queues. *Risk and Optimization in an Uncertain World*. INFORMS, 208–247.

Harrison, J Michael. 1998. Heavy traffic analysis of a system with parallel servers: asymptotic optimality of discrete-review policies. *The Annals of Applied Probability* **8**(3) 822–848.

Hartl, Richard F., Suresh P. Sethi, Raymond G. Vickson. 1995. A Survey of the Maximum Principles for Optimal Control Problems with State Constraints. *SIAM Review* **37**(2) 181–218. Publisher: Society for Industrial and Applied Mathematics.

Hu, Yue, Carri W. Chan, Jing Dong. 2019. Optimal scheduling of proactive service with customer deterioration and improvement. Working paper.

Hu, Yue, Carri W Chan, Jing Dong. 2021. Prediction-driven surge planning with application in the emergency department. Working paper.

Ibrahim, Rouba, Pierre L'Ecuyer. 2013. Forecasting call center arrivals: Fixed-effects, mixed-effects, and bivariate models. *Manufacturing & Service Operations Management* **15**(1) 72–85.

Iglehart, Donald L, Ward Whitt. 1970. Multiple channel queues in heavy traffic. i. *Advances in Applied Probability* **2**(1) 150–177.

Institute for Health Metrics and Evaluation. 2022. Covid-19 projections. `https://covid19.healthdata.org/united-states-of-america`. Accessed: 2022-05-22.

Jasin, Stefanus, Sunil Kumar. 2012. A re-solving heuristic with bounded revenue loss for network revenue management with customer choice. *Mathematics of Operations Research* **37**(2) 313–345.

Liu, Yunan, Ward Whitt. 2011. A network of time-varying many-server fluid queues with customer abandonment. *Operations research* **59**(4) 835–846.

Maglaras, Constantinos. 2000. Discrete-review policies for scheduling stochastic networks: Trajectory tracking and fluid-scale asymptotic optimality. *Annals of Applied Probability* **10**(3) 897–929.

Maman, Shimrit, Avishai Mandelbaum, S Zeltyn. 2009. Uncertainty in the demand for service: The case of call centers and emergency departments. Ph.D. thesis, Technion-Israel Institute of Technology, Faculty of Industrial and Management.

Mandelbaum, A., William A Massey, Martin I Reiman. 1998. Strong approximations for markovian service networks. *Queueing Systems* **30**(1) 149–201.

Mandelbaum, A., M.I. Reiman. 1998. On pooling in queueing networks. *Management Science* **44**(7) 971–981.

Mandelbaum, A., Alexander L. Stolyar. 2004. Scheduling Flexible Servers with Convex Delay Costs: Heavy-Traffic Optimality of the Generalized cμ-Rule. *Operations Research* **52**(6) 836–855.

Naor, Pinhas. 1969. The regulation of queue size by levying tolls. *Econometrica: journal of the Econometric Society* 15–24.

Pinker, Edieal J, Robert A Shumsky. 2000. The efficiency-quality trade-off of cross-trained workers. *Manufacturing & Service Operations Management* **2**(1) 32–48.

Sethi, Suresh P., Gerald L. Thompson. 2000. *Optimal Control Theory*. Springer.

Shi, Pengyi, Mabel C Chou, Jim G Dai, Ding Ding, Joe Sim. 2016. Models and insights for hospital inpatient operations: Time-dependent ed boarding time. *Management Science* **62**(1) 1–28.

Shu, Jia, Mabel C Chou, Qizhang Liu, Chung-Piaw Teo, I-Lin Wang. 2013. Models for effective deployment and redistribution of bicycles within public bicycle-sharing systems. *Operations Research* **61**(6) 1346–1359.

Simchi-Levi, David, Yehua Wei. 2012. Understanding the Performance of the Long Chain and Sparse Designs in Process Flexibility. *Operations Research* **60**(5) 1125–1141.

Smith, D.R., W. Whitt. 1981. Resource sharing for efficiency in traffic systems. *Bell Systems Technical Journal* **60**(1) 39–55.

Song, H., A.L. Tucker, K.L. Murrell. 2015. The diseconomies of queue pooling: An empirical investigation of emergency department length of stay. *Management Science* **61**(12) 3032–3053.

Song, H., Anita L. Tucker, Ryan Graue, Sarah Moravick, Julius J. Yang. 2019. Capacity Pooling in Hospitals: The Hidden Consequences of Off-Service Placement. *Management Science* **66**(9) 3825–3842.

Stein, Clifford, Van-Anh Truong, Xinshang Wang. 2020. Advance service reservations with heterogeneous customers. *Management Science* **66**(7) 2929–2950.

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

Tekin, Eylem, Wallace J Hopp, Mark P Van Oyen. 2002. Benefits of skill chaining in production lines with cross-trained workers. *Manufacturing & Service Operations Management* **4**(1) 17–20.

Tsitsiklis, J.N., K Xu. 2012. On the power of (even a little) resource pooling. *Stochastic Systems* **2**(1) 1–66.

Tzen, Belinda, Maxim Raginsky. 2019. Neural stochastic differential equations: Deep latent gaussian models in the diffusion limit. *arXiv preprint arXiv:1905.09883* .

van Mieghem, Jan A. 1995. Dynamic Scheduling with Convex Delay Costs: The Generalized $c\mu$ Rule. *The Annals of Applied Probability* **5**(3) 809–833.

van Mieghem, Jan A. 1998. Investment strategies for flexible resources. *Management Science* **44**(8) 1071–1078.

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

Yom-Tov, Galit B, Avishai Mandelbaum. 2014. Erlang-r: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing & Service Operations Management* **16**(2) 283–299.

Appendix A: 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 5. We make the following assumptions about the arrival rate functions.

ASSUMPTION 4. *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*

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

The proofs of the results in this section are in the E-companion.

A.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 4(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,

$$\bar{G}^t_{\text{exN1},1,3}(q) = G_3^t(q_3) + P_3^t(q) + G_1^{t+G_3^t(q)+P_3^t(q)}(\tilde{q}_1(t+G_3^t(q)+P_3^t(q))).$$

Similar to before, $G_1^{t+G_3^t(q)+P_3^t(q)}(\tilde{q}_1(t+G_3^t(q)+P_3^t(q))) = \frac{\phi_{31}}{h_3\mu_{31}}$ if $P_3^t(q) > 0$.

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 5. *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 5. *For the exN1-model, under Assumptions 4 and 5, the optimal control for* (10) *takes the following form. Pool 1 serves as many class 1 customers as possible.*

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

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

$$h_1\mu_{12}\bar{G}^t_{exN1,1,3}(q(t)) > h_2\mu_{22}G_2^t(q_2(t)) + \phi_{12}, \tag{17}$$

pool 2 gives priority to class 1.

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

$$h_1\mu_{12}G_1^t(q_1(t)) > h_2\mu_{22}G_2^t(q_2(t)) + \phi_{12}, \tag{18}$$

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*

$$h_1\mu_{13}\bar{G}^t_{exN1,1,2}(q(t)) > h_3\mu_{33}G_3^t(q_3(t)) + \phi_{13}, \tag{19}$$

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*

$$h_1\mu_{13}G_1^t(q_1(t)) > h_3\mu_{33}G_3^t(q_3(t)) + \phi_{13}, \tag{20}$$

pool 3 gives priority to class 1.

iiic. Otherwise, pool 3 serves class 3 only.

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

iia. If

$$G_2^t(q_2(t)) = 0 \text{ and } h_1\mu_{12}\bar{G}^t_{exN1,1,3}(q(t)) > \phi_{12}, \tag{21}$$

pool 2 provides partial help to class 1.

iib. Otherwise, pool 2 serves class 2 only.

For pool 3,

iiia. If

$$h_1\mu_{13}\bar{G}^t_{exN1,1,2}(q(t)) > h_3\mu_{33}G_3^t(q_3(t)) + \phi_{13}, \tag{22}$$

pool 3 prioritizes class 1.

iiib. Otherwise, pool 3 serves class 3 only.

III. If $h_1\mu_{12} < h_2\mu_{22}$ *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

$$G_2^t(q_2(t)) = 0, \text{ and } h_1\mu_{12}\bar{G}^t_{exN1,1,3}(q(t)) > \phi_{12}, \tag{23}$$

pool 2 provides partial help to class 1.

iib. Otherwise, pool 2 serves class 2 only.

For pool 3,

iiia. If

$$G_3^t(q_3(t)) = 0 \text{ and } h_1\mu_{13}\bar{G}^t_{exN1,1,2}(q(t)) > \phi_{13}, \tag{24}$$

pool 3 provides partial help to class 1.

iiic. Otherwise, pool 3 serves class 3 only.

To provide more intuition behind Theorem 5, 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 (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 (17).

Comparing the exN1-model to the N-model, we note that $\bar{G}^t_{\text{exN1},1,3}(q(t)) \le G^t_1(q_1(t))$. 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.

A.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 4(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:

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

where for $\tilde{q}$: $\tilde{q}(t)=q$; for $s\ge t$, 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 6. *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 6. *For the exN2-model, under Assumptions 4 and 6, the optimal control for* (10) *takes the following form. Pools 2 and 3 serve their primary classes as much as possible.*

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

a. If

$$h_2\mu_{21} G_2^t(q_2(t)) > h_1\mu_{11}\bar{G}_{exN2,1}^t(q(t)) + (h_3\mu_{31} - h_1\mu_{11})F^t(q(t)) + \phi_{21}, \tag{25}$$

pool 1 gives priority to class 2.

b. Otherwise, if

$$h_3\mu_{31} G_3^t(q_3(t)) > h_1\mu_{11} G_1^t(q_1(t)) + \phi_{31}, \tag{26}$$

pool 1 gives priority to class 3.

c. Otherwise, pool 1 serves class 1 only.

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

a. If

$$h_2\mu_{21} G_2^t(q_2(t)) > h_1\mu_{11} G_1^t(q_1(t)) + h_3\mu_{31} P^t(q(t)) + \phi_{21}, \tag{27}$$

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

a. If

$$G_1^t(q_1(t)) = 0 \text{ and } h_2\mu_{21} G_2^t(q_2(t)) > h_3\mu_{31} P^t(q(t)) + \phi_{21}, \tag{28}$$

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 6, 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 (25).

Comparing the exN2-model to the N-model, we note that $\bar{G}^t_{\text{exN2},1}(q(t)) \geq G^t_1(q_1(t))$, and

$$h_1\mu_{11}\bar{G}^t_{\text{exN2},1}(q(t)) + (h_3\mu_{31} - h_1\mu_{11})F^t(q(t)) + \phi_{21} \geq h_1\mu_{11}G^t_1(q_1(t)) + \phi_{21}.$$

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.

Appendix B: Proof of Optimal Fluid Control Results

The proof of Theorem 1 and subsequent fluid optimal control results (Theorems 4 - 6) 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 (Theorems 4 - 6) follow similar lines of analysis and are provided in the E-companion.

B.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 $g(q) = (g_1(q), \ldots, g_I(q))$. 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

$$\begin{aligned} \inf_z \quad & \int_0^\infty F\left(q(t), z(t)\right) dt \\ \text{s.t.} \quad & \dot{q}(t) = f\left(q(t), z(t), t\right), \quad q(0) = q_0 \\ & g(q(t)) \leq 0 \\ & l(z(t)) \leq 0 \end{aligned} \tag{29}$$

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 (29) 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) 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 Hartl et al. (1995) for (29).

THEOREM 7 **(Arrow-type sufficient condition)**. *Let (q^*, z^*) be a feasible pair for the optimal control problem (29). 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. *Augmented Hamiltonian condition:*

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

5. *Transversality condition:*

$$\lim_{t\to\infty} p^*(t)(q(t) - q^*(t)) \geq 0 \tag{T}$$

for any feasible state trajectory q.

6. *Complementarity condition:*

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

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

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

8. *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.

B.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 7 hold. Note that condition (T) holds trivially in our case because $p^*(t) \geq 0$, $q^*(t) = 0$ for t large enough, and for any feasible state trajectory, $q(t) \geq 0$. In fact, $p^*(t) = 0$ for large enough t in our case, which again ensures condition (T) holds.

We first prove an auxiliary lemma.

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 $t > \tau_1$.*

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 $t\geq\tau_1$. □

Proof of Theorem 1. In this case,

$$\begin{aligned}H(q(t),z(t),p(t),t)=&h_1q_1(t)+h_2q_2(t)+\phi_{12}z_{12}(t)\\&+p_1(t)\left(\lambda_1(t)-\mu_{11}z_{11}(t)-\mu_{12}z_{12}(t)\right)+p_2(t)\left(\lambda_2(t)-\mu_{22}z_{22}(t)\right)\end{aligned}$$

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 7.

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

$$\tau^*=\inf\{t\geq 0: h_1\mu_{12}G_1^t(q_1(t))-\phi_{12}\leq h_2\mu_{22}G_2^t(q_2(t))\} \tag{30}$$

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 $h_1\mu_{12}\geq h_2\mu_{22}$ and $\psi(t)$, 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

$$\tau_1^*=\tau^*+G_1^{\tau^*}(q_1^*(\tau^*)),\quad \tau_2^*=\tau^*+G_2^{\tau^*}(q_2^*(\tau^*)). \tag{31}$$

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 $n\geq 1$, $I_i=[V_{i-1},V_i)$ and $0=V_0<V_1<\cdots<V_n=\tau^*$, as follows. In the interior $t\in(V_{i-1},V_i)$ 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 V_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 $p_1^*(V_n) = p_1^*(\tau^*)$ defined, moving backwards in time, we recursively define $p_1^*(t)$ for $t \in [0, V_n)$. 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,

$$p_1^*(t)\mu_{12} - \phi_{12} - p_2^*(t)\mu_{22} \geq 0; \tag{32}$$

and (iii) in any boundary subinterval I_i,

$$p_1^*(t)\mu_{12} - \phi_{12} - p_2^*(t)\mu_{22} = 0. \tag{33}$$

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

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

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

for $t \in [V_{k-1}, V_k)$. That is, p_1^* is continuous at V_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 $p_1^*(V_{k-1})\mu_{12} - \phi_{12} - p_2^*(V_{k-1})\mu_{22} \geq 0$. If I_k is a boundary subinterval, we set $p_1^*(V_{k-1}) = p_2^*(V_{k-1})\mu_{22}/\mu_{12} + \phi_{12}/\mu_{12}$ and $p_1^*(t) = p_1^*(V_{k-1}) - \frac{h_2\mu_{22}}{\mu_{12}}(t - V_{k-1})$ for $t \in (V_{k-1}, V_k)$. That is, p_1^* has a jump at V_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 V_k is $p_1^*(V_k) - p_2^*(V_k)\mu_{22}/\mu_{12} - \phi_{12}/\mu_{12} \geq 0$, which is non-negative because $p_1^*(V_k)\mu_{12} - \phi_{12} - p_2^*(V_k)\mu_{22} \geq 0$. 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 $p_1^*(t)\mu_{12} - \phi_{12} - p_2^*(t)\mu_{22} \leq 0$ 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 $[V_{k-1}, V_k)$, $\xi_{22}^*(t) = 0$. This holds because of (33). (Note that $z_{22}^*(V_k) = 0$.)

For (AH), 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 $t \geq \tau^*$. 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 $t \geq \tau^*$.

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 $z_{11}^*(t)\mu_{11} + z_{12}^*(t)\mu_{12} \leq \lambda_1(t)$, 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 $p_1^*(t)\mu_{12} - \phi_{12} \geq p_2^*(t)\mu_{22}$, it is optimal to have $z_{12}^*(t)$ being maximal. When $t \geq \tau^*$, $p_1^*(t)\mu_{12} - \phi_{12} \leq p_2^*(t)\mu_{22}$, 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 $t \geq \tau_2^*$, $\phi_{12} - p_1^*(t)\mu_{12} \leq 0$, 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

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

$$p_2^*(t) = \begin{cases} h_2(\tau_2 - t) + h_1 \frac{\mu_{12}}{\mu_{22}} \tau^*, & t \in [0, \tau_2), \\ h_1 \frac{\mu_{12}}{\mu_{22}} (\tau_2 + \tau^* - t), & t \in [\tau_2, \tau_2 + \tau^*), \\ 0, & t \in [\tau_2 + \tau^*, \infty). \end{cases}$$

Lastly, define the multipliers

$$\eta_1^*(t) = \begin{cases} 0, & t \in [0, \tau_1^*), \\ h_1, & t \in [\tau_1^*, \infty), \end{cases}$$

$$\eta_2^*(t) = \begin{cases} 0, & t \in [0, \tau_2), \\ h_2 - h_1 \frac{\mu_{12}}{\mu_{22}}, & t \in [\tau_2, \tau_2 + \tau^*), \\ h_2, & t \in [\tau_2 + \tau^*, \infty); \end{cases}$$

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

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

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

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,

$$\phi_{12} - h_1\mu_{12}(\tau_1^* - t) + h_2\mu_{22}(\tau_2 - t) + h_1\mu_{12}\tau^* \geq 0 \tag{34}$$

and $\xi_{12}^*(t) \geq 0$.

The conditions (ODE), (ADJ), (C), (J), and (H) are verified straightforwardly by construction. For (AH),

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

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 $t \in [\tau_2, \tau_2 + \tau^*)$,

$$\phi_{12} - p_1^*(t)\mu_{12} + p_2^*(t)\mu_{22} = \phi_{12} - h_1\mu_{12}\left((\tau_1^* - t) - (\tau_2 + \tau^* - t)\right) = \phi_{12} - h_1\mu_{12}G_1^{\tau_2+\tau^*}(q_1^*(\tau_2 + \tau^*)) = 0.$$

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 (34)), 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. □

Appendix C: Proof of Asymptotic Optimality

In this section, we provide the proof of Theorem 3. Lemma 2 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 8 that there exists a fluid limit under any admissible control.

2. We then show in Theorem 9 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 9 is the continuity in the G and $\tilde{G}$ factors, which is established in Lemma 3 and Lemma 4, 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 $\tilde{G}_{i,n}^t(nx_i)$ as $\tilde{G}_{i,n}^t(nx_i) = \inf\left\{\Delta \geq 0 : \int_t^{t+\Delta}(s_i\mu_{ii} - \tilde{\lambda}_i^n(s))\,ds = x_i\right\}$. With a little abuse of notation, we redefine the input of the function as x_i, instead of nx_i, i.e.,

$$\tilde{G}_{i,n}^t(x_i) = \inf\left\{\Delta \geq 0 : \int_t^{t+\Delta}(s_i\mu_{ii} - \tilde{\lambda}_i^n(s))\,ds = x_i\right\}. \tag{35}$$

We start from proving the following lemma, which establishes the results in the first part of Theorem 3.

LEMMA 2. *For any admissible control π^n for system n, $\bar{V}^{n,\pi^n}(x) \geq \bar{V}^*(x)$.*

Proof. We suppress the superscript π^n from the corresponding processes to simplify the notation. Let $g_i^n(t,x) = \lambda_i^n(t) - \sum_j(\pi_t^n(x))_{ij}n\mu_{ij}$ for $t \in [0, T(x)], x \in \mathbb{N}_0^2$. Note that

$$M_i(\cdot) := X_i^n(\cdot) - nx_i - \int_0^{\cdot} g_i^n(s, X^n(s))\,ds$$

is a zero-mean martingale by the Dynkin formula. Taking expectation gives

$$\mathbb{E}_\pi[X_i^n(t)] = nx_i + \int_0^t \mathbb{E}_\pi[g_i^n(s, X^n(s))]\,ds = nx_i + \int_0^t \left(n\lambda_i(s) - n\sum_j \mathbb{E}_\pi[Z_{ij}^n(s)]\mu_{ij}\right)ds \tag{36}$$

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, $\sum_i z_{ij}(t) = \sum_i \mathbb{E}[Z^n_{ij}(t)] \le s_j$, and (36) implies that

$$0 \le \mathbb{E}_\pi[X^n_i(t)/n] = x_i + \int_0^t \left(\lambda_i(s) - \sum_j \mathbb{E}_\pi[Z^n_{ij}(s)]\mu_{ij}\right) ds$$

for $t \in [0, T(x)]$. Thus, u is an admissible control for the fluid problem and the corresponding fluid dynamics take 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,

$$\begin{aligned}\bar{V}^{n,\pi^n}(x) &= \mathbb{E}_\pi\left[\int_0^{T(x)} \left(\sum_i \frac{h_i}{n} X^n_i(t) + \sum_{i\ne j} \phi_{ij} Z^n_{ij}(t)\right) dt\right] \\ &= \int_0^{T(x)} \left(\sum_i h_i q_i(t) + \sum_{i\ne j} \phi_{ij} z_{ij}(t)\right) dt \\ &\ge \bar{V}^*(x). \quad \square\end{aligned}$$

To prove the second part of Theorem 3, we first introduce a notion of a fluid limit and show in Theorem 8 below that there exists a fluid limit under any admissible control. Let $\bar{Y} = (\bar{Y}_{11}, \bar{Y}_{12}, \bar{Y}_{22})$, where

$$\bar{Y}^n_{ij}(t) = \int_0^t Z^n_{ij}(s)\, ds$$

is the total amount of time spent by pool j servers on class i customers up to time t.

THEOREM 8. *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) $\bar{X}(0) = x$, $\bar{X}(t) \ge 0$ for $t \ge 0$;

(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) $\sum_i (\bar{Y}_{ij}(t) - \bar{Y}_{ij}(s)) \le s_j(t-s)$ for $j = 1, 2$ and $0 \le s < t$.

Proof. By Strong Law of Large Numbers, the scaled number of arrivals

$$\bar{A}^n_i(t) := \frac{1}{n} S_i\left(\int_0^t \lambda^n_i(s)\, ds\right) = \frac{1}{n} S_i\left(n \int_0^t \lambda_i(s)\, ds\right),$$

where S_i is a rate-1 Poisson process, satisfies

$$\bar{A}^n_i(t) \to \int_0^t \lambda_i(s)\, ds$$

uniformly on compact sets (u.o.c.) as $n \to \infty$. The rest of the proof follows from Theorem 6.5 in Dai and Harrison (2020). $\square$

From Theorem 8, there exists a fluid limit for the sequence of systems under the fluid translated control $\{\tilde{\nu}^n\}_{n\ge1}$ of Theorem 3. We next show that any fluid limit of the sequence of systems under policy $\{\tilde{\nu}^n\}_{n\ge1}$ is equal to the optimal fluid trajectory. Let (q^*, y^*) denote the optimal fluid trajectory, i.e., (q^*, z^*) is as defined in Theorem 1 and $y^*_{ij}(t) = \int_0^t z^*_{ij}(s)\, ds$.

THEOREM 9. *Let* $(\bar{X},\bar{Y}) = (\bar{X}_1,\bar{X}_2,\bar{Y}_{11},\bar{Y}_{12},\bar{Y}_{22})$ *be a fluid limit for the sequence of systems under policy* $\{\tilde{\nu}_n\}_{n\geq 1}$. *Then,* $(\bar{X},\bar{Y}) = (q^*,y^*)$.

Before we prove Theorem 9, we first present two auxiliary lemmas that will be used in the proof.

LEMMA 3. *If* $t \mapsto \bar{x}_i(t)$ *is continuous,* $t \mapsto G_i^t(\bar{x}_i(t))$ *is also continuous.*

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 $0 \leq \delta < \delta_2$, $\xi_2(\delta) > x$. Thus,

$$G_i^{t+\delta}(x) \leq G_i^t(x) + \epsilon.$$

Above all, for $0 \leq \delta < \epsilon \wedge \delta_2$, $|G_i^{t+\delta}(x) - G_i^t(x)| < \epsilon$, i.e., we have right-continuity.

Case II: $t < \kappa_i$. Note that for $0 \leq \delta \leq \kappa_i - t$, $\int_{t+\delta}^{t+G_i^t(x)} a_i(s)\,ds \geq x$. Thus,

$$G_i^{t+\delta}(x) \leq G_i^t(x) - \delta \leq G_i^t(x).$$

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 $0 \leq \delta \leq \delta_3$ $\xi_3(\delta) < x$ Thus,

$$G_i^{t+\delta}(x) \geq G_i^t(x) - \epsilon/2 - \delta.$$

Above all, for $0 \leq \delta \leq \delta_3 \wedge \epsilon/2$, $|G_i^{t+\delta}(x) - G_i^t(x)| = G_i^t(x) - G_i^{t+\delta}(x) \leq \epsilon$, i.e., we have right-continuity.

For the left continuity, we first note that for $0 \leq \delta \leq \kappa_i - t$, $\int_{t-\delta}^{t+G_i^t(x)} a_i(s)\,ds \leq x$. Thus,

$$G_i^{t-\delta}(x) \geq G_i^t(x) + \delta \geq G_i^t(x).$$

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 $0 \leq \delta \leq \delta_4$, $\xi_4(\delta) > x$ Thus,

$$G_i^{t-\delta}(x) \leq G_i^t(x) + \epsilon/2 + \delta.$$

Above all, for $0 \leq \delta \leq \delta_4 \wedge \epsilon/2$, $|G_i^{t-\delta}(x) - G_i^t(x)| = G_i^{t-\delta}(x) - G_i^t(x) \leq \epsilon$, 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 $\tilde{G}^t_{i,n}(x)$ in (35).

LEMMA 4. *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*

$$\tilde{G}^t_{1,n}(X_1^n(t))\to G_1^t(X_1(t))$$

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 of Lemma 4. 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

$$|G_1^t(x_1)-\tilde{G}^t_{1,n}(x_2)|\leq\alpha$$

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

$$\tilde{G}^t_{1,n}(x_2)<G_1^t(x_1)+\alpha.$$

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

$$\tilde{G}^t_{1,n}(x_2) > G^t_1(x_1) - \alpha.$$

(We have assumed above that $t + G^t_1(x_1) - \alpha \geq \kappa_1$, since otherwise it is trivial that $\tilde{G}^t_{1,n}(x_2) > G^t_1(x_1) - \alpha$.) Hence, if $|x_2 - x_1| < D(\alpha)/2$, then

$$|G^t_1(x_1) - \tilde{G}^t_{1,n}(x_2)| \leq \alpha$$

for $n > N_2$, as required. The proof for $[\kappa_1, A]$ and for class 2 are similar. □

Proof of Theorem 9. 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,

$$h_1\mu_{12}G^t_1(q^*_1(t)) - \phi_{12} > h_2\mu_{22}G^t_2(q^*_2(t))$$

for $t \in [0, T_1)$ and

$$h_1\mu_{12}G^t_1(q^*_1(t)) - \phi_{12} < h_2\mu_{22}G^t_2(q^*_2(t))$$

for $t > T_1$.

Supppose $T_1 > 0$, so that $h_1\mu_{12}G^{T_1}_1(q^*_1(T_1)) - \phi_{12} = h_2\mu_{22}G^{T_1}_2(q^*_2(T_1))$. We partition $[0, T_1)$ into finitely many subintervals $[V_i, V_{i+1})$ $(i = 0, \cdots, n)$ such that $0 = V_0 < \cdots < V_{n+1} = T_1$, and on each open subinterval $t \in (V_i, V_{i+1})$, 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 4 and the proof of Theorem 1.

We next show inductively that $(\bar{X}, \bar{Y}) = (q^*, y^*)$ on each $[V_i, V_{i+1})$. Suppose that $\bar{X}(V_i) = q^*(V_i)$ and $\bar{Y}(V_i) = y^*(V_i)$ for some i.

We first consider the case that $q^*_1(t) > 0$ on (V_i, V_{i+1}). Because $q^*_1(t)$ decreases at the maximum possible rate for $t < T_1$ under the optimal fluid control, we have that $\bar{X}_1(t) \geq q^*_1(t) > 0$ and $\bar{X}_2(t) \leq q^*_2(t)$ for all $t \in (V_i, V_{i+1})$. Hence, for $s \in (V_i, V_{i+1})$, $h_1\mu_{12}G^s_1(\bar{X}_1(s)) - \phi_{12} > h_2\mu_{22}G^s_2(\bar{X}_2(s))$. By continuity of $t \mapsto G^t_i(\bar{X}_i(t))$ (Lemma 3), we have

$$h_1\mu_{12}G^t_1(\bar{X}_1(t)) - \phi_{12} > \delta + h_2\mu_{22}G^t_2(\bar{X}_2(t))$$

for all $t \in [s - \epsilon, s + \epsilon]$, for some $\epsilon, \delta > 0$. Since $(X^n_1(t)/n, X^n_2(t)/n)) \to (\bar{X}_1(t), \bar{X}_2(t))$ u.o.c., we have by Lemma 4 that

$$h_1\mu_{12}\tilde{G}^t_{1,n}(X^n_1(t)/n) - \phi_{12} > \delta/2 + h_2\mu_{22}\tilde{G}^t_{2,n}(X^n_2(t)/n) \quad \text{and} \quad X^n_1(t) > s_1 + s_2$$

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 $h_1\mu_{12}G^t_1(q^*_1(t)) - \phi_{12} > h_2\mu_{22}G^t_2(q^*_2(t))$,

$$h_1\mu_{12}G^t_1(\bar{X}_1(t)) - \phi_{12} > h_2\mu_{22}G^t_2(\bar{X}_2(t))$$

for all $t \in (V_i, V_{i+1})$. Then, $d\bar{Y}(t)/dt = dy^*(t)/dt$ for all (regular) $t \in (V_i, V_{i+1})$, which implies that $(\bar{X}, \bar{Y}) = (q^*, y^*)$ on $t \in [V_i, V_{i+1}]$. In particular, $\bar{X}(V_{i+1}) = q^*(V_{i+1})$ and $\bar{Y}(V_{i+1}) = y^*(V_{i+1})$. This technique – applying

Lemma 4 to derive inequalities involving $\tilde{G}^t_{i,n}(X^n_i(t)/n)$ based on inequalities involving $G^t_i(\bar{X}_i(t))$ – is also used in subsequent cases in the proof.

We next consider the case $q^*_1(t) = 0$ on (V_i, V_{i+1}). Suppose that $\bar{X}_1(t) > 0$ for some $t \in (V_i, V_{i+1})$. Let $S = \sup\{s \le t : \bar{X}_1(s) = 0\}$. Note that $\bar{X}_1(S) = 0$ by continuity, and that $S \ge V_i$ because $\bar{X}_1(V_i) = 0$. By definition $\bar{X}_1(s) > 0$ 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 $d\bar{X}_1(t)/dt = \lambda_1(t) - s_1\mu_{11} - s_2\mu_{12} \le 0$. In this case, $\bar{X}_1(s)$ is non-increasing in $(S, t]$. Thus, $\bar{X}_1(s) \ge \bar{X}_1(t) > 0$ for $s \in (S, t]$. This implies that $\bar{X}_1(s)$ is not continuous at $s = S$, a contradiction. This implies that $\bar{X}_1(t) = 0$ for $t \in (V_i, V_{i+1})$. In addition, by Assumption 1, $\bar{X}_2(t) > 0$ on (V_i, V_{i+1}), which implies that $d(\bar{Y}_{12}(t) + \bar{Y}_{22}(t))/dt = s_2$ on (V_i, V_{i+1}). As $d\bar{Y}_{11}(t)/dt = s_1$ and $\bar{X}_1(t) = 0$, we have

$$\frac{d\bar{Y}_{12}(t)}{dt} = \frac{\lambda_1(t) - s_1\mu_{11}}{\mu_{12}} = \frac{dy^*_{12}(t)}{dt} \text{ for } t \in (V_i, V_{i+1}).$$

Thus, $(\bar{X}, \bar{Y}) = (q^*, y^*)$ on $[V_i, V_{i+1}]$.

By induction, $(\bar{X}, \bar{Y}) = (q^*, y^*)$ on $[0, T_1)$.

Lastly, we analyze $(\bar{X}(t), \bar{Y}(t))$ 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., $T_2 = \inf\{t \ge T_1 : q^*_1(t) = 0\}$. Let $S_2 > T_1$ be the first time $\bar{X}_1$ empties, i.e., $S_2 = \inf\{t \ge T_1 : \bar{X}_1(t) = 0\}$. For $T_1 < t < S_2$, $\bar{X}_1(t) > 0$. By the same reasoning as before, we have that there exists $\epsilon > 0$ such that $X^n_1(s) > s_1$ 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 $h_1\mu_{12}G^t_1(\bar{X}_1(t)) - \phi_{12}$ decreases at rate at least $h_1\mu_{12}$, whereas $h_2\mu_{22}G^t_2(\bar{X}_2(t))$ decreases at rate at most $h_2\mu_{22}$. Since $h_1\mu_{12} > h_2\mu_{22}$ and $h_1\mu_{12}G^{T_1}_1(\bar{X}_1(T_1)) - \phi_{12} \le h_2\mu_{22}G^{T_1}_2(\bar{X}_2(T_1))$ (strict inequality is possible if $T_1 = 0$), we have that for all $T_1 < t < S_2$,

$$h_1\mu_{12}G^t_1(\bar{X}_1(t)) - \phi_{12} < h_2\mu_{22}G^t_2(\bar{X}_2(t)).$$

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, $\bar{X}_1(t) = q^*_1(t)$ for $t \in (T_1, S_2)$ and $S_2 = T_2$. For $t > T_2$, $\bar{X}_1(t) = 0$. Hence, $\bar{X}_1(t) = q^*_1(t)$ for all t.

The above also establishes that $\bar{X}_2 = q^*_2$, $\bar{Y}_{12} = y^*_{12}$ and $\bar{Y}_{22} = y^*_{22}$ on (T_1, T_3), where $T_3 = \inf\{t \ge T_1 : q^*_2(t) = 0\}$ is the common emptying time of the class 2 queue for both q^*_2 and $\bar{X}_2$. For $t > T_3$, we again have that

$$h_1\mu_{12}G^t_1(\bar{X}_1(t)) - \phi_{12} < h_2\mu_{22}G^t_2(\bar{X}_2(t)),$$

as shown above if $t < S_2$, and trivially if $t \ge S_2$. Therefore $\bar{Y}_{12} = y^*_{12}$ for $t > T_3$, and pool 2 only serves its own class for large enough systems. Since $\bar{X}_1 = q^*_1$, this also implies that $\bar{Y}_{11} = y^*_{11}$. Finally, for $t > S_2$, since $\bar{X}'_2(t) \le 0$ whenever $\bar{X}_2(t) > 0$, $\bar{X}_2(t) = q^*_2(t) = 0$ 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 $T_1 = \inf\{t \ge \kappa_2 : q^*_2(t) = 0\}$. Similarly, let $S_1 = \inf\{t \ge \kappa_2 : \bar{X}_2(t) = 0\}$. For $t < S_1 \wedge T_1$, we have $\bar{X}_i(t) > 0$. 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, $(\bar{X}, \bar{Y}) = (q^*, y^*)$

on $t \in [0, T_1]$. For $t > T_1$, we can show as in Case I that $\bar{X}_2'(t) \le 0$ if $\bar{X}_2(t) > 0$, so that $\bar{X}_2(t) = 0 = q_2^*(t)$. Hence also $\bar{Y}_{22}(t) = y_{22}^*(t)$ for $t > T_1$.

Next, let $T_2 \ge T_1$ 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 $\bar{X}_2(t) = 0$ 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 $\bar{X}_1(t) \ge q_1^*(t) > 0$ for $t \in [T_1, T_2)$. Hence, $h_1\mu_{12}G_1^t(\bar{X}_1(t)) > \phi_{12}$ 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 $\bar{X}_1(t) = q_1^*(t)$ for $t \in [T_1, T_2)$.

For $t > T_2$, $h_1\mu_{12}G_1^t(\bar{X}_1(t)) < \phi_{12}$. 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 $\bar{X}_{11}(t) = q_{11}^*(t)$. □

With Theorem 9, we are now ready to prove the second part of Theorem 3.

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 $\bar{Y}_{12}^n(T(x)) \le s_2 T(x)$ is bounded. We have

$$\begin{aligned}
\bar{V}^{n,\tilde{\nu}^n}(x) =& \mathbb{E}\left[\int_0^{T(x)} \sum_i h_i \bar{X}_i^n(t)\, dt + \phi_{12}\bar{Y}_{12}^n(T(x))\right] \\
=& \int_0^{T(x)} \sum_i h_i \mathbb{E}[\bar{X}_i^n(t)]\, dt + \phi_{12}\mathbb{E}[\bar{Y}_{12}^n(T(x))] \text{ since } \bar{X}_i^n(t) \ge 0 \\
\to& \int_0^{T(x)} \sum_i h_i q_i^*(t)\, dt + \phi_{12} y_{12}^*(T(x)) \text{ as } n \to \infty \text{ by bounded convergence} \\
=& \bar{V}^*(x). \quad \square
\end{aligned}$$

E-companion for "Optimal Routing under Demand Surges: The Value of Future Arrival Rates"

In this E-companion, we provide results from additional numerical experiments as well as the proofs for Theorems 2, 4, 5, and 6. These proofs require constructing the optimal fluid trajectories q^* and the corresponding adjoint vectors p^* so that we can verify the conditions in Theorem 7. The derivations are similar to those of Theorem 1.

Appendix EC.1: Simple Priority Switching Structure for the N-model

The optimal scheduling policy derived in Theorem 1 can be characterized via a time-and-state-dependent switching curve, which is defined as

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

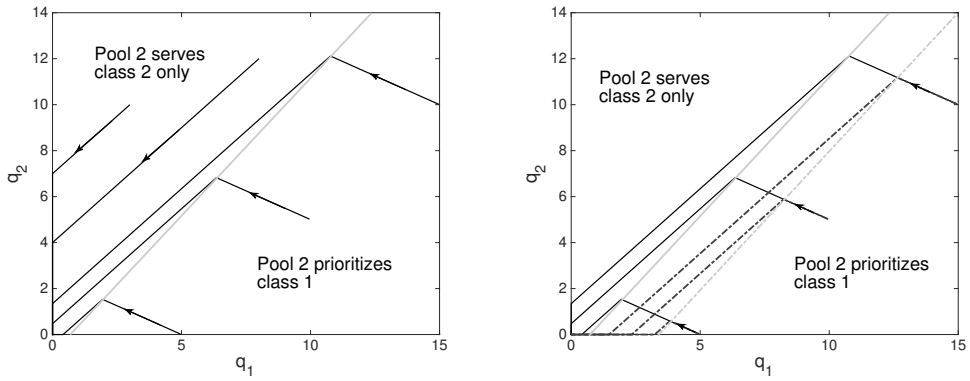
In Case I, when $\psi(t) > 0$, pool 2 gives priority to class 1; otherwise, pool 2 serves class 2 only. Furthermore, Lemma 1 indicates that in Case I, pool 2 switches priority *at most once* throughout the planning horizon. If it switches priority, it is from class 1 to class 2. If it does not switch 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 EC.1 (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 EC.1 (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".

Appendix EC.2: 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



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

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

experiencing a demand surge, the extra flexibility in the X-model is beneficial. In particular, as we explained in Section 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 EC.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 EC.1 Compare the N-model and the X-model under different levels of demand surge for class 1 ($s_1=3$**,** $s_2=4$**,** $\mu_{11}=\mu_{22}=0.25$**,** $\mu_{12}=0.18$**,** $h_1=2$**,** $h_2=1$**,** $\phi_{12}=1$**,** $\lambda_1(t)=\lambda_H\times\mathbf{1}\{0\le t\le 20\}+0.5\times\mathbf{1}\{t>20\}$**,** $\lambda_2(t)=0.6$**,** $q_1(0)=10, q_2(0)=0$**. For the X-model,** $\mu_{21}=0.18$ **and** $\phi_{21}=1$**.)**

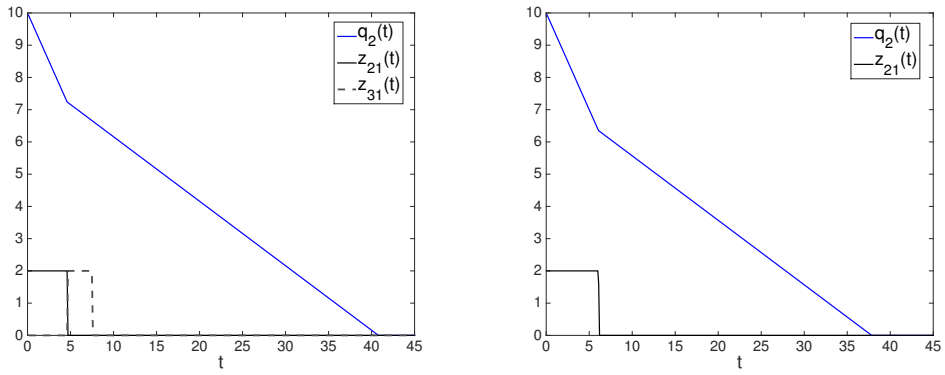
Appendix EC.3: Additional Numerical Experiments

EC.3.1. Fluid trajectory for exN2-model

Figure EC.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 EC.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 EC.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** $q_3(0) = 10$**.)**

EC.3.2. Performance of different policies in the stochastic systems

EC.3.2.1. Tuned versus untuned policies for the N-model Table EC.2 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 (12) and (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%.

EC.3.2.2. Additional comparisons for the N-model Table EC.3 reports the average costs under our policy as well as those under the benchmark policies in the following arrival rate setting:

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

		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 EC.2 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 (12) and (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 $X(0)=(30,40)$.)

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.

The costs for policies in Figure 7 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 as ϕ increases. The cost under modified maximum pressure policy is around 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. This is because too much help during the demand surge 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 Table 1. 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 a 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 EC.3 Expected total cost for the N-model under different scheduling 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.)

EC.3.2.3. Different policies for the X-model For the X-model, we consider the same two settings for the N-model: the baseline setting specified in Section 6.1, as well as the second setting of Section 6.2. In the X-model, pool 1 can also serve class 2 customers if necessary. Table EC.4 reports the cost comparison among different policies. The "Look-ahead (opt)" policy corresponds to the optimal fluid control derived in Section 5.1, while the "Look-ahead (heu)" policy is the heuristic policy – i.e., (12) and (13) – with tuning parameter $\theta = 0.8$. In all cases tested, our heuristic policy achieves performance that is comparable to (slightly worse or slightly better than) that of the fluid-optimal policy.

Table EC.5 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%.

		Look-ahead (opt)	Look-ahead (heu)	MaxPres	ModMaxP	Cmu	ModCmu
		Arrival Rate Setting I					
$\phi=2$	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
		Arrival Rate Setting II					
$\phi=2$	Holding	0.95	1.08	0.93	0.93	0.90	1.28
	Overflow	0.08	0.03	0.20	0.20	0.31	0.00
	Total	1.03	1.11	1.13	1.13	1.22	1.28
	SE	0.002	0.002	0.002	0.002	0.003	0.002
$\phi=10$	Holding	1.02	1.15	0.93	0.92	0.90	1.28
	Overflow	0.28	0.07	1.01	0.98	1.59	0.00
	Total	1.30	1.23	1.95	1.90	2.50	1.28
	SE	0.002	0.002	0.002	0.003	0.004	0.002
$\phi=25$	Holding	1.27	1.28	0.93	0.93	0.90	1.28
	Overflow	0.00	0.00	2.54	2.12	3.98	0.00
	Total	1.28	1.28	3.48	3.05	4.89	1.28
	SE	0.002	0.002	0.006	0.006	0.008	0.002

Table EC.4 Expected total cost for the X-model under different scheduling 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: $h=(1.5,1)$, $s_i=20$, $\mu_{ii}=0.25$, $\mu_{ij}=0.2$, $\phi_{ij}=\phi$ for $i\neq j$ 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 $\lambda_2(t)=3\times\mathbf{1}\{t<40\}+4.5\times\mathbf{1}\{t\geq 40\}$.)

EC.3.2.4. Tuned versus untuned heuristic policies for the 5-by-5 networks Table EC.6 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%.

EC.3.2.5. Clearing a large backlog with constant arrival rates We consider a setting where there is a large initial queue to be cleared, and where arrival rates are constant. Specifically, the initial queue lengths are $X(0)=(400,70)$, while the constant arrival rates are $\lambda_1=3$ and $\lambda_2=3$. As in the baseline setting, we set costs $h_1=1.5$ and $h_2=1$ and service rates $\mu_{11}=\mu_{22}=0.25$ and $\mu_{12}=0.2$. Also $s_1=s_2=20$. The results are shown in Table EC.7.

		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 EC.5 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 (12) and (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 $\phi_{21}=\phi$.)

Note that $G^t_i(q_i(t))=q_i(t)/2$ for $i=1,2$. Thus, when $\phi=2$ is small, our policy is almost the same as the (modified) maximum pressure policy. However, when ϕ is larger than 2, our policy performs better than the two maximum pressure policies. Our policy is better than the $c\mu$ policies throughout.

Appendix EC.4: Proof of the optimal fluid control for the N-model with multiple demand surges

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 7 are satisfied. We

		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 EC.6 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 (12) and (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 $X(0)=(30,40,50,60,70)$.)

first note that (T) holds trivially because $q^*(t)=0$ for t large enough and for any feasible state trajectory, $q(t)\geq 0$.

Case I: $h_1\mu_{12}>h_2\mu_{22}$**.** In this case,

$$\begin{aligned}H(q(t),z(t),p(t),t)=&h_1q_1(t)+h_2q_2(t)+\phi_{12}z_{12}(t)\\&+p_1(t)\left(\lambda_1(t)-\mu_{11}z_{11}(t)-\mu_{12}z_{12}(t)\right)+p_2(t)\left(\lambda_2(t)-\mu_{22}z_{22}(t)\right)\end{aligned}$$

and

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

		Look-ahead	MaxPres	ModMaxP	Cmu	ModCmu
$\phi = 2$	Holding	3.47	3.47	3.47	4.11	5.55
	Overflow	0.16	0.17	0.17	0.25	0.00
	Total	3.63	3.64	3.64	4.36	5.55
	SE	0.004	0.004	0.004	0.005	0.007
$\phi = 10$	Holding	3.52	3.47	3.47	4.11	5.55
	Overflow	0.68	0.84	0.83	1.26	0.00
	Total	4.20	4.31	4.30	5.37	5.55
	SE	0.004	0.004	0.004	0.006	0.007
$\phi = 25$	Holding	3.89	3.47	3.47	4.11	5.55
	Overflow	1.11	2.11	2.01	3.14	0.00
	Total	5.00	5.58	5.48	7.25	5.55
	SE	0.005	0.005	0.005	0.007	0.007

Table EC.7 Expected total cost for the N-model under different scheduling 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: $h = (1.5, 1)$, $\phi_{12} = \phi$, $s_i = 20$, $\lambda_1 = 3$, $\lambda_2 = 3$, $X(0) = (400, 70)$.)

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 $q_i^*, p_i^*, z_{ij}^*, \eta_i^*, \xi_{ij}^*, \gamma_i^*$ 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 7 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

$$\tau^* = \inf\{t \geq 0 : h_1\mu_{12}G_1^t(q_1(t)) - \phi_{12} \leq h_2\mu_{22}G_2^t(q_2(t))\}. \tag{EC.1}$$

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

$$\tau^* = \inf\{t \geq 0 : h_1\mu_{12}G_1^t(q_1(t)) - \phi_{12} \leq h_2\mu_{22}G_2^t(q_2(t))\}. \tag{EC.2}$$

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

$$\tau' = \inf\{t \geq \kappa_b : h_1\mu_{12}G_1^t(q_1(t)) - \phi_{12} \leq h_2\mu_{22}G_2^t(q_2(t))\} \tag{EC.3}$$

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 $n \geq 1$, $I_i = [V_{i-1}, V_i)$ and $0 = V_0 < V_1 < \cdots < V_n = \tau^*$. The subintervals are defined such that in the interior of each subinterval, i.e., $t \in (V_{i-1}, V_i)$, 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 V_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 $p_1^*(V_n) = p_1^*(\tau^*)$ defined, we recursively define $p_1^*(t)$ for $t \in [0, V_n)$. 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,

$$p_1^*(t)\mu_{12} - \phi_{12} - p_2^*(t)\mu_{22} \geq 0; \tag{EC.4}$$

and (iii) in boundary subintervals I_i,

$$p_1^*(t)\mu_{12} - \phi_{12} - p_2^*(t)\mu_{22} = 0. \tag{EC.5}$$

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 $[\kappa_b, \tau_d)$,

$$p_1^*(t)\mu_{12} - \phi_{12} - p_2^*(t)\mu_{22} \geq 0; \tag{EC.6}$$

and (iii) in boundary subintervals I_i of $[\kappa_b, \tau_d)$,

$$p_1^*(t)\mu_{12} - \phi_{12} - p_2^*(t)\mu_{22} = 0. \tag{EC.7}$$

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 $G_1^t(q_1^*(t))$ 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 $t \in [V_{k-1}, V_k)$, $\xi_{22}^*(t) = 0$. This holds because of (EC.5) and (EC.7). We now verify (AH). 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, $q_1^*(\kappa_b) = q_2^*(\kappa_b) = 0$. □

Appendix EC.5: Optimal control for the X-Model

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 7 are satisfied. We first note that (T) holds trivially because $q^*(t) = 0$ for t large enough and for any feasible state trajectory, $q(t) \geq 0$.

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 (11) holds at $t = 0$.

Case I: Pool 2 does not initially prioritize class 1, i.e., (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 $b \le a$ 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 $\tau_2 \le \tau_1$, 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 $\phi_{21} - h_2\mu_{21}G^{\tau_1}_2(q_2(\tau_1)) \geq 0$ 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 (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 (AH). 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

$$\phi_{21} - p^*_2(t)\mu_{21} + p^*_1(t)\mu_{11} = \phi_{21} - h_2\mu_{21}\left((\tau^*_2 - t) - (\tau_1 + \tau^* - t)\right) = \phi_{21} - h_2\mu_{21}G^{\tau_1+\tau^*}_2(q^*_2(\tau_1 + \tau^*)) = 0.$$

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 $t < \tau_1$, $G^t_1(q_1(t)) > 0$, 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 (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., (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 (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 $G^t_1(q^*_1(t)) > 0$.

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 $n \geq 1$, $I_i = [V_{i-1}, V_i)$ and $0 = V_0 < V_1 < \cdots < V_n = T_1$, as follows. In the interior $t \in (V_{i-1}, V_i)$ 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 4 rules out the case $q^*_1(t) = q^*_2(t) = 0$ (such a subinterval cannot occur after $\kappa_1 \vee \kappa_2$, because then $G^t_i(q_i(t)) = 0$ for $i = 1, 2$ and (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 V_k. The fact that there are finitely many such subintervals follows from piecewise monotonicity in Assumption 4, because the class 1 queue length can only leave zero once during each monotone period.

With $p^*_1(V_n) = p^*_1(T_1)$ defined, we recursively define $p^*_1(t)$ for $t \in [0, V_n)$. 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,

$$p^*_1(t)\mu_{12} - \phi_{12} - p^*_2(t)\mu_{22} \geq 0; \tag{EC.8}$$

and (iii) in boundary subintervals I_i,

$$p^*_1(t)\mu_{12} - \phi_{12} - p^*_2(t)\mu_{22} = 0. \tag{EC.9}$$

Note that

$$p^*_1(T_1)\mu_{12} - \phi_{12} - p^*_2(T_1)\mu_{22} = 0. \tag{EC.10}$$

Indeed, this statement is equivalent to equality for (11) at $t = T_1$, which follows from continuity.

Suppose $p^*_1(V_k)$ has been defined for some k, with $p^*_1(V_k)\mu_{12} - \phi_{12} - p^*_2(V_k)\mu_{22} \geq 0$. If I_k is an interior subinterval, we set

$$p^*_1(t) = h_1(V_k - t) + p^*_1(V_k)$$

for $t \in [V_{k-1}, V_k)$. That is, p^*_1 is continuous at V_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 $p^*_1(V_{k-1})\mu_{12} - \phi_{12} - p^*_2(V_{k-1})\mu_{22} \geq 0$.

Suppose instead I_k is a boundary subinterval. We set $p^*_1(V_{k-1}) = p^*_2(V_{k-1})\mu_{22}/\mu_{12} + \phi_{12}/\mu_{12}$ and $p^*_1(t) = p^*_1(V_{k-1}) - \frac{h_2\mu_{22}}{\mu_{12}}(t - V_{k-1})$ for $t \in (V_{k-1}, V_k)$. That is, p^*_1 has a jump at V_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 V_k is $p^*_1(V_k) - p^*_2(V_k)\mu_{22}/\mu_{12} - \phi_{12}/\mu_{12} \geq 0$, which is non-negative because $p^*_1(V_k)\mu_{12} - \phi_{12} - p^*_2(V_k)\mu_{22} \geq 0$. 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

$$\gamma_1^*(t) = \begin{cases} p_1^*(t)\mu_{11}, & t \in [0, T_3), \\ 0, & t \in [T_3, \infty), \end{cases}$$

$$\gamma_2^*(t) = \begin{cases} p_1^*(t)\mu_{12} - \phi_{12}, & t \in [0, T_1) \\ p_2^*(t)\mu_{22}, & t \in [T_1, T_4), \\ 0, & t \in [T_4, \infty). \end{cases}$$

$$\xi_{12}^*(t) = \begin{cases} 0, & t \in [0, T_1), \\ \phi_{12} - p_1^*(t)\mu_{12} + p_2^*(t)\mu_{22}, & t \in [T_1, \infty), \end{cases}$$

$$\xi_{21}^*(t) = \begin{cases} \phi_{21} - p_2^*(t)\mu_{21} + p_1^*(t)\mu_{11}, & t \in [0, T_2), \\ 0, & t \in [T_2, T_3), \\ \phi_{21} - p_2^*(t)\mu_{21}, & t \in [T_3, \infty). \end{cases}$$

$$\xi_{22}^*(t) = \begin{cases} p_1^*(t)\mu_{12} - \phi_{12} - p_2^*(t)\mu_{22}, & t \in [0, T_1), \\ 0, & t \in [T_1, \infty), \end{cases}$$

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 (EC.10). 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 $t \in [V_{k-1}, V_k)$, $\xi_{22}^*(t) = 0$. This holds because of (EC.9). (Note that $z_{22}^*(V_k) = 0$.)

We now verify (AH). 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

$$\tau^* = \inf\left\{ t \geq 0 : G^{\tau_1 + t}_2(q_2(\tau_1 + t)) \leq \frac{\phi_{21}}{h_2\mu_{21}} \right\}.$$

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 (AH). 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

$$\phi_{21} - p^*_2(t)\mu_{21} + p^*_1(t)\mu_{11} = \phi_{21} - h_2\mu_{21}\left((\tau^*_2 - t) - (\tau_1 + \tau^* - t)\right) = \phi_{21} - h_2\mu_{21} G_2^{\tau_1+\tau^*}(q^*_2(\tau_1+\tau^*)) = 0.$$

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 $t \ge \tau_1 + \tau^*$, $\phi_{21} - p^*_2(t)\mu_{21} \ge 0$, 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 $\phi_{12} - p^*_1(t)\mu_{12} + p^*_2(t)\mu_{22} = \xi^*_{12}(t) \ge 0$, 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 $t \ge \tau^*_2$, $\phi_{12} - p^*_1(t)\mu_{12} = \xi^*_{12}(t) \ge 0$, and so it is again Hamiltonian-minimal to have $z^*_{12}(t) = 0$. This completes the proof. $\square$

Appendix EC.6: Optimal control for the exN1-Model

Proof of Theorem 5. We will construct the optimal primal and dual trajectories under the policy characterized in Theorem 5 and show that the conditions in Theorem 7 are satisfied. We first note that (T) holds trivially because $q^*(t) = 0$ for t large enough and for any feasible state trajectory, $q(t) \geq 0$.

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 $h_1\mu_{12} \geq h_2\mu_{22}$, (17) 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 (18) 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 (18) is violated, which is precisely the same condition as in the N-model. If (ii), note that pool 2 stops helping class 1 when (17) 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 $G_1^{t+G^t_3(q)+P^t_3(q)}(\tilde{q}_1(t + G^t_3(q) + P^t_3(q)))$ 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), (AH), (C) and (J) follow directly from the N-model analysis. It remains to verify (M).

The Hamiltonian is

$$H(q(t), z(t), p(t), t) = \sum_i h_i q_i(t) + \sum_{j\neq 1} \phi_{1j} z_{1j}(t) + \sum_i p_i(t) \left(\lambda_i(t) - \sum_j \mu_{ij} z_{ij}(t) \right). \quad \text{(EC.11)}$$

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 $G_i^t(q_i(t)) > 0$ and $\phi_{1i} - p_1^*(t)\mu_{1i} \geq 0$ whenever pool $G_i^t(q_i(t)) = 0$. Moreover, by Assumption 5, $q_1(t) > 0$ 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 (21) is violated. At this time, pool 3 will continue to prioritize class 1 until (22) is violated (if it has not yet stopped prioritizing class 1). Thus, the class 1 queue will take an additional

$$\bar{G}_{\text{exN1,1,3}}^t(q) = G_3^t(q_3) + P_3^t(q) + G_1^{t+G_3^t(q)+P_3^t(q)}(\tilde{q}_1(t + G_3^t(q) + P_3^t(q)))$$

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), (AH), (C), and (J) all follow directly from the N-model analysis. It remains to verify (M).

The Hamiltonian is

$$H(q(t), z(t), p(t), t) = \sum_i h_i q_i(t) + \sum_{j \neq 1} \phi_{1j} z_{1j}(t) + \sum_i p_i(t) \left(\lambda_i(t) - \sum_j \mu_{ij} z_{ij}(t) \right). \quad \text{(EC.12)}$$

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 $G_3^t(q_3(t)) > 0$ and $\phi_{13} - p_1^*(t)\mu_{13} \geq 0$ whenever pool $G_3^t(q_3(t)) = 0$. 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 5, $q_1(t) > 0$ 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), (AH), (C), and (J) all follow directly from the N-model analysis. It remains to verify (M).

The Hamiltonian is

$$H(q(t), z(t), p(t), t) = \sum_i h_i q_i(t) + \sum_{j\neq 1} \phi_{1j} z_{1j}(t) + \sum_i p_i(t) \left(\lambda_i(t) - \sum_j \mu_{ij} z_{ij}(t) \right). \quad \text{(EC.13)}$$

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 5, $q_1(t) > 0$ whenever pool i is providing help to class 1. This establishes (M). $\square$

Appendix EC.7: Optimal control for the exN2-Model

Proof of Theorem 6. We will construct the optimal primal and dual trajectories under the policy characterized in Theorem 6 and show that the conditions in Theorem 7 are satisfied. We first note that (T) holds trivially because $q^*(t) = 0$ for t large enough and for any feasible state trajectory, $q(t) \geq 0$.

Let $q_1(0) = q_1$ and $q_2(0) = q_2$. For the exN2-model, the Hamiltonian takes the form:

$$H(q(t), z(t), p(t), t) = \sum_i h_i q_i(t) + \sum_{j=2}^{3} \phi_{j1} z_{j1}(t) + \sum_i p_i(t) \left(\lambda_i(t) - \sum_j \mu_{ij} z_{ij}(t) \right). \quad \text{(EC.14)}$$

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 (25) 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 (26) 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:

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

$$q_2^*(t) = \begin{cases} q_2 + \int_0^t (\lambda_2(s) - s_2\mu_{22} - z_{21}^*(s)\mu_{21})\,ds, & t \in [0, \tau_1),\\ q_2^*(\tau_1) + \int_{\tau_1}^t (\lambda_2(s) - s_2\mu_{22})\,ds, & t \in [\tau_1, \tau_2^*),\\ 0, & t \in [\tau_2^*, \infty), \end{cases}$$

$$q_3^*(t) = \begin{cases} q_3 + \int_0^t (\lambda_3(s) - s_3\mu_{33} - z_{31}^*(s)\mu_{31})\,ds, & t \in [0, \min\{\tau_3^*, \tau_1+\tau_2\}),\\ q_3^*(\tau_1+\tau_2) + \int_{\tau_1+\tau_2}^t (\lambda_3(s) - s_3\mu_{33})\,ds, & t \in [\tau_1+\tau_2, \tau_3^*),\\ 0, & t \in [\tau_3^*, \infty). \end{cases}$$

Note that by Assumption 6, $q_2^*(t) > 0$ 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 (AH). 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 (27) 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 6 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 (AH). 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 $\tau_1^* = G_1^0(q_1(0)) \geq 0$ 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 (28) 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 6 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 (AH). 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. $\square$