Capacity Allocation and Scheduling in Two-Stage Service Systems with Multi-Class Customers

Zhiheng Zhong[1], Ping Cao[2], Junfei Huang[3] and Sean X. Zhou[3]
[1]Department of Electronic Business, South China University of Technology, Guangzhou, China, zhihengzhong@hnu.edu.cn
[2]School of Management, University of Science and Technology of China, Hefei, China, pcao@ustc.edu.cn
[3]Department of Decision Sciences and Managerial Economics, CUHK Business School, The Chinese University of Hong Kong, Hong Kong SAR, China, {junfeih@cuhk.edu.hk, zhoux@baf.cuhk.edu.hk}

This paper considers a tandem queueing system, in which stage 1 has one station serving multiple classes of arriving customers differing in their service requirements and related delay costs, and stage 2 has multiple parallel stations, each of which provides one type of service. Each station has many statistically identical servers. The objective is to design a joint capacity allocation between the stages/stations and scheduling rule of different classes of customers to minimize the long-run average cost.

Using fluid approximation, we convert the stochastic problem into a fluid optimization problem and develop a procedure to solve it. Based on the solution to the fluid optimization problem, we propose a simple and easy-to-implement capacity allocation and scheduling policy, and establish its asymptotic optimality for the stochastic system. The policy has explicit index-based forms for two special system structures, namely, the many-to-one and one-to-many systems. We further propose a grouping and pooling strategy to streamline the operations of the service system. Finally, we conduct numerical experiments to validate the accuracy of the fluid approximation, and quantify the effect of grouping and pooling based on fluid optimal solution.

Tandem queueing systems are ubiquitous. Our results provide guidelines on the allocation of limited resource and scheduling of customer service in those systems. Our proposed policy would improve the system's operational efficiency as well as customers' service quality.

Key words: tandem queue; fluid analysis; resource allocation; scheduling.

1. Introduction

We study a multi-class service system that consists of two stages connected in tandem. Stage 1 has a single station, serving multiple classes of customers, whereas stage 2 has multiple stations, each providing one type of service. Each station comprises many statistically identical servers. Customers arrive at stage 1 and are first served by one of the servers in stage 1. Upon completing service at stage 1, some customers move to one of the stations at stage 2 for more service, and leave the system after that service is completed. Customers are impatient in the sense that they may leave a queue while waiting before the service there starts.

Service systems with the aforementioned tandem structure are prevalent in practice and have been studied extensively in literature (see, e.g., Harrison 1978). One example is call centers, in which an incoming call interacts with a general service representative for routine inquiries and then

is guided to a specific department if further service is requested, such as investment, mortgage, and credit card services, etc. (see, e.g., Gans et al. 2003, Wang et al. 2019a). Another example is the critical care systems, in which patients are first admitted to the intensive care unit (ICU) for receiving treatment, and then transferred to a step-down unit (SDU) if intermediate level of care (the care that is between the intensive care provided by ICU and general care provided by general wards) is needed. Patients in the SDU generally require less amount of resources (e.g., lower nurse-to-bed ratio requirement). Such patient segmentation ensures the quality of treatment in the critical care system while taking pressure off the ICU by alleviating its congestion level (Armony et al. 2018). Our study of the two-stage system is mainly motivated by the challenge arising from the resource allocation and patient flow control with the presence of SDU in critical care systems.

The manager concerns about the system efficiency, which is mostly affected by two types of cost: holding (customer waiting) costs and abandonment costs. The manager's goal is to minimize the expected long-run average costs. As resource is often limited, the success of managing the aforementioned two-stage system depends on the appropriate allocation of capacity between the two stages. In the healthcare context, the manpower (e.g., nurses) shortage, remains a constraining bottleneck (Green 2010). Hence, given a pre-specified amount of manpower, how to allocate these limited resources between the two stages becomes crucial in alleviating the tension between outstripping customer demand and limited manpower supply, as well as improving the efficiency of systems and quality of care. An effective allocation rule shall take into consideration various customer-side features, such as arriving intensity and delay sensitivity, as well as supply-side parameters, such as heterogeneous operational efficiency and conversion rate from manpower resources to service capacity. Meanwhile, customer flow control is another useful lever to improve operational efficiency. The flow control considered in this paper is in the form of scheduling at the first stage. That is, if there is a server becoming available upon service completion, from which class the next waiting customer is admitted? The objective of this paper is to identify an appropriate allocation scheme among the stations, as well as an effective and easy-to-implement scheduling rule.

Jointly optimizing the capacity allocation and scheduling control for the above stochastic system is technically challenging. For example, if no capacity is allocated to the second stage, then the system becomes a pseudo-single stage system with impatient customers, and one might expect that the classic $c\mu/\theta$-index (Atar et al. 2010) is suitable. Nevertheless, our results reveal that this is not true (see Example 1). For the two-stage service system, even with exponential inter-arrival times, service times and patience times, an exact optimality analysis seems impossible. This is because, even with a given capacity allocation decision, the system state is multi-dimensional due to the multiple customer classes and multiple stations. Identifying the optimal scheduling rule in

a multi-class queueing system is in general difficult due to the large state and action spaces in the underlying Markov decision process, and one can hardly get any managerial insights from the complex exact solution if not impossible (Hu et al. 2022). Moreover, the capacity allocation decision lays another difficulty for deriving the explicit-form solution. Hence, in this paper we resort to asymptotic analysis by using fluid approximation to derive simple yet effective policies and draw managerial insights. Fluid approximations have been proven to be a useful and accurate tool to analyze queueing models with impatient customers (Whitt 2006, Bassamboo and Randhawa 2010), especially when the system is under the efficiency-driven (ED) limiting regime (Garnett et al. 2002, Whitt 2004), that is, when capacity is insufficient to satisfy all service requirements. The main results and contributions of the paper are summarized as follows.

First, we develop a fluid approximation for the tandem queueing system with multiple customer classes. We formulate a fluid optimization problem, and use it to derive a sufficient condition for the asymptotic optimality of capacity allocation and scheduling policies for the stochastic systems.

Second, we solve the fluid optimization problem. Based on its solution, we propose a practical policy for the stochastic system and establish its asymptotic optimality. The policy reduces to explicit index-based ones for two special system structures. These policies provide guidelines for capacity allocation between two stages and prioritization of customer classes in stage 1.

Finally, we propose a fluid-based heuristic for grouping customer classes in stage 1 and pooling stations in stage 2 to form clusters, and numerically demonstrate that such a procedure will not deteriorate system performance.

The rest of the paper is organized as follows. We provide a literature review in Section 2. In Section 3, we set up the model for the general system and formulate a fluid optimization problem, and establish the relationship between the fluid optimization problem and the original stochastic problem. In Section 4, we solve the fluid optimization problem and interpret the optimal solution as a joint resource allocation and scheduling policy for the original system. We also examine two special system structures, namely, the many-to-one and one-to-many systems, and propose index-based scheduling rules. Numerical studies are conducted in Section 5. Section 6 concludes. All proofs are relegated to the appendix.

1.1. Notation

We use the standard notation $\mathbb{R}_+$ to denote the set of nonnegative real numbers, and $\mathbb{R}_+^N$ to denote the N-times product of $\mathbb{R}_+$. For a real number x, we let $\lceil x \rceil$ ($\lfloor x \rfloor$) represent the least (greatest) integer that is not less (not larger) than x. For a pair of numbers a and b, $a \vee b = \max\{a, b\}$,

$a \wedge b = \min\{a, b\}$ and $a^+ = \max\{a, 0\}$. Vectors are assumed to be column vectors. With abuse of notation, we denote by $X := (X_1, X_2)$ as the column vector constructed by concatenating column vectors X_1 and X_2 horizontally. For a set $\mathcal{N}$, we use $|\mathcal{N}|$ to denote the number of its elements. Throughout the paper, we use $\|\cdot\|$ to denote the L^1 norm for any vector.

2. Literature Review

We sketch three streams of research that are related to our work: (i) analysis of tandem queues; (ii) scheduling in queueing systems; and (iii) capacity allocation or staffing in many-server systems. It is not our intention to be exhaustive. For a comprehensive literature review on complex service networks, we refer the readers to the recent work Momčilović et al. (2022), who build a novel framework to model and analyze large and complex service systems.

2.1. Analysis of Tandem Queues

Tandem queueing systems have received a lot of attentions in the early development of queueing networks. Examples in this line of research include Harrison and Shepp (1984) and Rosberg et al. (1982), which study the diffusion approximation and optimal control, respectively, in single-server tandem queues. Some recent research on the optimal controls of single-server tandem queues under different scenarios can be found in Sheu and Ziedins (2010) and Wang et al. (2019b). Using exact analysis, Baumanna and Sandmann (2017), Wang et al. (2019a) formulate the tandem queueing systems with multiple servers as a level-dependent quasi-birth-and-death process, and apply the matrix-analytic method to evaluate system performances. Zychlinski et al. (2018) use time-varying fluid model to assess performance and gain operational insights of tandem queues with finite buffer. Recently, Zychlinski et al. (2020) use tandem-queue models to facilitate geriatric bed allocation so as to minimize the cost incurred due to bed blocking in a hospital-institution system. Different from the aforementioned papers, we consider tandem queueing systems with multiple classes of customers and multiple stations, and consider joint resource allocation and scheduling decisions.

2.2. Scheduling in Queueing Systems

Scheduling for queueing models has been long studied in the literature. For example, Cox and Smith (1961) establish the optimality of the classical $c\mu$ rule. van Mieghem (1995) and Mandelbaum and Stolyar (2004) propose the generalized $c\mu$ rule and established its asymptotic optimality under the conventional heavy traffic framework. For the many-server setting, Gurvich and Whitt (2009) propose the queue-and-idleness ratio rules for parallel service systems, and show its asymptotic optimality in the many-server heavy-traffic regime. Atar et al. (2010) and Atar et al. (2011) propose

the $c\mu/\theta$ rule for overloaded systems with customer abandonment, and establish the asymptotic optimality by investigating a related fluid optimization problem. Kim et al. (2018) study the dynamic scheduling problem with general patience time distribution using diffusion approximation, and Long et al. (2020) study the problem with general delay-related cost via fluid analysis. Recently, Hu et al. (2022) consider a multi-class queueing model that allows customers to transition among classes. Under both the long-run average and transient cost criteria, they solve the corresponding fluid optimal control problems, and interpret the optimal solutions into steady-state and transient scheduling policies for the stochastic system. Zychlinski et al. (2022) derive the optimal scheduling policy in multi-server queues with multiple customer classes in which customers from different classes require different amounts of resources. Our paper serves a non-trivial extension to Atar et al. (2010). Different from Atar et al. (2010) who consider a one-stage many-server setting with a pre-determined number of servers, we consider a two-stage queueing system, in which allocating resource in different stations is an important decision. Our scheduling rule involves the interaction between two stages, which differs from the $c\mu/\theta$ rule in one-stage setting.

2.3. Many-Server Capacity Allocation and Staffing

Capacity planning in queueing systems has received considerable attention in the literature. One important area in this stream is capacity allocation, which concerns how to split a fixed amount of resources among service stations. Chao et al. (2003) consider a multi-site service system in the single-sever setting, and find the optimal allocation policies by solving convex programs under different customer switching scenarios. Kostami and Ward (2009) discuss how to allocate capacity between two queues when customers can choose to wait in a physical line or a virtual line (and return at a specified future time). Best et al. (2015) study a hospital bed allocation problem, and derive optimal care-partitioning policy. To address the time-varying demand, Shone et al. (2019) consider a dynamic resource allocation problem in the single-server setting. They formulate the problem as a dynamic program and propose several heuristic allocation policies via approximate dynamic programming techniques. Another area is staffing in queueing systems, in which capacity is to be determined to trade off service quality against staffing cost. Borst et al. (2004) establish the optimality of the square-root staffing principle. Mandelbaum and Zeltyn (2009) study optimal staffing in many-server queues with abandonment subject to various service-level constraints. Baron and Milner (2009) demonstrate how to staff an outsourced call center that operates in an environment of uncertain arrival rate, with the objective to maximize expected profit subject to service-level agreements. The joint staffing and scheduling problem in many-server setting has also been studied in the literature (see, e.g., Gurvich and Whitt 2010, Armony and Mandelbaum 2011,

Kocaga et al. 2014). Differently, we consider a joint capacity allocation and scheduling problem with a given amount of resource in a two-stage service system.

The most related work to ours is Armony et al. (2018) who consider resource allocation between SDUs and ICUs in hospitals using both fluid and diffusion approximations. Their proposed policy performs fairly well numerically, and is robust with respect to system parameters. Our model is different from theirs in the following prospects: (i) Armony et al. (2018) focus on fulfilling two-stage service requirement for customers from a single class, whereas we consider multiple classes of customers for each stage, thereby scheduling is an essential ingredient of our model; (ii) Armony et al. (2018) assume no waiting for the second-stage customers, so that customers finishing the first-stage service will be blocked if all servers in stage 2 are occupied. In contrast, our model allows customers to wait in stage 2; (iii) Armony et al. (2018) provide the optimal balking threshold and resource allocation in a tandem queue that arises in critical care context, while we propose a joint capacity allocation and scheduling policy in two-stage service systems.

3. General Model: Many-to-Many System

We consider a multi-class two-stage queueing system as depicted in Figure 1. As our work is mainly motivated by the application in the critical-care context, in which critical and semi-critical patients are treated in the first (ICU) and second (SDU) stage respectively, we will append the subscripts "c" and "s" to the quantities in stage 1 and stage 2, respectively.

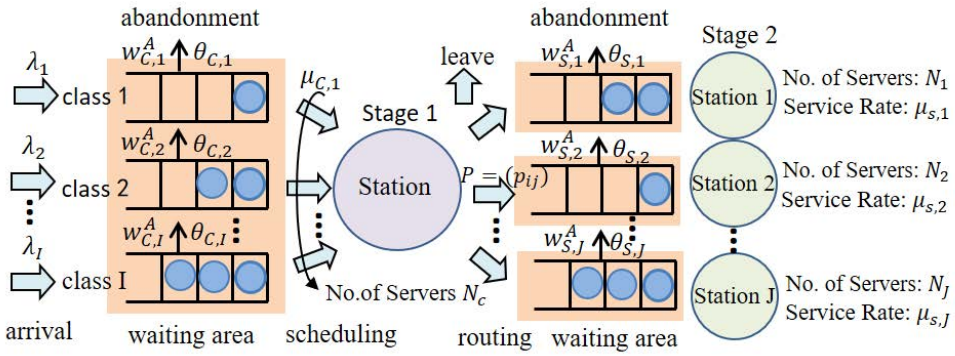


Figure 1 **Many-to-Many System**

3.1. System Dynamics

We describe the model in detail and provide the notation. Stage 1 has a single station with N_c servers. There are I classes of customers who will arrive at stage 1, and each customer class has its

own queue. Denote by $\mathcal{I} := \{1, \ldots, I\}$, the set of customer classes at stage 1. Let $E_c(\cdot) = \{E_{c,i}(t); i \in \mathcal{I}, t \geq 0\}$ denote the arrival process, with $E_{c,i}(t)$ being the number of class-i customers who have arrived by time t. We assume that $E_{c,i}(\cdot) = \{E_{c,i}(t); t \geq 0\}$ is a Poisson process with rate $\lambda_{c,i}$ and

$$E_{c,i}(t) = A_{c,i}(\lambda_{c,i} t), \quad t \geq 0, \tag{1}$$

where $A_{c,i}(\cdot) = \{A_{c,i}(t); t \geq 0\}$ is a unit rate Poisson process. The service times of class-i customers are assumed to be i.i.d., and exponentially distributed with mean $1/\mu_{c,i}$. Introduce $Z_c(\cdot) = \{Z_{c,i}(t); i \in \mathcal{I}, t \geq 0\}$, in which $Z_{c,i}(t)$ is the number of class-i customers in service at stage 1 at time t, satisfying

$$\sum_{i \in \mathcal{I}} Z_{c,i}(t) \leq N_c. \tag{2}$$

Let $D_c(\cdot) = \{D_{c,i}(t); i \in \mathcal{I}, t \geq 0\}$, in which $D_{c,i}(t)$ is the number of class-i customers who have departed stage 1 after completing service by time t. Then

$$D_{c,i}(t) = S_{c,i}\left(\mu_{c,i} \int_0^t Z_{c,i}(s) ds\right), \tag{3}$$

with $S_{c,i}(\cdot) = \{S_{c,i}(t); t \geq 0\}$ being a unit rate Poisson process. Any customer waiting in the queue has a patience time, and will leave the system if the patience gets exhausted. The patience time of a class-i customer is assumed to be exponentially distributed with parameter $\theta_{c,i}$. Denote $Q_c(\cdot) = \{Q_{c,i}(t); i \in \mathcal{I}, t \geq 0\}$, in which $Q_{c,i}(t)$ is the number of class-i customers waiting in queue at stage 1 at time t, and denote $R_c(\cdot) = \{R_{c,i}(t); i \in \mathcal{I}, t \geq 0\}$, in which $R_{c,i}(t)$ is the number of class-i customers that have abandoned stage 1 by time t and can be represented by

$$R_{c,i}(t) = G_{c,i}\left(\theta_{c,i} \int_0^t Q_{c,i}(s) ds\right), \tag{4}$$

with $G_{c,i}(\cdot) = \{G_{c,i}(t); t \geq 0\}$ being a unit rate Poisson process. Finally, let $X_c(\cdot) = \{X_{c,i}(t); i \in \mathcal{I}, t \geq 0\}$, in which $X_{c,i}(t) := Q_{c,i}(t) + Z_{c,i}(t)$ is the total number of class-i customers in stage 1 at time t. Clearly, the dynamics of $X_c(\cdot)$ is

$$X_c(t) = X_c(0) + E_c(t) - D_c(t) - R_c(t). \tag{5}$$

After completing the service in stage 1, some customers will move to stage 2, each with a specific request of service. Stage 2 has J parallel stations with station j having $N_{s,j}$ servers for $j \in \mathcal{J} := \{1, \ldots, J\}$. For $i \in \mathcal{I}$, let $\upsilon_i(\cdot) = \{\upsilon_i(\ell); \ell \in \mathbb{N}\}$ be a sequence of i.i.d. random vectors with $\upsilon_i(\ell) := (\upsilon_{ij}(\ell); j \in \mathcal{J})$, in which $\upsilon_{ij}(\ell) \in \{0, 1\}$ indicates whether the ℓth class-i customer completing service in stage 1 requests a type j service and thus moves to station j in stage 2. We assume $\mathbb{P}\{\upsilon_{ij}(\ell) = 1\} = p_{ij}$, $j \in \mathcal{J}$, and thus $p_{i0} := 1 - \sum_{j \in \mathcal{J}} p_{ij}$ is the probability that a class-i customer

would leave the system after completing service in stage 1. Let $\Upsilon(\cdot) = \{\Upsilon_{ij}(\ell); i \in \mathcal{I}, j \in \mathcal{J}, \ell \in \mathbb{N}\}$ with $\Upsilon_{ij}(\ell) = \sum_{k=1}^{\ell} \upsilon_{ij}(k)$ representing the cumulative number out of the first ℓ class-i customers transferring to station j. Then the cumulative number of class-i customers in stage 1 transferring to station j in stage 2 till time t is given by

$$E_{ij}(t) := \Upsilon_{ij}(D_{c,i}(t)). \tag{6}$$

Let $E_{ij}(\cdot) = \{E_{ij}(t); t \geq 0\}$. Introduce $E_s = \{E_{s,j}(t); j \in \mathcal{J}, t \geq 0\}$, in which $E_{s,j}(t)$ is the cumulative number of arrivals to station j in stage 2. Then we have

$$E_{s,j}(t) = \sum_{i \in \mathcal{I}} E_{ij}(t).$$

We assume that each station in stage 2 is also work-conserving. The service time for each customer in station j is assumed to be exponentially distributed with mean $1/\mu_{s,j}$. Denote by $Z_s(\cdot) = \{Z_{s,j}(t); j \in \mathcal{J}, t \geq 0\}$ with $Z_{s,j}(t)$ representing the number of customers in service at station j at time t. Then, $Z_{s,j}(t) \leq N_{s,j}$ for all $t \geq 0$. Let $D_s(\cdot) = \{D_{s,j}(t); j \in \mathcal{J}, t \geq 0\}$, with $D_{s,j}(t)$ representing the number of customers who have departed station j in stage 2 by time t, given by

$$D_{s,j}(t) = S_{s,j}\left(\mu_{s,j} \int_0^t Z_{s,j}(s)ds\right). \tag{7}$$

Here $S_{s,j}(\cdot) = \{S_{s,j}(t); t \geq 0\}$ is a unit rate Poisson process. Customers waiting in station j may also abandon the system, and the patience times for these customers are assumed to be exponentially distributed with mean $1/\theta_{s,j}$. Denote by $Q_s(\cdot) = \{Q_{s,j}(t); j \in \mathcal{J}, t \geq 0\}$, with $Q_{s,j}(t)$ representing the number of customers waiting in queue at station j at time t. Let $R_s(\cdot) = \{R_{s,j}(t); j \in \mathcal{J}, t \geq 0\}$, in which $R_{s,j}(t)$ is the number of customers having abandoned station j by time t and is given by

$$R_{s,j}(t) = G_{s,j}\left(\theta_{s,j} \int_0^t Q_{s,j}(s)ds\right). \tag{8}$$

Here $G_{s,j}(\cdot) = \{G_{s,j}(t); t \geq 0\}$ is a unit rate Poisson process. Let $X_s(\cdot) = \{X_{s,j}(t); j \in \mathcal{J}, t \geq 0\}$, in which $X_{s,j}(t) := Q_{s,j}(t) + Z_{s,j}(t)$ is the total number of customers in station j at time t. Then due to work-conserving, we have

$$Z_{s,j}(t) = \min\{X_{s,j}(t), N_{s,j}\} \text{ and } \ Q_{s,j}(t) = (X_{s,j}(t) - N_{s,j})^+. \tag{9}$$

The dynamics of $X_s(\cdot)$ is

$$X_s(t) = X_s(0) + E_s(t) - D_s(t) - R_s(t). \tag{10}$$

For notational convenience, we denote by $X = (X_c, X_s)$, $Z = (Z_c, Z_s)$ and $Q = (Q_c, Q_s)$. Then, Equations (5) and (10) can be rewritten as

$$X(t) = X(0) + E(t) - D(t) - R(t) \geq 0, \tag{11}$$

where $D = (D_c, D_s)$, $R = (R_c, R_s)$ and $E = (E_c, E_s)$. We assume that the Poisson processes $A_{c,i}$, $S_{c,i}$, $G_{c,i}$, $S_{s,j}$, $G_{s,j}$, $i \in \mathcal{I}, j \in \mathcal{J}$, are mutually independent, which are also independent of Υ.

3.2. Resource Allocation and Scheduling

The system manager can control the system dynamics via two decisions. The first decision is, given the total number of available resources $m \in \mathbb{R}_+$, to decide the capacity allocation vector $N := (N_c, N_{s,j}; j \in \mathcal{J}) \in \mathbb{R}_+^{J+1}$. (In practice, the number of servers should be integers; however, as we consider fluid approximation in large-scale systems, we allow for non-integer values which can be rounded to the nearest integers.) A capacity allocation vector N is called *feasible* if

$$\frac{N_c}{r_c} + \sum_{j \in \mathcal{J}} \frac{N_{s,j}}{r_{s,j}} \leq m. \tag{12}$$

It is worth pointing out that we distinguish between resource and service capacity: the former is flexible in converting into different types of servers to meet various customers' needs, whereas the latter captures a stations' ability to fulfill customers' service requirements, and is proportional to the number of servers as well as their service rates. We denote by r_c and $r_{s,j}$, with $j \in \mathcal{J}$, the *conversion rates* that measure efficiency in converting the resources into number of servers. That is, if one unit of resource is assigned to stage 1 or station j in stage 2, it can be converted into r_c and $r_{s,j}$ servers, respectively. For example, in the critical care context, m represents the number of available nurses while N_c (resp. $N_{s,j}$) represents the number of beds in the ICU (resp. SDU), and r_c and $r_{s,j}$ are the nurse-to-bed ratios for ICU or SDU j, which specify the numbers of beds that a nurse can simultaneously attend to in these units respectively. These ratios are stipulated by healthcare authority to ensure the quality of care. Then, given a fixed number of nurses m, the capacity allocation vector prescribes the number of beds assigned to ICU and SDUs while satisfying the nurse-to-bed ratios.

The second decision is to set the scheduling rule for stage 1. A scheduling rule ω determines how to allocate available servers to different classes of customers in stage 1. In this paper we only consider preemptive scheduling rules, which are determined by specifying the quantity $Z_c(t) = (Z_{c,i}(t); i \in \mathcal{I})$, the number of servers to different classes of customers, at any time t. Given the system primitives, the process Z_c can determine the system dynamics according to (5)–(9). We call a scheduling rule *admissible* if

1. it is *non-anticipating*, i.e., it depends only on the current and historical system states; and
2. it satisfies (2) as well as the following constraints

$$Z_{c,i}(t) \geq 0, \text{ for } i \in \mathcal{I}, \text{ and } \int_0^\infty \left(N_c - \sum_{i \in \mathcal{I}} Z_{c,i}(t) \right) d \left(\sum_{i \in \mathcal{I}} Q_{c,i}(t) \right) = 0. \tag{13}$$

The first constraint is the non-negative requirement, while the second constraint is usually referred to as the *work-conserving* condition, which specifies that a server at stage 1 is not allowed to be idle if there are customers waiting at stage 1.

We introduce the concept of feasibility for the joint resource allocation and scheduling policy.

DEFINITION 1 (FEASIBLE POLICIES). A policy $\psi := (N, \omega)$ is called *feasible* if

1. the capacity allocation vector N satisfies the constraint (12), and
2. the scheduling rule ω is admissible.

Denote by Ψ the set of all feasible policies.

Finally, we describe the costs incurred in the system under a given policy $\psi \in \Psi$. Let $w^H_{c,i}$ and $w^H_{s,j}$ be the waiting cost rates for a class-i customer at stage 1 and a customer at station j in stage 2, respectively. Then, for any $T > 0$, the expected cumulative waiting cost till time T is

$$\mathbb{E}^{\psi}\left[\int_0^T \left(\sum_{i\in\mathcal{I}} w^H_{c,i} Q_{c,i}(t) + \sum_{j\in\mathcal{J}} w^H_{s,j} Q_{s,j}(t)\right) dt\right],$$

where $\mathbb{E}^{\psi}$ indicates that the expectation is taken with respective to both $Q_{c,i}(t)$ and $Q_{s,j}(t)$ under policy ψ. Denote by $w^A_{c,i}$ and $w^A_{s,j}$ the cost of a class-i customer abandoning stage 1, and cost of a customer abandoning station j in stage 2, respectively. Then, in view of (4) and (8), the expected cumulative abandonment cost till time T is given by

$$\mathbb{E}^{\psi}\left[\sum_{i\in\mathcal{I}} w^A_{c,i} R_{c,i}(T) + \sum_{j\in\mathcal{J}} w^A_{s,j} R_{s,j}(T)\right] = \mathbb{E}^{\psi}\left[\int_0^T \left(\sum_{i\in\mathcal{I}} \theta_{c,i} w^A_{c,i} Q_{c,i}(t) + \sum_{j\in\mathcal{J}} \theta_{s,j} w^A_{s,j} Q_{s,j}(t)\right) dt\right].$$

Let $w_{c,i} := w^H_{c,i} + \theta_{c,i} w^A_{c,i}$ and $w_{s,j} := w^H_{s,j} + \theta_{s,j} w^A_{s,j}$, which can be considered as the effective waiting cost rates for class-i customers at stage 1 and for customers at station j in stage 2, respectively. We assume that $w_{c,i} > 0$ for $i \in \mathcal{I}$ and $w_{s,j} > 0$ for $j \in \mathcal{J}$; otherwise, there is no need to allocate any capacity to such a class or station, hence the class or station can be removed from the system. Then, the total expected cost up to time T under policy $\psi \in \Psi$ is given by

$$J(\psi, T) = \mathbb{E}^{\psi}\left[\int_0^T \left(\sum_{i\in\mathcal{I}} w_{c,i} Q_{c,i}(t) + \sum_{j\in\mathcal{J}} w_{s,j} Q_{s,j}(t)\right) dt\right]. \tag{14}$$

The manager seeks to minimize the system's expected long-run average cost

$$\mathrm{AC}^* := \inf_{\psi\in\Psi} J(\psi). \tag{15}$$

where

$$J(\psi) = \limsup_{T\to\infty} \frac{1}{T} J(\psi, T). \tag{16}$$

Directly solving problem (15) is a challenging task, if not impossible. This is mainly due to the large state and action spaces, the complex decisions (capacity allocation and dynamic scheduling), as well as the complexity of the system structure and the feature of customer abandonment (Atar et al. 2010, 2011, Kim et al. 2018). Therefore, we resort to asymptotic analysis and consider a related fluid approximation problem in the subsequent sections.

3.3. Asymptotic Framework and Fluid Optimization

Consider a sequence of queueing systems with the above-mentioned structures, indexed by m, the total number of resources. The relevant parameters and processes in the mth system will be appended superscript m, except for the service and abandonment rates, routing probabilities and cost parameters, which are assumed unchanged. We assume that for each $i \in \mathcal{I}$, there exists a number $\lambda_{c,i} > 0$ such that

$$\frac{\lambda_{c,i}^m}{m} \to \lambda_{c,i}, \quad \text{as} \quad m \to \infty. \tag{17}$$

We also introduce the fluid-scaled processes $\bar{X}^m(\cdot) = \{\bar{X}^m(t); t \geq 0\}$, $\bar{Z}^m(\cdot) = \{\bar{Z}^m(t); t \geq 0\}$ and $\bar{Q}^m(\cdot) = \{\bar{Q}^m(t); t \geq 0\}$, defined by

$$\bar{X}^m(t) := \frac{1}{m} X^m(t), \quad \bar{Z}^m(t) := \frac{1}{m} Z^m(t), \quad \text{and} \quad \bar{Q}^m(t) := \frac{1}{m} Q^m(t).$$

Under an appropriate sequence of feasible policies $\{\psi^m\}$, we expect that $\bar{X}^m(\cdot) \approx \bar{X}(\cdot)$, $\bar{Z}^m(\cdot) \approx \bar{Z}(\cdot)$, $\bar{Q}^m(\cdot) \approx \bar{Q}(\cdot)$ for some continuous processes $\bar{X}(\cdot)$, $\bar{Z}(\cdot)$, and $\bar{Q}(\cdot)$ when m is large. In addition, we expect that $N^m/m = (N_c^m/m, N_{s,j}^m/m; j \in \mathcal{J})$ can be approximated by $n = (n_c, n_{s,j}; j \in \mathcal{J})$ for sufficiently large m. Then, from (1)–(10), we expect the following relationships to hold:

$$\bar{X}_{c,i}(t) = \bar{X}_{c,i}(0) + \lambda_{c,i} t - \mu_{c,i} \int_0^t \bar{Z}_{c,i}(s) ds - \theta_{c,i} \int_0^t \bar{Q}_{c,i}(s) ds, \tag{18}$$

$$\bar{X}_{s,j}(t) = \bar{X}_{s,j}(0) + \sum_{i \in \mathcal{I}} p_{ij} \mu_{c,i} \int_0^t \bar{Z}_{c,i}(s) ds - \mu_{s,j} \int_0^t \bar{Z}_{s,j}(s) ds - \theta_{s,j} \int_0^t \bar{Q}_{s,j}(s) ds, \tag{19}$$

$$\sum_{i \in \mathcal{I}} \bar{Z}_{c,i}(t) \leq n_c, \tag{20}$$

$$\bar{Q}_{c,i}(t) \geq 0, \quad \bar{Z}_{c,i}(t) \geq 0, \quad \bar{Q}_{c,i}(t) + \bar{Z}_{c,i}(t) = \bar{X}_{c,i}(t), \ i \in \mathcal{I}, \ \text{and} \tag{21}$$

$$\bar{Q}_{s,j}(t) \geq 0, \quad 0 \leq \bar{Z}_{s,j}(t) \leq n_{s,j}, \quad \bar{Q}_{s,j}(t) + \bar{Z}_{s,j}(t) = \bar{X}_{s,j}(t), \ j \in \mathcal{J}. \tag{22}$$

Here, Equation (18) is formally obtained by plugging (3) and (4) into (5) and applying the strong law of large numbers (SLLN). Similarly, Equation (19) is formally obtained by plugging (7) and (8) into (10) and using the SLLN. Relations (20) to (22) are from (2) and the non-negativity of Q^m and Z^m, respectively.

As we focus on the long-run average cost (15), we expect that a *stationary* version of the fluid model (18)–(22) would help the analysis. We assume that there exist vectors $x_c = (x_{c,i}; i \in \mathcal{I})$, $q_c = (q_{c,i}; i \in \mathcal{I})$, and $z_c = (z_{c,i}; i \in \mathcal{I})$ such that $(\bar{X}_c(t), \bar{Z}_c(t), \bar{Q}_{c,i}(t)) = (x_c, z_c, q_c)$ for all $t \geq 0$. Similarly, we assume that there exist vectors $x_s = (x_{s,j}; j \in \mathcal{J})$, $q_s = (q_{s,i}; j \in \mathcal{J})$ and $z_s = (z_{s,i}; j \in \mathcal{J})$ such that $(\bar{X}_s(t), \bar{Z}_s(t), \bar{Q}_s(t)) = (x_s, z_s, q_s)$ for all $t \geq 0$. For notational convenience, we denote $x = (x_c, x_s)$,

$z=(z_c,z_s)$ and $q=(q_c,q_s)$, which can be used to approximate the long-run average fluid contents for customers in the system, being served and waiting, respectively, for large m. Note that because x can be derived from z and q, we can omit it to simplify the presentation. From (12) and (18)–(22), we expect that the vector (z,q,n) satisfies

$$\mu_{c,i}z_{c,i}+\theta_{c,i}q_{c,i}=\lambda_{c,i},\ i\in\mathcal{I}, \tag{23}$$

$$\mu_{s,j}z_{s,j}+\theta_{s,j}q_{s,j}-\sum_{i\in\mathcal{I}}\mu_{c,i}z_{c,i}p_{ij}=0,\ j\in\mathcal{J}, \tag{24}$$

$$\frac{n_c}{r_c}+\sum_{j\in\mathcal{J}}\frac{n_{s,j}}{r_{s,j}}\le 1, \tag{25}$$

$$\sum_{i\in\mathcal{I}}z_{c,i}\le n_c, \tag{26}$$

$$q_{c,i},z_{c,i}\ge 0,\ i\in\mathcal{I},\ \text{and} \tag{27}$$

$$q_{s,j}\ge 0,\quad 0\le z_{s,j}\le n_{s,j},\ j\in\mathcal{J}. \tag{28}$$

We will refer to (23)–(28) a stationary fluid model, and any vector (z,q,n) satisfying (23)–(28) a feasible solution. Denote by $\Xi(\lambda_c)=\{(z,q,n):(z,q,n)\text{ is a feasible solution for a given }\lambda_c\}$. From (14), we define the cost of a stationary fluid model as $\sum_{i\in\mathcal{I}}w_{c,i}q_{c,i}+\sum_{j\in\mathcal{J}}w_{s,j}q_{s,j}$. Therefore, we get a stationary fluid optimization problem which is formulated as a linear program (LP):

$$\bar{J}^*:=\min_{(z,q,n)\in\Xi(\lambda_c)}\ \sum_{i\in\mathcal{I}}w_{c,i}q_{c,i}+\sum_{j\in\mathcal{J}}w_{s,j}q_{s,j}. \tag{29}$$

Before solving (29), we present the following lemma, which identifies a sufficient and necessary condition under which the optimal value of problem (29) is zero. That is, the system's long-run average cost vanishes under fluid scale.

LEMMA 1. *We have* $\bar{J}^*=0$ *if and only if* $\sum_{i\in\mathcal{I}}\lambda_{c,i}\left(\frac{1}{r_c\mu_{c,i}}+\sum_{j\in\mathcal{J}}\frac{p_{ij}}{r_{s,j}\mu_{s,j}}\right)\le 1$.

The proof of Lemma 1 can be found in Appendix EC.1.1. To focus on the more interesting case with non-zero average cost, we impose the following condition in the remaining of this paper:

$$\sum_{i\in\mathcal{I}}\lambda_{c,i}\left(\frac{1}{r_c\mu_{c,i}}+\sum_{j\in\mathcal{J}}\frac{p_{ij}}{r_{s,j}\mu_{s,j}}\right)>1. \tag{30}$$

Condition (30) means that the system is overloaded, which is satisfied in various service systems. For example, in healthcare, service requests from patients often outstrip the supply of manpower resources (see, e.g., Armony et al. 2018) due to stringent training requirements. In addition, modern call centers typically avoid over-staffing for cost-saving purpose, and occasionally face highly variable arrival patterns, thereby operating in an overloaded regime (Gans et al. 2003, Whitt 2004).

We have the following result, which establishes a close relationship between the original stochastic problem (15) and the fluid optimization problem (29). In what follows, we denote by (z^*, q^*, n^*) the optimal solution to problem (29), and let $\bar{J}^* = \sum_{i\in\mathcal{I}} w_{c,i} q^*_{c,i} + \sum_{j\in\mathcal{J}} w_{s,j} q^*_{s,j}$. Moreover, we assume that $\mathbb{E}[\|X^m(0)\|^2] < \infty$ for any m, to simplify the technical analysis.

THEOREM 1. (i) *For any sequence of feasible policies* $\{\psi^m; m \in \mathbb{N}\}$, *we have* $\liminf_{m\to\infty} J^m(\psi^m)/m \geq \bar{J}^*$. *As a result,*

$$\liminf_{m\to\infty} \frac{1}{m} AC^{m,*} \geq \bar{J}^*. \tag{31}$$

(ii) *If a sequence of feasible policies* $\{\psi^{m,*}; m \in \mathbb{N}\}$ *satisfies*

$$\limsup_{t\to\infty} \mathbb{E}\left[\|m^{-1} Q^{m,*}(t) - q^*\|\right] \leq \delta_m \tag{32}$$

for some sequence of positive constants $\{\delta_m\}$ *such that* $\lim_{m\to\infty} \delta_m = 0$, *then we have* $\lim_{m\to\infty} J^m(\psi^{m,*})/m = \bar{J}^*$. *In the above,* $Q^{m,*}(t)$ *is the queue length vector at time* t *under* $\psi^{m,*}$ *in the* m*th system.*

The proof of Theorem 1 can be found in Appendix EC.1.2. We have the following corollary, whose proof is omitted.

COROLLARY 1. *If a sequence of feasible policies* $\{\psi^{m,*}; m \in \mathbb{N}\}$ *satisfies* (32), *then it is asymptotically optimal in the sense that*

$$\limsup_{m\to\infty} \frac{1}{m} \left(J^m(\psi^{m*}) - AC^{m,*}\right) = 0.$$

We close this section with a simplification of problem (29). By writing q in terms of z by (23) and (24) and noting that $n_c = \sum_{i\in\mathcal{I}} z_{c,i}$ and $n_{s,j} = z_{s,j}$ at optimality (otherwise, we can always reduce n_c or $n_{s,j}$ to bind the corresponding constraints without changing the objective function in (29) or destroying the feasibility of the solution), we convert problem (29) into the following one:

$$\begin{aligned}
\max_{z} \quad & \sum_{i\in\mathcal{I}} \left(\frac{w_{c,i}}{\theta_{c,i}} - \sum_{j\in\mathcal{J}} \frac{w_{s,j}}{\theta_{s,j}} p_{ij} \right) \mu_{c,i} z_{c,i} + \sum_{j\in\mathcal{J}} \frac{w_{s,j}}{\theta_{s,j}} \mu_{s,j} z_{s,j} \\
\text{s.t.} \quad & \sum_{i\in\mathcal{I}} \frac{\mu_{c,i} z_{c,i}}{r_c \mu_{c,i}} + \sum_{j\in\mathcal{J}} \frac{\mu_{s,j} z_{s,j}}{r_{s,j} \mu_{s,j}} \leq 1, \\
& 0 \leq \mu_{c,i} z_{c,i} \leq \lambda_{c,i}, \ i \in \mathcal{I}, \\
& 0 \leq \mu_{s,j} z_{s,j} \leq \sum_{i\in\mathcal{I}} \mu_{c,i} z_{c,i} p_{ij}, \ j \in \mathcal{J}.
\end{aligned} \tag{33}$$

Hence, the coefficients $\frac{w_{c,i}}{\theta_{c,i}} - \sum_{j\in\mathcal{J}} \frac{w_{s,j}}{\theta_{s,j}} p_{ij}$ and $\frac{w_{s,j}}{\theta_{s,j}}$ in the objective function can be regarded as the marginal cost reduction when assigning additional unit of capacity to class i in stage 1 and station

j in stage 2, respectively. Note that instead of minimizing total costs as stated in problem (29), we maximize the total cost reduction in problem (33).

Our work extends the single-station model considered in Atar et al. (2010) to a two-stage setting. Several challenges arise. First, instead of only concerning the scheduling rule, we consider a joint capacity allocation and scheduling policy in the tandem structure. Second, because the source of arrivals at the second stage is the departing customers from the first stage, the upper bound of capacity allocation at the second stage depends on both the total service capacity and the associated first-stage throughput. This is different from the one in Atar et al. (2010) in that the upper bound of capacity allocation for each class in their model only depends on its own arrival and service capacity, irrespective of those of other classes. Finally, unlike the nonnegative objective coefficient in Atar et al. (2010), the coefficients of the first term in the objective function of (33) might be negative, which means assigning capacity to the first stage does not always contribute to cost reduction of the overall system. This complicates the analysis of the LP.

Given Corollary 1, in the following sections, we will solve problem (33) and explore the structure of the optimal solution to design a sequence of feasible policies satisfying (32).

4. Fluid-Based Scheduling and Resource Allocation Heuristics

We solve problem (33) in Section 4.1. Though problem (33) can be solved efficiently using generic algorithms such as the simplex method, we provide an alternative solution procedure, which shows how the optimal solution is determined by the system parameters. Based on the fluid optimal solutions, we construct policies for the stochastic systems. We then consider two special model structures to derive simple index-based scheduling rules for the stochastic systems in Section 4.2.

4.1. Optimal Solution to the Fluid Problem

We introduce $\varphi = (\varphi_c, \varphi_s)$ with $\varphi_c = (\varphi_{c,i}; i \in \mathcal{I}) := (\mu_{c,i} z_{c,i}; i \in \mathcal{I})$ and $\varphi_s = (\varphi_{s,j}; j \in \mathcal{J}) := (\mu_{s,j} z_{s,j}; j \in \mathcal{J})$, where $\varphi_{c,i}$ and $\varphi_{s,j}$ can be interpreted as the stationary fluid service capacities occupied by class i in stage 1 and station j in stage 2, respectively. Then, problem (33) becomes

$$
\begin{aligned}
\max_{\varphi} \quad & \sum_{i\in\mathcal{I}} \left(\frac{w_{c,i}}{\theta_{c,i}} - \sum_{j\in\mathcal{J}} \frac{w_{s,j}}{\theta_{s,j}} p_{ij} \right) \varphi_{c,i} + \sum_{j\in\mathcal{J}} \frac{w_{s,j}}{\theta_{s,j}} \varphi_{s,j} \\
\text{s.t.} \quad & \sum_{i\in\mathcal{I}} \frac{\varphi_{c,i}}{r_c \mu_{c,i}} + \sum_{j\in\mathcal{J}} \frac{\varphi_{s,j}}{r_{s,j}\mu_{s,j}} \le 1, \\
& 0 \le \varphi_{c,i} \le \lambda_{c,i}, \ i \in \mathcal{I}, \\
& 0 \le \varphi_{s,j} \le \sum_{i\in\mathcal{I}} \varphi_{c,i} p_{ij}, \ j \in \mathcal{J}.
\end{aligned}
\tag{34}
$$

We solve problem (34) in two steps, i.e.,

$$\max_{\varphi_c \in \Pi_c} \left\{ \sum_{i\in\mathcal{I}} \left(\frac{w_{c,i}}{\theta_{c,i}} - \sum_{j\in\mathcal{J}} \frac{w_{s,j}}{\theta_{s,j}} p_{ij} \right) \varphi_{c,i} + \max_{\varphi_s \in \Pi_s(\varphi_c)} \sum_{j\in\mathcal{J}} \frac{w_{s,j}}{\theta_{s,j}} \varphi_{s,j} \right\} \tag{35}$$

where

$$\Pi_c := \left\{ \varphi_c \in \mathbb{R}^I : 0 \le \varphi_{c,i} \le \lambda_{c,i}, \sum_{i\in\mathcal{I}} \frac{\varphi_{c,i}}{r_c \mu_{c,i}} \le 1 \right\}.$$

and

$$\Pi_s(\varphi_c) := \left\{ \varphi_s \in \mathbb{R}^J : \sum_{j\in\mathcal{J}} \frac{\varphi_{s,j}}{r_{s,j}\mu_{s,j}} \le 1 - \sum_{i\in\mathcal{I}} \frac{\varphi_{c,i}}{r_c \mu_{c,i}},\ 0 \le \varphi_{s,j} \le \sum_{i\in\mathcal{I}} \varphi_{c,i} p_{ij},\ j \in \mathcal{J} \right\}.$$

Denote by $\varphi_s^*(\varphi_c)$ the optimal solution to the inner optimization of (35), which is a bin packing problem. In the sequel, we assume that the indices in $\mathcal{J}$ have been rearranged such that

$$\frac{w_{s,1} r_{s,1} \mu_{s,1}}{\theta_{s,1}} \ge \frac{w_{s,2} r_{s,2} \mu_{s,2}}{\theta_{s,2}} \ge \cdots \ge \frac{w_{s,J} r_{s,J} \mu_{s,J}}{\theta_{s,J}}. \tag{36}$$

To facilitate the presentation, for $\phi \in [\mathcal{J}] := \mathcal{J} \cup \{0\}$, we define

$$\chi_{i,\phi} := \frac{1}{r_c \mu_{c,i}} + \sum_{j=1}^{\phi} \frac{p_{ij}}{r_{s,j}\mu_{s,j}}, \tag{37}$$

where $\sum_1^0 = 0$. Here $\chi_{i,\phi}$ represents the amount of resources required for serving one class-i customer in stage 1, as well as its (expected) subsequent stage-2 service requirements up to station ϕ. One can verify that $\chi_{i,\phi}$ is nondecreasing in ϕ; that is, $\chi_{i,\phi} \le \chi_{i,\phi+1}$ for $\phi < J$.

For any $\varphi_c \in \Pi$, define

$$\underline{j}(\varphi_c) = \max \left\{ j' \in \mathcal{J} : \sum_{i\in\mathcal{I}} \chi_{i,j'} \varphi_{c,i} \le 1 \right\}. \tag{38}$$

If the above set is empty, let $\underline{j}(\varphi_c) = 0$. The following proposition, whose proof is omitted, characterizes $\varphi_s^*(\varphi_c)$ using the standard solution to the bin packing problem.

LEMMA 2. *For any $\varphi_c \in \Pi$, we have*

(i) *if $\underline{j}(\varphi_c) = J$, then $\varphi_{s,j}^*(\varphi_c) = \sum_{i\in\mathcal{I}} \varphi_{c,i} p_{ij}$ for all $j \in \mathcal{J}$;*

(ii) *otherwise,*

$$\varphi_{s,j}^*(\varphi_c) = \begin{cases} \sum_{i\in\mathcal{I}} \varphi_{c,i} p_{ij}, & j < \underline{j}(\varphi_c) + 1, \\ \left(1 - \sum_{i\in\mathcal{I}} \chi_{i,\underline{j}(\varphi_c)} \varphi_{c,i}\right) r_{s,j}\mu_{s,j}, & j = \underline{j}(\varphi_c) + 1, \\ 0, & j > \underline{j}(\varphi_c) + 1. \end{cases}$$

From Lemma 2, it is optimal to fully accommodate the service requirements of first $\underline{j}(\varphi_c)$ stations in stage 2 given the stage-1 capacity assignment φ_c. If $\underline{j}(\varphi_c)=J$, then it is clear that all $\varphi_{s,j}$ can attain the upper bound $\sum_{i\in\mathcal{I}}\varphi_{c,i}p_{ij}$ at optimum. If $\underline{j}(\varphi_c)<J$, then the first $\underline{j}(\varphi_c)$ stations will get capacity $\sum_{i\in\mathcal{I}}\varphi_{c,i}p_{ij}$ while station $\underline{j}(\varphi_c)+1$ receives the remaining resources.

In the second step, we solve the optimization problem

$$\max_{\varphi_c\in\Pi_c}\quad \sum_{i\in\mathcal{I}}\left(\frac{w_{c,i}}{\theta_{c,i}}-\sum_{j\in\mathcal{J}}\frac{w_{s,j}}{\theta_{s,j}}p_{ij}\right)\varphi_{c,i}+\sum_{j\in\mathcal{J}}\frac{w_{s,j}}{\theta_{s,j}}\varphi^*_{s,j}(\varphi_c). \tag{39}$$

Denote by φ_c^* the optimal solution to (39). By comparing the objective functions of (35) and (39) with that of (34), one can verify that $(\varphi_c^*,\varphi_s^*(\varphi_c^*))$ is an optimal solution to problem (34).

In the following, we provide the idea of solving problem (39), and will then summarize the solution to (34) in Theorem 2. Noting that the expression of $\varphi_s^*(\varphi_c)$ in problem (39) depends on $\underline{j}(\varphi_c)\in[\mathcal{J}]$, we separate the feasible region Π_c into $J+1$ disjoint sub-regions: $\Pi_c=\cup_{\phi=0}^{J}\Pi_\phi$ with $\Pi_\phi:=\{\varphi_c\in\Pi_c:\underline{j}(\varphi_c)=\phi\}$. Then we get $J+1$ sub-problems

$$\max_{\varphi_c\in\Pi_\phi}\quad \sum_{i\in\mathcal{I}}\left(\frac{w_{c,i}}{\theta_{c,i}}-\sum_{j\in\mathcal{J}}\frac{w_{s,j}}{\theta_{s,j}}p_{ij}\right)\varphi_{c,i}+\sum_{j\in\mathcal{J}}\frac{w_{s,j}}{\theta_{s,j}}\varphi^*_{s,j}(\varphi_c), \tag{40}$$

for $\phi\in[\mathcal{J}]$. The optimal solution will be the one that attains the maximum of these $J+1$ sub-problems. Then, by substituting the expression of $\varphi^*_{s,j}(\varphi_c)$ in Lemma 2 and noticing that $\underline{j}(\varphi_c)=\phi$ for $\varphi_c\in\Pi_\phi$, one can verify that problem (40) becomes

$$\max_{\varphi_c\in\Pi_\phi}\sum_{i\in\mathcal{I}}\xi_{i,\phi}\varphi_{c,i}+\mathbb{1}\{\phi<J\}\cdot\left(\frac{w_{s,\phi+1}r_{s,\phi+1}\mu_{s,\phi+1}}{\theta_{s,\phi+1}}\right), \tag{41}$$

where

$$\xi_{i,\phi}:=\begin{cases}\frac{w_{c,i}}{\theta_{c,i}}-\sum_{j=\phi+1}^{J}\frac{w_{s,j}}{\theta_{s,j}}p_{ij}-\frac{w_{s,\phi+1}r_{s,\phi+1}\mu_{s,\phi+1}}{\theta_{s,\phi+1}}\chi_{i,\phi}, & \phi<J,\\ \frac{w_{c,i}}{\theta_{c,i}}, & \phi=J.\end{cases} \tag{42}$$

We will provide the physical meaning for $\xi_{i,\phi}$ in the one-to-many setting (after Proposition 3).

Note that for $\phi<J$, $\underline{j}(\varphi_c)=\phi$ is equivalent to

$$\sum_{i\in\mathcal{I}}\chi_{i,\phi}\varphi_{c,i}\le 1,\quad\text{and}\quad\sum_{i\in\mathcal{I}}\chi_{i,\phi+1}\varphi_{c,i}>1.$$

In addition, Π_c is a closed set and $\chi_{i,\phi}\ge 1/(r_c\mu_{c,i})$. Hence the closure of Π_ϕ, denoted by $\bar{\Pi}_\phi$, is still a subset of Π_c and can be represented as

$$\bar{\Pi}_\phi=\left\{\varphi_c\in\mathbb{R}^I:\sum_{i\in\mathcal{I}}\chi_{i,\phi}\varphi_{c,i}\le 1,\ \sum_{i\in\mathcal{I}}\chi_{i,\phi+1}\varphi_{c,i}\ge 1,\ 0\le\varphi_{c,i}\le\lambda_{c,i},\ i\in\mathcal{I}\right\}.$$

For $\phi = J$, using $\chi_{i,J} \geq 1/(r_c \mu_{c,i})$ for each i, we have

$$\bar{\Pi}_J = \Pi_J = \left\{ \varphi_c \in \mathbb{R}^I : \sum_{i \in \mathcal{I}} \chi_{i,J} \varphi_{c,i} \leq 1,\ 0 \leq \varphi_{c,i} \leq \lambda_{c,i},\ i \in \mathcal{I} \right\}.$$

We relax the constraint in (41) from Π_ϕ to $\bar{\Pi}_\phi$; that is, for $\phi \in [\mathcal{J}]$, we consider

$$\max_{\varphi_c \in \bar{\Pi}_\phi} \sum_{i \in \mathcal{I}} \xi_{i,\phi} \varphi_{c,i} + \mathbb{1}\{\phi < J\} \cdot \left(\frac{w_{s,\phi+1} r_{s,\phi+1} \mu_{s,\phi+1}}{\theta_{s,\phi+1}} \right). \tag{43}$$

We call (43) the sub-problem ϕ, and denote by $\varphi_c^*(\phi)$ and $\pi^*(\phi)$ the optimal solution and optimal value to sub-problem ϕ, respectively.

We have the following observation, whose proof is omitted:

LEMMA 3. *Denote by V^* the optimal value of problem (39). Then, there exists one ϕ^* such that $\varphi_c^* \in \bar{\Pi}_{\phi^*}$, and $V^* = \pi^*(\phi^*)$; that is, V^* is the optimal value of the sub-problem ϕ^*.*

To find the optimal ϕ^* (which hints up to which station in stage 2 the service requirement can be fully satisfied under the optimal policy), for any $\phi \in [\mathcal{J}]$, we separate $\mathcal{I}$ into two disjoint subsets: $\mathcal{I}_+(\phi) = \{i \in \mathcal{I} : \xi_{i,\phi} \geq 0\}$ and $\mathcal{I}_-(\phi) = \{i \in \mathcal{I} : \xi_{i,\phi} < 0\}$, and let $i_+(\phi) = |\mathcal{I}_+(\phi)|$. Then, we have the following monotonicity results, whose proof can be found in Appendix EC.2.1.

LEMMA 4. *Fix $i \in \mathcal{I}$. Then*

(i) *$\xi_{i,\phi}$ is nondecreasing in ϕ; that is, $\xi_{i,\phi} \leq \xi_{i,\phi+1}$ for $\phi < J$;*
(ii) *$\mathcal{I}_+(\phi)$ is nondecreasing in ϕ; that is, $\mathcal{I}_+(\phi) \subseteq \mathcal{I}_+(\phi+1)$. Hence $i_+(\phi)$ is nondecreasing in ϕ.*

The solutions to (43) are summarized in Proposition EC.1 in Appendix EC.2. The form of the solution to sub-problem ϕ depends on whether ϕ is in one of the following sets:

$$\begin{aligned}
\mathcal{S}_0 &= \left\{ \phi \in [\mathcal{J}] : \sum_{i \in \mathcal{I}} \lambda_{c,i} \chi_{i,\phi+1} < 1 \right\}, \\
\mathcal{S}_1 &= \mathcal{S}_0^c \cap \left\{ \phi \in [\mathcal{J}] : \sum_{i \in \mathcal{I}_+(\phi)} \lambda_{c,i} \chi_{i,\phi+1} < 1 \right\}, \\
\mathcal{S}_2 &= \mathcal{S}_0^c \cap \left\{ \phi \in [\mathcal{J}] : \sum_{i \in \mathcal{I}_+(\phi)} \lambda_{c,i} \chi_{i,\phi} \leq 1 \leq \sum_{i \in \mathcal{I}_+(\phi)} \lambda_{c,i} \chi_{i,\phi+1} \right\}, \\
\mathcal{S}_3 &= \mathcal{S}_0^c \cap \left\{ \phi \in [\mathcal{J}] : \sum_{i \in \mathcal{I}_+(\phi)} \lambda_{c,i} \chi_{i,\phi} > 1 \right\},
\end{aligned} \tag{44}$$

where $\mathcal{S}_0^c = [\mathcal{J}] \setminus \mathcal{S}_0$ is the complement of $\mathcal{S}_0$. Specifically, if $\phi \in \mathcal{S}_0$, then sub-problem ϕ contains no feasible solutions; if $\phi \in \mathcal{S}_1$ (resp., $\phi \in \mathcal{S}_3$), then the second (resp., first) constraint of sub-problem

ϕ is binding at optimality; if $\phi \in \mathcal{S}_2$, both the first and second constraints of sub-problem ϕ can be non-binding at optimality. Moreover, $J \in \mathcal{S}_3$ because $\mathcal{I}_+(J) = \mathcal{I}$ and condition (30) is imposed.

Using Lemma 4, we have the following proposition that states several monotonicity properties of $\pi^*(\phi)$, which will be useful to characterize the optimal solution to problem (34).

PROPOSITION 1. (i) *We have* $\phi_0 < \phi_1 < \phi_2 < \phi_3$ *for* $\phi_\ell \in \mathcal{S}_\ell$, *with* $\ell = 0, 1, 2, 3$.

(ii) *If* $\phi_1 \in \mathcal{S}_1$, *then* $\pi^*(\phi_1) \le \pi^*(\phi_1 + 1)$.

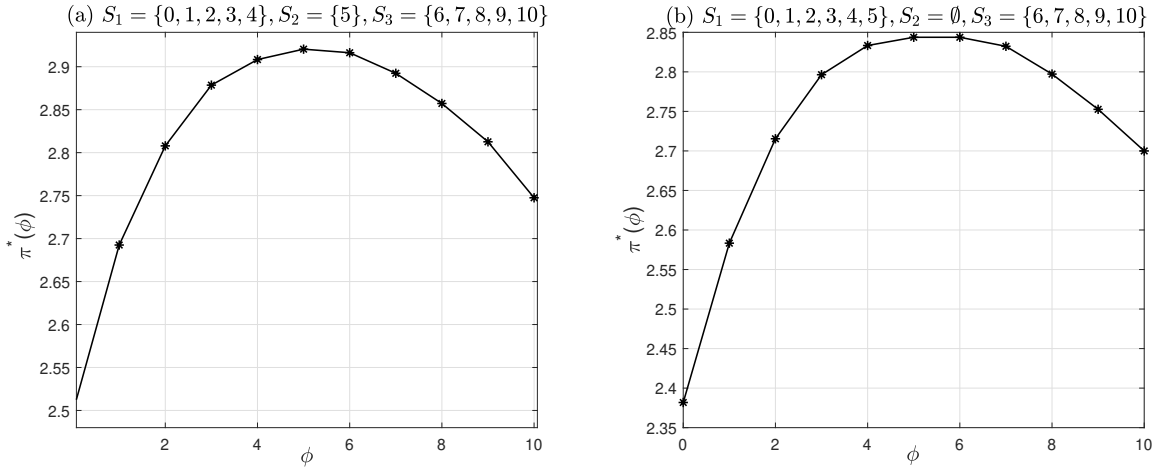
(iii) *For* $\phi_2, \phi_2' \in \mathcal{S}_2$, $\pi^*(\phi_2) = \pi^*(\phi_2')$.

(iv) *If* $\phi_3 \in \mathcal{S}_3$, *then* $\pi^*(\phi_3) \le \pi^*(\phi_3 - 1)$.

The proof of Proposition 1 can be found in Appendix EC.2.3. Note that Proposition 1(i) provides an ordering for the sets given in (44). We consider two cases to illustrate Proposition 1(ii)-(iv).

1. If $\mathcal{S}_2 \neq \emptyset$, then from Proposition 1(ii), $\pi^*(\phi_2) \ge \pi^*(\phi_1)$ for any $\phi_2 \in \mathcal{S}_2$ and $\phi_1 \in \mathcal{S}_1$, and from Proposition 1(iv), $\pi^*(\phi_2) \ge \pi^*(\phi_3)$ for any $\phi_2 \in \mathcal{S}_2$ and $\phi_3 \in \mathcal{S}_3$. This is demonstrated in Figure 2(a). In this case, we can choose $\phi^* = \phi_2^*$ in which ϕ_2^* is the smallest index in $\mathcal{S}_2$.
2. If $\mathcal{S}_2 = \emptyset$, denote by ϕ_3^* the smallest index in $\mathcal{S}_3$ (recall that $J \in \mathcal{S}_3$ so $\mathcal{S}_3 \neq \emptyset$). Then from Proposition 1(ii) $\pi^*(\phi_3^*) \ge \pi^*(\phi_1)$ for any $\phi_1 \in \mathcal{S}_1$ and from Proposition 1(iv) $\pi^*(\phi_3^*) \ge \pi^*(\phi_3)$ for any $\phi_3 \in \mathcal{S}_3$. This is demonstrated in Figure 2(b). In this case, we can choose $\phi^* = \phi_3^*$.

Figure 2 **Properties of** $\pi^*(\phi)$



Note. The parameters are as follows: $I = J = 10$; $\mu_{c,i} = \mu_{s,j} = \theta_{c,i} = \theta_{s,j} = r_c = r_{s,j} = 1$ and $p_{ij} = 1/10$ for all i and j; $w_c^H = w_c^A = (0.5, 0.75, 1, 1.25, 1.5, 1.75, 2, 2.25, 2.5, 5)$, $w_s^H = w_s^A = 0.5 \times w_c^H$; $\lambda = 1.05$ for panel (a) and $\lambda = 1.00$ for panel (b) with $\lambda_{c,i}$ being identical across all i.

As a result, one can choose ϕ^* as the smallest index in $\mathcal{S}_2 \cup \mathcal{S}_3$, which can be represented as

$$\phi^* := \min\left\{\phi \in [\mathcal{J}] : \sum_{i\in\mathcal{I}_+(\phi)} \lambda_{c,i}\chi_{i,\phi+1} \geq 1\right\}. \tag{45}$$

The choice of ϕ^* might not be unique. For example, we can also choose $\phi^* = 5$ in Figure 2(b).

With ϕ^* given in (45) and the solution to sub-problem ϕ^* in Proposition EC.1, we can get the optimal solution to (34) in the following theorem, whose proof can be found in Appendix EC.2.4.

THEOREM 2. *Let ϕ^* be given in (45) and arrange the indices in $\mathcal{I}$ such that*

$$\frac{\xi_{1,\phi^*}}{\chi_{1,\phi^*}} \geq \cdots \geq \frac{\xi_{i_+(\phi^*),\phi^*}}{\chi_{i_+(\phi^*),\phi^*}} \geq 0 > \frac{\xi_{i_+(\phi^*)+1,\phi^*}}{\chi_{i_+(\phi^*)+1,\phi^*+1}} \geq \cdots \geq \frac{\xi_{I,\phi^*}}{\chi_{I,\phi^*+1}}. \tag{46}$$

Then, the optimal solution to problem (34) can be characterized as follows.

(i) *If $\mathcal{S}_2 \neq \emptyset$, then $\phi^* \in \mathcal{S}_2$. Let $\varphi^*_{c,i} = \lambda_{c,i}$ for $i \in \mathcal{I}_+(\phi^*)$ and $\varphi^*_{c,i} = 0$ for $i \in \mathcal{I}_-(\phi^*)$; $\varphi^*_{s,j} = \sum_{i\in\mathcal{I}_+(\phi^*)} \lambda_{c,i}p_{ij}$ for $j \leq \phi^*$ and $\varphi^*_{s,j} = 0$ for $j > \phi^*+1$, with φ^*_{s,ϕ^*+1} being set so that the first constraint in problem (34) is binding, i.e., $\varphi^*_{s,\phi^*+1} = (1 - \sum_{i\in\mathcal{I}_+(\phi^*)} \chi_{i,\phi^*}\varphi^*_{c,i})r_{s,\phi^*+1}\mu_{s,\phi^*+1}$.*

(ii) *Otherwise, $\phi^* \in \mathcal{S}_3$. Let $i_{\phi^*} := \min\{i' \in \mathcal{I}_+(\phi^*) : \sum_{i\leq i'} \lambda_{c,i}\chi_{i,\phi^*} \geq 1\} \leq i_+(\phi^*)$, $\varphi^*_{c,i} = \lambda_{c,i}$ for $i < i_{\phi^*}$ and $\varphi^*_{c,i} = 0$ for $i > i_{\phi^*}$, with $\varphi^*_{c,i_{\phi^*}}$ being set so that the first constraint in problem (34) is binding, i.e., $\varphi^*_{c,i_{\phi^*}} = (1 - \sum_{i<i_{\phi^*}} \chi_{i,\phi^*}\lambda_{c,i})/\chi_{i_{\phi^*},\phi^*}$; $\varphi^*_{s,j} = \sum_{i\in\mathcal{I}_+(\phi^*)} \varphi^*_{c,i}p_{ij}$ for $j \leq \phi^*$, and $\varphi^*_{s,j} = 0$ for $j > \phi^*$. In particular, if $\phi^* = J$, then $\varphi^*_{s,j} = \sum_{i\in\mathcal{I}} \varphi^*_{c,i}p_{ij}$ for all $j \in \mathcal{J}$.*

Based on Theorem 2, we can propose a sequence of policies for the sequence of stochastic systems considered in Section 3.2. For the mth system, a natural candidate is to allocate $N_c^{m,*} := m \cdot \sum_{i\in\mathcal{I}} \frac{\varphi^*_{c,i}}{\mu_{c,i}}$ (resp., $N_j^{m,*} = m \cdot \frac{\varphi^*_{s,j}}{\mu_{s,j}}$) servers to stage 1 (resp., station j in stage 2). Then, among the servers in stage 1, dedicate a fixed proportion $\frac{\varphi^*_{c,i}/\mu_{c,i}}{\sum_{i\in\mathcal{I}} \varphi^*_{c,i}/\mu_{c,i}}$ of the $N_c^{m,*}$ servers to class i. However, this policy has obvious shortcomings, that is, the servers in stage 1 are dedicated and thus service capacity cannot be shared among classes (Atar et al. 2010), which will potentially lead to unnecessary server idleness. To overcome this, we propose the following policy for the mth system, denoted by $\psi^{m,*} = (N^{m,*}, \omega^{m,*})$, in which we use an index-based priority scheduling rule for the first stage. Under this index-based scheduling rule, once a server completes service, it always admits the customer that is currently waiting in line and has the highest priority. This can avoid server idling when there are customers waiting.

(MM1) The capacity allocation is given by $N^{m,*} = (N_c^{m,*}, N_s^{m,*})$, in which $N_c^{m,*} = m \cdot \sum_{i\in\mathcal{I}} \varphi^*_{c,i}/\mu_{c,i}$ and $N_j^{m,*} = m \cdot \varphi^*_{s,j}/\mu_{s,j}$, with $j \in \mathcal{J}$.

(MM2) For classes in stage 1, a class with a smaller index is given a higher priority.

The following theorem establishes the asymptotic optimality of the above policy. Its proof can be found in Appendix EC.2.5.

THEOREM 3. *Under the sequence of policies* $\{\psi^{m,*}; m \in \mathbb{N}\}$, *condition* (32) *holds. As a result, the sequence of policies* $\{\psi^{m,*}; m \in \mathbb{N}\}$ *is asymptotically optimal.*

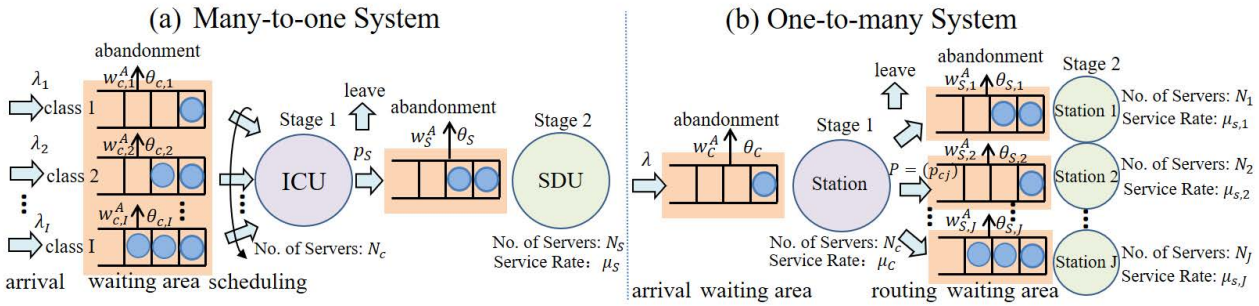
Theorems 2 and 3 suggest that the classes in stage 1 and stations in stage 2 can be categorized into three groups: (i) the first group includes the classes or stations whose service requirements can be fully satisfied (i.e. class i in stage 1 with $\varphi^*_{c,i} = \lambda_{c,i}$, or station j in stage 2 with $\varphi^*_{s,j} = \sum_{i\in\mathcal{I}} \varphi^*_{c,i} p_{ij}$); (ii) the second group includes at most a single class or station whose service requirement is partially satisfied (i.e., station ϕ^*+1 in stage 2 if $\phi^* \in \mathcal{S}_2$ and $0 < \varphi^*_{s,\phi^*+1} < \sum_{i\in\mathcal{I}} \varphi^*_{c,i} p_{i(\phi^*+1)}$, or class i_{ϕ^*} in stage 1 if $\phi^* \in \mathcal{S}_3$ and $\varphi^*_{c,i_{\phi^*}} < \lambda_{c,i_{\phi^*}}$, that consumes all the remaining resources); and (iii) the third group includes classes or stations receiving no service capacity (i.e., classes i with $\varphi^*_{c,i} = 0$, or station j with $\varphi^*_{s,j} = 0$).

Finally, we note that the scheduling part of the policy $\psi^{m,*}$ relies on ϕ^*, i.e., the index-based scheduling rule is ϕ^*-dependent. To better understand the underlying factors of the optimal policy, we derive explicit scheduling indexes for two simpler settings in the following section: (i) the many-to-one system in which the scheduling rule is ϕ^*-independent; and (ii) the one-to-many system in which the scheduling rule is absent.

4.2. Two Special Systems

In this subsection, we consider two special systems as depicted in Figure 3: (i) a many-to-one system ($J = 1$), where stage 2 has one station with many identical servers serving all customers from stage 1 that need additional service; and (ii) a one-to-many system ($I = 1$), where there is only one class of customers arriving at stage 1, and there are J stations in stage 2.

Figure 3 **The Special Systems**



4.2.1. Many-to-one System This structure is relevant to the scenarios in which the service in stage 2 does not require specialized skills as in stage 1. One example is the critical-care system in hospitals, in which stage 1 represents the ICU and stage 2 represents the SDU. Patients of various severity levels are evaluated using different scoring systems (Hu et al. 2018) and first treated in the ICU, often by different specialists depending on the patients' illness. After their conditions are stabilized and less intensive care is required, the patients are then transferred to a common SDU, which provides intermediate level of care (Armony et al. 2018). For notational simplicity, we replace the subscript "(s,j)" or "j" by "s" to indicate that there is just one station at stage 2. For example, p_{is} is the probability of a class-i customer in stage 1 transferring to stage 2.

One can verify that the parameters in (37) and (42) are respectively as follows:

$$\chi_{i,\phi} = \begin{cases} \frac{1}{r_c\mu_{c,i}}, & \phi = 0, \\ \frac{1}{r_c\mu_{c,i}} + \frac{p_{is}}{r_s\mu_s}, & \phi = 1, \end{cases} \quad \text{and} \quad \xi_{i,\phi} = \begin{cases} \frac{w_{c,i}}{\theta_{c,i}} - \frac{w_s r_s \mu_s}{\theta_s}\chi_{i,1}, & \phi = 0, \\ \frac{w_{c,i}}{\theta_{c,i}}, & \phi = 1. \end{cases}$$

The following proposition formalizes the solution to the fluid optimization problem in the many-to-one system. It serves as a corollary to Theorem 2 and therefore the proof is omitted. Note that the optimal solution depends on whether $\phi^* \in \mathcal{S}_2$ or $\phi^* \in \mathcal{S}_3$. Referring to the results in Section 4.1 and by condition (30), we have $\mathcal{S}_0 = \emptyset$, and $1 \in \mathcal{S}_3$. Hence, the choice of ϕ^* depends on whether $0 \in \mathcal{S}_1$, $0 \in \mathcal{S}_2$ or $0 \in \mathcal{S}_3$.

PROPOSITION 2. (i) *If* $\sum_{i\in\mathcal{I}_+(0)} \lambda_{c,i}\chi_{i,1} < 1$, *then* $0 \in \mathcal{S}_1$, $\mathcal{S}_2 = \emptyset$, *and* $\phi^* = 1 \in \mathcal{S}_3$. *By substituting* $\phi^* = 1$ *into* (46), *we arrange indices in* $\mathcal{I}$ *such that* $\frac{\xi_{1,1}}{\chi_{1,1}} \geq \cdots \geq \frac{\xi_{I,1}}{\chi_{I,1}} \geq 0$, *by noting that* $\mathcal{I}_+(1) = \mathcal{I}$ *and* $i_+(\phi^*) = I$ *for* $\phi^* = 1$. *Then we have*

$$\varphi_c^* = \left(\lambda_{c,1}, \ldots, \lambda_{c,i^*-1}, \frac{1 - \sum_{i=1}^{i^*-1} \lambda_{c,i}\chi_{i,1}}{\chi_{i^*,1}}, 0, \ldots, 0\right),$$

with $i^* = \min\{i' \in \mathcal{I} : \sum_{i=1}^{i'} \lambda_{c,i}\chi_{i,1} > 1\}$, *and* $\varphi_s^* = \sum_{i\in\mathcal{I}} \varphi_{c,i}^* p_{is}$.

(ii) *If* $\sum_{i\in\mathcal{I}_+(0)} \lambda_{c,i}\chi_{i,1} \geq 1 \geq \sum_{i\in\mathcal{I}_+(0)} \lambda_{c,i}\chi_{i,0}$, *then* $0 \in \mathcal{S}_2$, $\mathcal{S}_1 = \emptyset$ *and thus* $\phi^* = 0 \in \mathcal{S}_2$. *By substituting* $\phi^* = 0$ *into* (46), *we arrange indices in* $\mathcal{I}$ *such that*

$$\frac{\xi_{1,0}}{\chi_{1,0}} \geq \cdots \geq \frac{\xi_{i_+(0),0}}{\chi_{i_+(0),0}} \geq 0 > \frac{\xi_{i_+(0)+1,0}}{\chi_{i_+(0)+1,1}} \geq \cdots \geq \frac{\xi_{I,0}}{\chi_{I,1}} \tag{47}$$

with $i_+(0) = \max\{i \in \mathcal{I} : \xi_{i,0} \geq 0\}$. *Then we have*

$$\varphi_c^* = \left(\lambda_{c,1}, \ldots, \lambda_{c,i_+(0)}, 0, \ldots, 0\right),$$

and $\varphi_s^* = r_s\mu_s(1 - \sum_{i\in\mathcal{I}_+(0)} \varphi_{c,i}^*\chi_{i,0})$.

(iii) *If* $\sum_{i\in\mathcal{I}_+(0)}\lambda_{c,i}\chi_{i,1}\geq\sum_{i\in\mathcal{I}_+(0)}\lambda_{c,i}\chi_{i,0}>1$, *then* $0\in\mathcal{S}_3$, $\mathcal{S}_2=\emptyset$ *and thus* $\phi^*=0\in\mathcal{S}_3$. *By substituting* $\phi^*=0$ *into* (46), *we arrange* $\mathcal{I}$ *using* (47). *Then we have*

$$\varphi_c^*=\left(\lambda_{c,1},\ldots,\lambda_{c,i^*-1},\frac{1}{\chi_{i^*,0}}\left(1-\sum_{i=1}^{i^*-1}\lambda_{c,i}\chi_{i^*,0}\right),0,\ldots,0\right),$$

with $i^*=\min\{i'\in\mathcal{I}:\sum_{i=1}^{i'}\lambda_{c,i}\chi_{i,0}\geq 1\}\leq i_+(0)$, *and* $\varphi_s^*=0$.

Though the indexes are still arranged differently for different cases in Proposition 2, we can get an index-based rule which does not depend on the cases for scheduling stage-1 classes in the stochastic system. For this, recall that $\mathcal{I}_+(0):=\{i\in\mathcal{I}:\xi_{i,0}\geq 0\}$ and $\mathcal{I}_-(0):=\{i\in\mathcal{I}:\xi_{i,0}<0\}$. Note that because $\xi_{i,1}=\xi_{i,0}+\frac{w_s r_s\mu_s}{\theta_s}\chi_{i,1}$, one has

$$\frac{\xi_{i,1}}{\chi_{i,1}}=\frac{\xi_{i,0}}{\chi_{i,1}}+\frac{w_s r_s\mu_s}{\theta_s}.$$

Therefore, for Case (i) of Proposition 2, the ordering using $\xi_{i,1}/\chi_{i,1}$ is the same as the one using $\xi_{i,0}/\chi_{i,1}$, that is, the indexes can be arranged as $\xi_{i,0}/\chi_{i,1}\geq\xi_{i+1,0}/\chi_{i+1,1}$ for $i<I$. That is,

$$\frac{\xi_{1,0}}{\chi_{1,1}}\geq\cdots\geq\frac{\xi_{i_+(0),0}}{\chi_{i_+(0),1}}\geq 0>\frac{\xi_{i_+(0)+1,0}}{\chi_{i_+(0)+1,1}}\geq\cdots\geq\frac{\xi_{I,0}}{\chi_{I,1}}.$$

Because $\sum_{i\in\mathcal{I}_+(0)}\lambda_{c,i}\chi_{i,1}<1$, one has $i^*>i_+(0)$, which implies that $\varphi_{c,i}^*=\lambda_{c,i}$ for $i\in\mathcal{I}_+(0)$; that is, a sufficient amount of capacity is allocated to $\mathcal{I}_+(0)$. Thus, the ordering in $\mathcal{I}_+(0)$ does not matter, and for those classes, we re-order them such that $\frac{\xi_{1,0}}{\chi_{1,0}}\geq\cdots\geq\frac{\xi_{i_+(0),0}}{\chi_{i_+(0),0}}$. For classes in $\mathcal{I}_-(0)$, we keep the ordering unchanged. Then, we have the ordering in (47). As a result, the ordering in (47) can also be applied to Case (i) of Proposition 2, and hence all three cases in Proposition 2.

Based on the above discussion, we propose a sequence of policies for the sequence of stochastic systems, which is (with some abuse of notation) denoted by $\psi^{m,*}=(N^{m,*},\omega^{m,*})$ for the mth system.

(MO1) The allocation vector is $N^{m,*}=(N_c^{m,*},N_s^{m,*})$, in which $N_c^{m,*}=m\cdot r_c\sum_{i\in\mathcal{I}}(\varphi_{c,i}^*/\mu_{c,i})$, and $N_s^{m,*}=m\cdot r_s\varphi_s^*/\mu_s$.

(MO2) The scheduling rule $\omega^{m,*}$ prioritizes classes in $\mathcal{I}$ according to (47), with a smaller index receiving a higher priority.

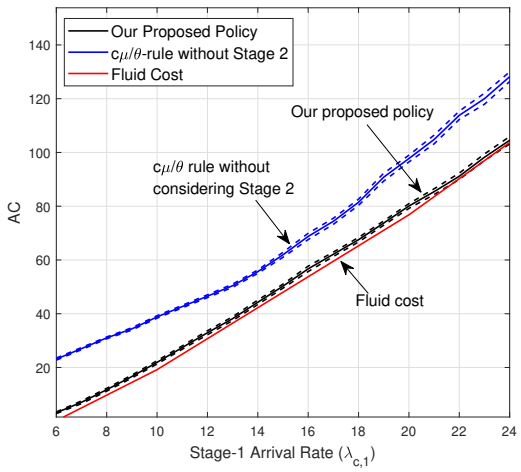
Because this is a special case of the many-to-many system, from Theorems 1 and 3, we can claim that the sequence of policies $\psi^{m,*}$ is asymptotically optimal when m becomes large. Due to the simpler structure of the many-to-one system, we get an index-based scheduling rule (47), whose values can be calculated using system primitives.

We remark that the scheduling part is different from the $c\mu/\theta$ rule in Atar et al. (2010), even when no resource is assigned to the second-stage station. This is interesting because if no resource

is assigned to the second stage, then one may want to ignore the second stage and reduce the two-stage system into a single-stage one, and then apply the $c\mu/\theta$ rule from Atar et al. (2010). However, our results indicate that doing so would result in a sub-optimal policy. This is because the parameters associated with stage 2 play important roles in our proposed policy (i.e., how the customers would be prioritized and the resource would be allocated between stage 1 and 2), and hence stage 2 shall not be ignored even if it does not receive any resource.

EXAMPLE 1. Here we provide an example to illustrate this observation. Consider a many-to-one system with two classes in stage 1, in which $m = 100$ and $\lambda^m_{c,1} = \lambda^m_{c,2}$ with $\lambda^m_{c,1}$ varying from 6 to 24. Assume that $w_{c,1} = 5 > w_{c,2} = w_s = 2$, $\theta_{c,1} = \theta_{c,2} = \theta_s = 1$, $r_c\mu_{c,1} = r_s\mu_s = \frac{1}{5} < r_c\mu_{c,2} = \frac{5}{9}$, and $p_{is} = 1$ for $i = 1, 2$. Then we have $\mathcal{I}_+(0) = \{1\}$ and $\mathcal{I}_-(0) = \{2\}$ by calculating the indices in (47) as $\frac{\xi_{1,0}}{\chi_{1,0}} = 5 > \frac{\xi_{2,0}}{\chi_{2,1}} = -\frac{9}{85}$. Therefore, our proposed policy prioritizes class 1 over class 2, in contrast the $c\mu/\theta$ rule prioritizes class 2 because $\frac{w_{c,2}r_c\mu_{c,2}}{\theta_{c,2}} = \frac{10}{9} > \frac{w_{c,1}r_c\mu_{c,1}}{\theta_{c,1}} = 1$ (where the unit cost $w_{c,i}$ and service rate per unit of resource $r_c\mu_{c,i}$ correspond to c and μ, respectively in the $c\mu/\theta$ rule). Figure 4 shows that our proposed policy outperforms the $c\mu/\theta$ rule within the entire parameter range, especially for the case when $\lambda^m_{c,1} > 20$, in which we have $\sum_{i\in\mathcal{I}_+(0)} \lambda_{c,i}\chi_{i,0} > 1$ and hence $\varphi^*_s = 0$ (i.e., stage 2 receives no capacity and thus is absent by Proposition 2). This illustrates the cost of simply ignoring stage 2 when designing the scheduling rule.

Figure 4 The Value of Considering Stage 2 in Scheduling



Note. The fluid cost (red line) is the optimal objective value to problem (34), whereas the other (simulated) costs are obtained by running the system under the corresponding policies.

4.2.2. One-to-many System This special structure is relevant when it is difficult or too costly to identify customer classes and provide differentiated services at stage 1. For example,

incoming calls to a customer service center, with their service requests unknown, are placed in a single queue and answered by an operator in stage 1. After the service in stage 1 is completed, further service requests may be needed with clearly specified service requests, and then the customer is routed to a specific department in stage 2. Another example is from the healthcare context, where stage 1 service corresponds to the triage/registration for patients and stage 2 service corresponds to the medical treatment.

For this system, because there is only one class in stage 1, there is no need to consider scheduling. Hence, the manager only needs to decide the capacity allocation. For notational simplicity, we replace the subscript "(c,i)" or "i" by "c" to indicate that there is just one class at stage 1. For example, p_{cj} is the probability of a customer in stage 1 transferring to station j in stage 2.

The parameters in (37) and (42) for the one-to-many system are as follows:

$$\chi_{c,\phi} = \frac{1}{r_c\mu_c} + \sum_{j=1}^{\phi} \frac{p_{cj}}{r_{s,j}\mu_{s,j}} \quad \text{and} \quad \xi_{c,\phi} = \begin{cases} \frac{w_c}{\theta_c} - \sum_{j=\phi+1}^{J} \frac{w_{s,j}}{\theta_{s,j}} p_{cj} - \frac{w_{s,\phi+1} r_{s,\phi+1}\mu_{s,\phi+1}}{\theta_{s,\phi+1}} \chi_{c,\phi}, & \phi < J, \\ \frac{w_c}{\theta_c}, & \phi = J. \end{cases}$$

Note that $\mathcal{I} = \{c\}$. One can verify that $\mathcal{S}_1 = \emptyset$. Define

$$\phi_c := \min\{\phi \in \mathcal{J} : \xi_{c,\phi} \geq 0\}.$$

Then $\mathcal{I}_+(\phi) = \{c\}$ for $\phi \geq \phi_c$, and $\mathcal{I}_+(\phi) = \emptyset$ for $\phi < \phi_c$ by Lemma 4. Then (45) becomes $\phi^* := \min\{\phi \geq \phi_c : \lambda_c \chi_{c,\phi+1} \geq 1\}$. The following proposition gives the solution to the fluid optimization problem of the one-to-many system. It follows directly from Theorem 2, hence the proof is omitted.

PROPOSITION 3. *The optimal solutions are*

(i) *If $\lambda_c \chi_{c,\phi_c} < 1$, then $\mathcal{S}_2 \neq \emptyset$, and $\phi^* = \min\{\phi \geq \phi_c : \lambda_c \chi_{c,\phi+1} \geq 1\} \in \mathcal{S}_2$. We have $\varphi_c^* = \lambda_c$, and*

$$\varphi_s^* = \left(\lambda_c p_{c1}, \ldots, \lambda_c p_{c\phi^*}, \left(1 - \sum_{j=1}^{\phi^*} \lambda_c \chi_{c,j}\right) r_{s,\phi^*+1}\mu_{s,\phi^*+1}, 0, \ldots, 0\right).$$

(ii) *If $\lambda_c \chi_{c,\phi_c} \geq 1$, then $\mathcal{S}_2 = \emptyset$, and $\phi^* = \phi_c \in \mathcal{S}_3$. We have $\varphi_c^* = 1/\chi_{c,\phi^*}$, and*

$$\varphi_s^* = \left(\frac{p_{c1}}{\chi_{c,\phi^*}}, \ldots, \frac{p_{c\phi^*}}{\chi_{c,\phi^*}}, 0, \ldots, 0\right).$$

In particular, if $\phi^* = J$, then we have $\varphi_{s,j}^* = p_{cj}/\chi_{c,j}$ for all $j \in \mathcal{J}$.

Proposition 3 can be translated to a sequence of policies $\{\psi^{m,*}\}$. The policy $\psi^{m,*}$ only contains the allocation vector $N^{m,*} = (N_c^{m,*}, N_s^{m,*})$ with $N_c^* = m \cdot r_c \varphi_c^*/\mu_c$ and $N_{s,j}^{m,*} = m \cdot r_{s,j}\varphi_{s,j}^*/\mu_{s,j}$ for the mth stochastic system. The scheduling rule is absent for this special case because only a single class arrives in stage 1.

Recall that the allocation rule in stage 2 is an index-based rule according to (36). Note that $\xi_{c,\phi}$ captures the cost reduction (increment if negative) when transferring resources from station $\phi+1$ to serve one stage-1 customer and its subsequent stage-2 service requirements up to station ϕ. To be specific, it comprises of three parts: (i) w_c/θ_c measures the cost reduction by fulfilling a customer's service requirement in stage 1; (ii) $\sum_{j=\phi+1}^{J} w_{s,j}p_{cj}/\theta_{s,j}$ measures the expected cost increment due to the transfer of the customer to stations $\phi+1,\cdots,J$ in stage 2; (iii) $(w_{s,\phi+1}/\theta_{s,\phi+1})r_{s,\phi+1}\mu_{s,\phi+1}\chi_{c,\phi}$ measures the cost increment due to resources being transferred out of station $\phi+1$ (here $r_{s,\phi+1}\mu_{s,\phi+1}\chi_{c,\phi}$ measures the loss in service capacity due to resources being transferred out of station $\phi+1$ by noting that $\chi_{c,\phi}$ represents the amount of resources required for serving one stage-1 customer, as well as its expected subsequent stage-2 service requirements up to station ϕ). Hence, transferring resources from station $\phi+1$ in stage 2 with $\phi \geq \phi_c$ to stage 1 is beneficial because the cost is reduced, i.e., $\xi_{c,\phi} \geq 0$; while transferring any resources from station $\phi+1$ in stage 2 with $\phi < \phi_c$ to stage 1 is not, because it leads to cost increment, i.e., $\xi_{c,\phi} < 0$ (equivalently, it is beneficial to transfer resources from stage 1 to station $\phi+1$ in stage 2 with $\phi < \phi_c$). Also note that $\lambda_c\chi_{c,\phi_c}$ represents the total fluid-scaled resources that are needed for serving all customers in stage 1 as well as their subsequent service requirements till station ϕ_c, then we have the following explanation for the two scenarios in Proposition 3:

1. If $\lambda_c\chi_{c,\phi_c} < 1$, then there are sufficient resources to serve all customers in stage 1 as well as their expected subsequent service requirements till station ϕ_c. Then, we can allocate sufficient resources to stage 1 as well as those stations, i.e., $\varphi_c^* = \lambda_c$ and $\varphi_{s,j}^* = \lambda_c p_{cj}$ for $j \leq \phi_c$. For the remaining resources, one can allocate them to stations according to the $wr\mu/\theta$-index as specified in (36) until it is exhausted, i.e., $\varphi_{s,j}^* = \lambda_c p_{cj}$ for $\phi_c < j \leq \phi^*$, with the service requirement in station ϕ^*+1 being partially satisfied. This corresponds to the first scenario in Proposition 3.
2. If $\lambda_c\chi_{c,\phi_c} \geq 1$, then there are no sufficient resources to serve all customers in stage 1 as well as their expected subsequent service requirements till station ϕ_c. Then, because it is beneficial to transfer resources from stage 1 to station $\phi+1$ in stage 2 with $\phi < \phi_c$, it is better to reserve sufficient amount of resources to station ϕ with $\phi \leq \phi_c$ in stage 2 to guarantee their service requirements being fully satisfied. Hence, we have $\varphi_c^* = 1/\chi_{c,\phi_c} < \lambda_c$ and $\varphi_{s,j}^* = p_{cj}/\chi_{c,\phi_c}$ for $j \leq \phi_c$. This case corresponds to the second scenario in Proposition 3.

4.3. Grouping and Pooling

The discussion after Theorem 2 inspires us to consider the following strategy to streamline the operations of the many-to-many system. First, merge multiple classes in stage 1 to form (at most)

three service groups as follows: $\mathcal{I}_1 := \{i \in \mathcal{I} : \varphi^*_{c,i} = \lambda_{c,i}\}$, $\mathcal{I}_3 := \{i \in \mathcal{I} : \varphi^*_{c,i} = 0\}$, and $\mathcal{I}_2 := \mathcal{I} \setminus (\mathcal{I}_1 \cup \mathcal{I}_3)$. Here, group $\mathcal{I}_2$ may be empty and contains at most one element only if $\phi^* \in \mathcal{S}_3$. Second, pool the stations in the second stage to form at most three clusters: $\mathcal{J}_1 := \{j \in \mathcal{J} : \varphi^*_{s,j} = \sum_{i\in\mathcal{I}} \varphi^*_{c,i} p_{ij}\}$, $\mathcal{J}_3 := \{j \in \mathcal{J} : \varphi^*_{s,j} = 0\}$, and $\mathcal{J}_2 := \mathcal{J} \setminus (\mathcal{J}_1 \cup \mathcal{J}_3)$. Cluster $\mathcal{J}_2$ may be empty and contains at most one element only if $\phi^* \in \mathcal{S}_2$. Therefore, at least one of $\mathcal{I}_2$ or $\mathcal{J}_2$ is empty.

With the grouping and pooling, we propose the following policy, denoted by $\psi^{m,\mathrm{GP}} = (N^{m,\mathrm{GP}}, \phi^{m,\mathrm{GP}})$, for the resulting mth system (GP).

(GP1) The capacity allocation is given by $N^{m,\mathrm{GP}} = (N_c^{m,\mathrm{GP}}, N_{s,\mathcal{J}_1}^{m,\mathrm{GP}}, N_{s,\mathcal{J}_2}^{m,\mathrm{GP}}, N_{s,\mathcal{J}_3}^{m,\mathrm{GP}})$, in which $N_c^{m,\mathrm{GP}} = N_c^{m,*}$ and $N_{s,\mathcal{J}_k}^{m,\mathrm{GP}} = \sum_{j\in\mathcal{J}_k} N_j^{m,*}$ for $k = 1, 2, 3$. (Here, $N_{s,\mathcal{J}_3}^{m,\mathrm{GP}} = 0$.)

(GP2) The scheduling decision $\phi^{m,\mathrm{GP}}$ assigns the highest (resp., lowest) priority to classes in group $\mathcal{I}_1$ (resp., $\mathcal{I}_3$), while adopting the FCFS discipline among classes within the same group.

In Section 5.2, we will use numerical experiment to illustrate that the GP system under $\psi^{m,\mathrm{GP}}$ performs similarly to the original system under $\psi^{m,*}$.

5. Numerical Studies

In this section, we conduct numerical experiments to illustrate our results and discuss additional insights. In particular, Section 5.1 examines how well the fluid-approximated cost predicts the actual system performance; and Section 5.2 evaluates the system performance under optimal pooling and grouping based on fluid solution discussed in Section 4.3.

5.1. Validity of the Fluid Approximation

Theorem 3 claims that under the sequence of proposed policy $\{\psi^{m,*}; m \in \mathbb{N}\}$, the expected long-run average cost of the stochastic system converges to the fluid model solution. Below, we investigate the empirical accuracy of the fluid approximation. Specifically, we consider systems that have 3 classes of customers arriving in stage 1 and 3 stations in stage 2. Denote by λ the total arrival rate to stage 1 and assume that $\lambda_{c,i} = \lambda/3$ for $i \in \mathcal{I}$; that is, the arrival rates of these three customer classes in stage 1 are the same. Let the service rate vectors be $\mu_c = (2, 2, 2)$ and $\mu_s = (1, 1, 1)$. The routing matrix (p_{ij}) is set to be $p_{ij} = 0.3$ for all i and j. The conversion rates from resource to servers are $r_c = 1$ and $r_{s,j} = 2$ for $j \in \{1, 2, 3\}$. The system size, i.e., amount of resource m, varies from 300 to 1,050 in increments of 50, and λ is chosen so that $\sum_{i\in\mathcal{I}} \frac{\lambda_{c,i}}{m} \left(\frac{1}{r_c\mu_{c,i}} + \sum_{j\in\mathcal{J}} \frac{p_{ij}}{r_j\mu_{s,j}}\right) = 1.5$ (correspondingly, λ varies from 473 to 1,658). For simplicity, we assume $\theta_{c,i} = \theta_{s,j} = 1$ for $i \in \mathcal{I}$ and $j \in \mathcal{J}$, that is, the abandonment rates of all customer classes are the same. We set the waiting cost

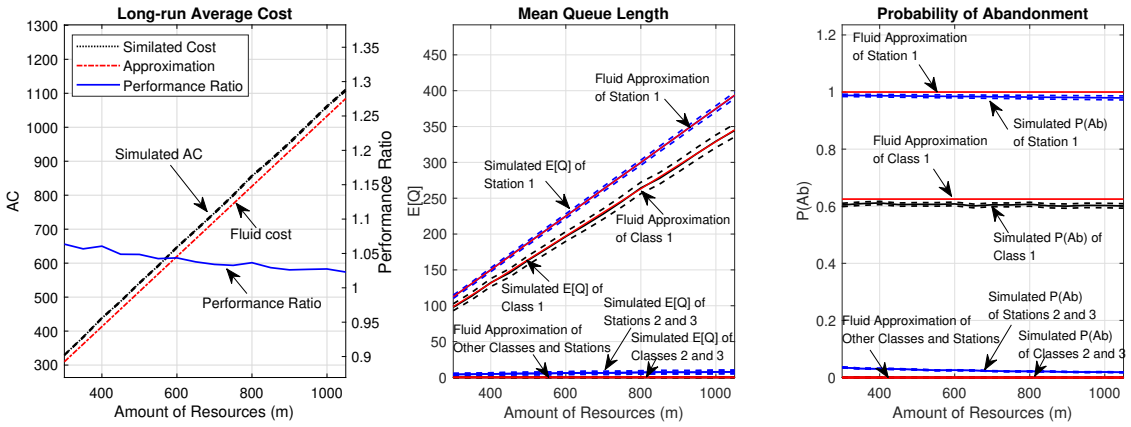
rate vectors $w_c^H = (1,2,5)$ and $w_s^H = (0.5,1,2)$, and the abandonment cost rate vectors $w_c^A = (1,2,5)$ and $w_s^A = (0.5,1,2)$ for stage 1 and 2, respectively.

In this numerical study, we consider the following performance measures: (i) long-run average cost, which is approximated by the objective value of Problem (29); (ii) expected queue length, which is approximated by $m \cdot q$ with $q = (q_c, q_s)$; and (iii) probability of abandonment $\mathbb{P}\{Ab\}$, which is the percentage of abandoned customers among all the arrivals. The probability of abandonment $\mathbb{P}\{Ab\}$ is approximated by

$$\mathbb{P}\{Ab\} \approx \frac{\text{Arrival rate to each class or station} - \text{Total capacity allocated to that class or station}}{\text{Arrival rate to each class or station}},$$

i.e., $(\lambda_{c,i} - \varphi_{c,i})/\lambda_{c,i}$ for $i \in \mathcal{I}$, and $(\sum_{i\in\mathcal{I}} \varphi_{c,i} p_{ij} - \varphi_{s,j})/\sum_{i\in\mathcal{I}} \varphi_{c,i} p_{ij}$ for $j \in \mathcal{J}$.

Figure 5 **Accuracy of the Fluid Approximation**



In Figure 5(a), the simulated costs (point estimates using solid lines together with the corresponding 95% confidence intervals using dashed lines), which are obtained by running the system repeatedly for 100 times under the policy based on the fluid solutions, are used to represent the true system costs. We also present the performance ratio in the figures, which is the ratio of the simulated cost to fluid-approximated cost. As can be seen from the figures, the fluid-approximated costs are close to, but generally lower than, the simulated costs under different m. Moreover, the performance ratios are less than 1.1 and have a converging trend to 1 as the system size m increases. Figures 5(b) and (c) demonstrate that the fluid approximation of queue lengths and probability of abandonment are quite accurate when comparing to their simulated counterparts. This suggests that our fluid model provides good approximations to the stochastic system under the proposed policy.

5.2. Performance of the GP System

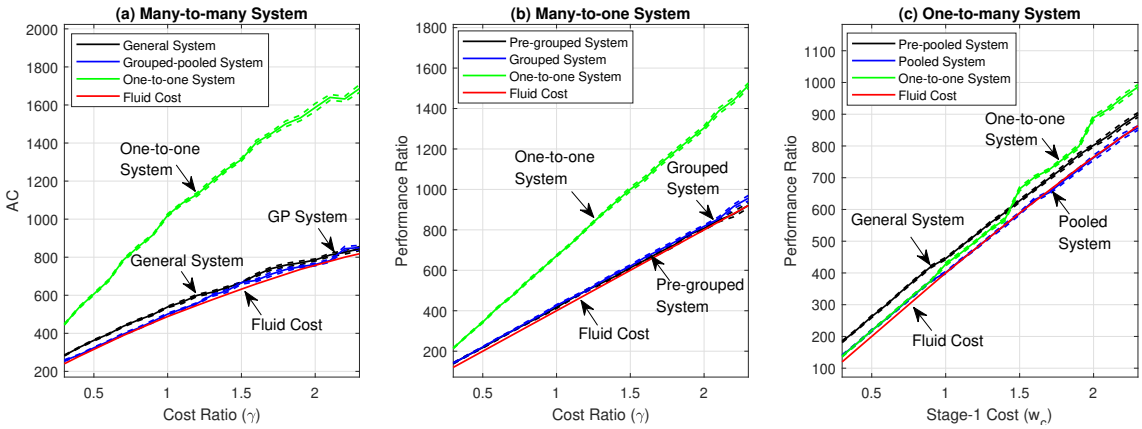
In Section 4.3, we propose a GP system, which has a simpler system structure than the original many-to-many one. Here we numerically illustrate that the GP system under $\psi^{m,\mathrm{GP}}$ performs similarly to the original system under $\psi^{m,*}$.

We first consider a many-to-many system, with 10 classes in stage 1 and 10 stations in stage 2, i.e., $\mathcal{I}=\mathcal{J}=\{1,\ldots,10\}$. The parameters for the original system are as follows: the total arrival rate is $\lambda=400$, with equal arrival rate to each class, i.e., $\lambda_{c,i}=\lambda/10$ for $i\in\mathcal{I}$. The service rates and abandonment rates are $\mu_{c,i}=\mu_{s,j}=1$ and $\theta_{c,i}=\theta_{s,j}=1$ for $i\in\mathcal{I}$ and $j\in\mathcal{J}$. The routing probabilities are $p_{ij}=1/10$ for all $i\in\mathcal{I}$ and $j\in\mathcal{J}$. The total amount of resource is $m=400$ and the conversion rates are $r_c=1$ and $r_{s,j}=1$ for $j\in\mathcal{J}$. The respective holding and abandonment costs are $w_s^H=w_s^A=(0.5,0.75,1,1.25,1.5,1.75,2,2.25,2.5,5)$, and $w_c^H=w_c^A=\gamma w_s^H$ with γ ranging from 0.5 to 2.5 in increments of 0.1. For the GP system, stage-1 classes form at most three service groups and are subsequently attended by at most two stations in stage 2. We also construct a one-to-one system by further combining all stage-1 classes into a single queue and pooling all stage-2 servers into one station. Customers within the same queue are served following the FCFS discipline. From the comparisons in Figure 6(a), the average cost incurred under the GP system (blue line) is almost indistinguishable from that under the original system (black line); that is, a well-designed scheme of grouping and pooling can achieve comparable performance to the original one in terms of simulated cost. Moreover, the GP system significantly outperforms the corresponding one-to-one system (green line), which suggests that simply grouping all customer classes to stage 1 and pooling all stations in stage 2 to construct a one-to-one system compromises system performance.

The cost efficiency in the GP system is attributed to the combined effect of grouping and pooling. In what follows, we investigate their effect separately using many-to-one and one-to-many systems, respectively. For the many-to-one system, we keep its stage-1 parameters the same as those in the general system, while setting the transition probability as $p_{is}=1$ and the stage-2 parameters as follows: $\mu_s=1$, $\theta_s=1$, $r_s=1$ and $w_s^H=w_s^A=1.85$; for the one-to-many system, we keep the stage-2 parameters the same as those in the preceding general system, while setting the transition probability as $p_{cj}=1/10$ and the stage-1 parameters as follows: $\lambda=400$, $\mu_c=1$, $\theta_c=1$, $r_c=1$ and $w_c^H=w_c^A=\gamma$ with γ ranging from 0.5 to 2.5. For these two systems, we repeat the grouping and pooling process to construct the corresponding grouped system and pooled system, respectively.

The results are plotted in Figures 6(b) and (c). From panel (b), the grouped system incurs the long-run average cost (blue line) that is very close to that in the original system (black line). This is because, for the grouped system, $\mathcal{I}_1$ consists of the classes whose service requirements can be

Figure 6 The Value of Optimal Grouping and Pooling



fully satisfied. Therefore, holding customers from those classes wait in a single line and following FCFS service discipline does not have much impact on their service fulfillment in the long run. This prevents the queue dedicated to $\mathcal{I}_1$ from building up, resulting in a zero long-run average cost. On the other hand, the classes in $\mathcal{I}_3$ receive almost no service capacity in the grouped system (because the service capacity is mainly used for serving customers from the other groups), similar to the original system in which the least priority is assigned to them. Hence, nearly all customers from these classes eventually abandon the queue, incurring similar abandonment penalties in both systems. For the class (if there exists one) in $\mathcal{I}_2$, its steady-state queue length remains unchanged as it receives the same amount of service capacity after grouping, resulting in similar delay-related cost. Analogously, panel (c) reveals that the pooled system incurs long-run average cost that is comparable to the original system. This can be explained using some similar argument. To be specific, the aggregate resources assigned to service types belonging to $\mathcal{J}_1$ (resp., $\mathcal{J}_3$) remain unchanged after pooling, with their service requirements being fully (resp., never) accommodated. For the service type (if there exists one) in $\mathcal{J}_2$, it receives the same amount of resource and thus its long-run average cost is not affected after pooling. Moreover, panel (c) shows that the pooled system (red line) slightly outperforms the original system (black line). This is due to the *pooling effect*: having servers attending to multiple service types allows for capacity sharing, and thus results in better cost efficiency.

6. Conclusion

In this paper, we consider a joint capacity allocation and scheduling problem in a two-stage many-server queueing system with multiple customer classes. By solving the corresponding fluid optimization problem, we propose a capacity allocation and (static priority) scheduling policy, which

is shown to be asymptotically optimal when the system size grows large. We obtain clean priority rules for two special system structures. We also discuss how to combine classes and stations that streamlines the system operations while not worsening the system performance much. Finally, we run numerical experiments to validate the accuracy of the fluid approximation, and evaluate performance of the heuristic for grouping customers classes and pooling stations to form clusters.

Our work suggests several potential directions for future research. First, one may consider diffusion approximation, which is well known to be more accurate than fluid approximation and is able to capture the error terms of order $\mathcal{O}(\sqrt{m})$ due to stochastic variability. However, according to the results in Bassamboo and Randhawa (2010) and Bassamboo et al. (2023), if the system is overloaded (as for our model), it is expected that the fluid-based solution is within $\mathcal{O}(1)$ optimality. It is worthwhile to justify this for our setting. Second, a rigorous analysis for the GP system is called for. Note that in the GP system, a customer's service time and patience time are dependent through the customer's class, which is unknown to the system manager. As a result, the traditional fluid limit results (see e.g., Whitt 2006), which typically use the independence of service times and patience times, cannot be directly applied. Finally, one may also consider general queueing network structures, other control forms (e.g., routing), as well as other features (e.g., time-varying arrivals and non-exponential service times or patience times).

References

Armony, M., C. W. Chan, and B. Zhu (2018). Critical care capacity management: Understanding the role of a step down unit. *Production and Operations Management 27*(5), 859–883.

Armony, M. and A. Mandelbaum (2011). Routing and staffing in large-scale service systems: The case of homogeneous impatient customers and heterogeneous servers. *Operations Research 59*(1), 50–65.

Atar, R., C. Giat, and N. Shimkim (2010). The $c\mu/\theta$ rule for many-server queues with abandonment. *Operations Research 58*(5), 1427–1439.

Atar, R., C. Giat, and N. Shimkim (2011). On the asymptotic optimality of the $c\mu/\theta$ rule under ergodic cost. *Queueing Systems 67*(2), 127–144.

Baron, O. and J. Milner (2009). Staffing to maximize profit for call centers with alternate service-level agreements. *Manufacturing & Service Operations Management 57*(3), 685–700.

Bassamboo, A. and R. S. Randhawa (2010). On the accuracy of fluid models for capacity sizing in queueing systems with impatient customers. *Operations Research 58*(5), 1398–1413.

Bassamboo, A., R. Randhawa, and C. A. Wu (2023). Optimally scheduling heterogeneous impatient customers. *Manufacturing & Service Operations Management* Forthcoming.

Baumanna, H. and W. Sandmann (2017). Multi-server tandem queue with Markovian arrival process, phase-type service times, and finite buffers. *European Journal of Operational Research 256*(1), 187–195.

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

Borst, S., A. Mandelbaum, and M. I. Reiman (2004). Dimensioning large call centers. *Operations Research 52*(1), 17–34.

Chao, X., L. Liu, and S. Zheng (2003). Resource allocation in multisite service systems with intersite customer flows. *Management Science 49*(12), 1739–1752.

Cox, D. and W. Smith (1961). *Queues*. Methuen, London; Wiley, New York.

Gans, N., G. Koole, and A. Mandelbaum (2003). Telephone call centers: Tutorial, review and research prospects. *Manufacturing and Service Operations Management 5*(2), 79–141.

Garnett, O., A. Mandelbaum, and M. Reiman (2002). Desining a call center with impatient customers. *Manufacturing and Service Operations Management 3*(4), 208–227.

Green, L. V. (2010). Using queueing theory to alleviate emergency department overcrowding. *Wiley Encyclopedia of Operations Research and Management Science*.

Gurvich, I. and W. Whitt (2009). Scheduling flexible servers with convex delay costs in many-server service systems. *Manufacturing & Service Operations Management 11*(2), 237–253.

Gurvich, I. and W. Whitt (2010). Service-level differentiation in many-server service systems via queue-ratio routing. *Operations Research 58*(2), 316–328.

Harrison, J. M. (1978). The diffusion approximation for tandem queues in heavy traffic. *Advances in Applied Probability 10*(4), 886–905.

Harrison, J. M. and L. Shepp (1984). A tandem storage system and its diffusion limit. *Stochastic Processes and their Applications 16*(3), 257–274.

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

Hu, Y., C. W. Chan, and J. Dong (2022). Optimal scheduling of proactive service with customer deterioration and improvement. *Management Science 68*(4), 2533–2578.

Kim, J., R. S. Randhawa, and A. R. Ward (2018). Dynamic scheduling in a many-server, multiclass system: The role of customer impatience in large systems. *Manufacturing & Service Operations Management 20*(2), 285–301.

Koçağa, Y. L., M. Armony, and A. R. Ward (2014). Staffing call centers with uncertain arrival rates and co-sourcing. *Production and Operations Management 7*(24), 1101–1117.

Kostami, V. and A. R. Ward (2009). Managing service systems with an offline waiting option and customer abandonment. *Manufacturing & Service Operations Management 4*(11), 644–656.

Long, Z., N. Shimkin, H. Zhang, and J. Zhang (2020). Dynamic scheduling of multiclass many-server queues with abandonment: The generalized $c\mu/h$ rule. *Operation Research 68*(4), 1218–1230.

Mandelbaum, A. and A. L. Stolyar (2004). Scheduling flexible servers with convex delay costs: Heavy-traffic optimality of the generalized $c\mu$-rule. *Operations Reserch 52*(6), 836–855.

Mandelbaum, A. and S. Zeltyn (2009). Staffing many-server queues with impatient customers: Constraint satisfaction in call centers. *Operations Research 57*(5), 1189–1205.

Momčilović, P., A. Mandelbaum, N. Carmeli, M. Armony, and G. Yom-Tov (2022). Resource-driven Activity-Networks (RANs): A modelling framework for complex operations. *Submitted*.

Rosberg, Z., P. Varaiya, and J. Walrand (1982). Optimal control of service in tandem queues. *IEEE Transactions on Automatic Control 27*(3), 600–610.

Sheu, R.-S. and I. Ziedins (2010). Asymptotically optimal control of parallel tandem queues with loss. *Queueing Systems 65*, 211–227.

Shone, R., K. Glazebrook, and K. G. Zografos (2019). Resource allocation in congested queueing systems with time-varying demand: An application to airport operations. *European Journal of Operational Research 276*(2), 566–581.

van Mieghem, J. A. (1995). Dynamic scheduling with convex delay costs: The generalized $c\mu$ rule. *Annals of Applied Probability 5*(3), 809–833.

Wang, J., H. Abouee-Mehrizi, O. Baron, and O. Berman (2019a). Tandem queues with impatient customers. *Performance Evaluation 135*, 102011.

Wang, X., S. Andradottir, and H. Ayhan (2019b). Optimal pricing for tandem queues with finite buffers. *Queueing Systems 92*, 323–396.

Whitt, W. (2004). Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Science 50*(10), 1449–1461.

Whitt, W. (2006). Fluid models for multiserver queues with abandonments. *Operations Research 54*(1), 27–54.

Zychlinski, N., C. W. Chan, and J. Dong (2022). Managing queues with different resource requirements. *Operation Research*, Forthcoming.

Zychlinski, N., A. Mandelbaum, P. Momčilović, and I. Cohen (2020). Bed blocking in hospitals due to scarce capacity in geriatric institutions – cost minimization via fluid models. *Manufacturing & Service Operations Management 22*(2), 396–411.

Zychlinski, N., P. Momčilović, and A. Mandelbaum (2018). Time-varying many-server finite-queues in tandem: Comparing blocking mechanisms via fluid models. *Operations Research Letters 46*(5), 492–499.

Appendix for "Capacity Allocation and Scheduling in Two-Stage Service Systems with Multi-Class Customers"

EC.1. Proof of the Results in Section 3

EC.1.1. Proof of Lemma 1

We first prove the "if" part. Suppose $\sum_{i\in\mathcal{I}}\lambda_{c,i}\left(\frac{1}{r_c\mu_{c,i}}+\sum_{j\in\mathcal{J}}\frac{p_{ij}}{r_{s,j}\mu_{s,j}}\right)\le 1$, then let (z',q',n') be

$$z'_{c,i}=\frac{\lambda_{c,i}}{\mu_{c,i}},\quad z'_{s,j}=\frac{1}{\mu_{s,j}}\sum_{i\in\mathcal{I}}\lambda_{c,i}p_{ij},\quad n'_c=\sum_{i\in\mathcal{I}}z'_{c,i},\quad n'_{s,j}=z'_{s,j},$$
$$q'_{c,i}=\frac{\lambda_{c,i}-\mu_{c,i}z'_{c,i}}{\theta_{c,i}}\quad\text{and}\quad q'_{s,j}=\frac{\sum_{i\in\mathcal{I}}\mu_{c,i}z'_{c,i}p_{ij}-\mu_{s,j}z'_{s,j}}{\theta_{s,j}}.$$

It is easy to verify that $q'_{c,i}=0$ for all i and $q'_{s,j}=0$ for all j. In addition, one can verify that $(z',q',n')\in\Xi(\lambda_c)$. The objective function at (z',q',n') is 0. As a result, $\bar{J}^*=0$.

Next, we prove the "only if" part. Suppose that an optimal solution is (z^*,q^*,n^*). Then we have $q^*_{c,i}=0$ for all $i\in\mathcal{I}$ and $q^*_{s,j}=0$ for all $j\in\mathcal{J}$ as $\bar{J}^*=0$. (Recall that we have $w_{c,i}>0$ for $i\in\mathcal{I}$ and $w_{s,j}>0$ for $j\in\mathcal{J}$.) Then from (23)–(24), we have

$$z^*_{c,i}=\frac{\lambda_{c,i}}{\mu_{c,i}}\quad\text{and}\quad z^*_{s,j}=\frac{1}{\mu_{s,j}}\sum_{i\in\mathcal{I}}\lambda_{c,i}p_{ij}.$$

By (25), (26) and (28),

$$\frac{1}{r_c}\sum_{i\in\mathcal{I}}z^*_{c,i}+\sum_{j\in\mathcal{J}}\frac{z^*_{s,j}}{r_{s,j}}\le 1.$$

Therefore, $\sum_{i\in\mathcal{I}}\lambda_{c,i}(\frac{1}{r_c\mu_{c,i}}+\sum_{j\in\mathcal{J}}\frac{p_{ij}}{r_{s,j}\mu_{s,j}})\le 1$. The proof is hence complete. □

EC.1.2. Proof of Theorem 1

Proof of Part (i). We follow the idea in the proof of Proposition 2.1 in Atar et al. (2011). Let

$$U^m:=\liminf_{t\to\infty}\frac{1}{t}\int_0^t\left(\sum_{i\in\mathcal{I}}w_{c,i}Q^m_{c,i}(s)+\sum_{j\in\mathcal{J}}w_{s,j}Q^m_{s,j}(s)\right)ds.$$

By Fatou's lemma, it suffices to show that

$$\liminf_{m\to\infty}\frac{1}{m}U^m\ge\bar{J}^*,\quad\text{almost surely.}\tag{EC.1}$$

Fix one sample point. Let $\{t_k;k\in\mathbb{N}\}$ be a sequence of real values increasing to infinity, such that

$$U^m=\lim_{k\to\infty}\frac{1}{t_k}\int_0^{t_k}\left(\sum_{i\in\mathcal{I}}w_{c,i}Q^m_{c,i}(s)+\sum_{j\in\mathcal{J}}w_{s,j}Q^m_{s,j}(s)\right)ds.$$

By (2), (9) and (12), the random vector $\int_0^{t_k} Z^m(s)ds/t_k$ is uniformly bounded. Hence, we can choose a subsequence of $\{t_k; k\in\mathbb{N}\}$, still denoted by $\{t_k; k\in\mathbb{N}\}$ without loss of generality, along which random vectors $\int_0^{t_k} Z^m(s)ds/t_k$ converge. Denote the corresponding limiting random vector as $\widehat{Z}^m := \lim_{k\to\infty}\int_0^{t_k} Z^m(s)ds/t_k$.

We use the following lemma, whose proof is deferred to the end of this section.

LEMMA EC.1. *As $t\to\infty$, $X^m(t)/t\to 0$ almost surely.*

Dividing both sides of (5) by t_k, taking $t_k\to\infty$, using Lemma EC.1 and following the same argument as in Atar et al. (2011), we can claim that $\widehat{Q}_c^m := \lim_{t_k\to\infty}\int_0^{t_k} Q_c^m(s)ds/t_k$ exists and

$$0 = \lambda_{c,i}^m - \mu_{c,i}\widehat{Z}_{c,i}^m - \theta_{c,i}\widehat{Q}_{c,i}^m, \quad \text{for } i\in\mathcal{I}. \tag{EC.2}$$

Similarly, from (10), we can get the existence of $\widehat{Q}_s^m := \lim_{t_k\to\infty}\int_0^{t_k} Q_s^m(s)ds/t_k$ and

$$0 = \sum_{i\in\mathcal{I}} p_{ij}\mu_{c,i}\widehat{Z}_{c,i}^m - \mu_{s,j}\widehat{Z}_{s,j}^m - \theta_{s,j}\widehat{Q}_{s,j}^m, \quad \text{for } j\in\mathcal{J}. \tag{EC.3}$$

From (20) to (22), we have

$$\widehat{Q}^m\ge 0, \quad \widehat{Z}^m\ge 0, \quad \sum_{i\in\mathcal{I}}\widehat{Z}_{c,i}^m \le N_c^m, \text{ and } \widehat{Z}_{s,j}^m\le N_{s,j}^m \text{ for } j\in\mathcal{J}. \tag{EC.4}$$

Recalling the definition of feasible region $\Xi(\lambda_c)$ defined below (28), we have

$$\left(\widehat{Z}^m/m, \widehat{Q}^m/m, N^m/m\right)\in\Xi(\lambda_c^m/m)$$

by (12) and (EC.2)–(EC.4). Therefore, we have

$$\begin{aligned}\frac{U^m}{m} &= \sum_{i\in\mathcal{I}} w_{c,i}\frac{\widehat{Q}_{c,i}^m}{m} + \sum_{j\in\mathcal{J}} w_{s,j}\frac{\widehat{Q}_{s,j}^m}{m}\\ &\ge \min_{(z,q,n)\in\Xi(\lambda_c^m/m)} \sum_{i\in\mathcal{I}} w_{c,i}q_{c,i} + \sum_{j\in\mathcal{J}} w_{s,j}q_{s,j}.\end{aligned}$$

From the continuity of the LP problem (29) in its feasible region and (17), we then have $\liminf_{m\to\infty} U^m/m\ge \bar{J}^*$, establishing (EC.1). □

Proof of Part (ii). We use a similar idea as in the proof of Theorem 2.2 in Atar et al. (2011). From Condition (32), there exists $T_m>0$, such that $\mathbb{E}[\|m^{-1}Q^{m,*}(t)-q^*\|]\le\delta_m$ for $t\ge T_m$. Let $\bar{J}^* := \sum_{i\in\mathcal{I}} w_{c,i}q_{c,i}^* + \sum_{j\in\mathcal{J}} w_{s,j}q_{s,j}^*$. Then

$$\mathbb{E}\int_0^T \Big(\sum_{i\in\mathcal{I}} w_{c,i}Q_{c,i}^{m,*}(t)/m + \sum_{j\in\mathcal{J}} w_{s,j}Q_{s,j}^{m,*}(t)/m\Big)dt - T\cdot\bar{J}^*$$

$$\leq \big(\max_{i\in\mathcal{I}} w_{c,i} \vee \max_{j\in\mathcal{J}} w_{s,j}\big)\mathbb{E}\Big[\int_0^T \|Q^{m,*}(t)/m - q^*\|dt\Big]$$
$$\leq \big(\max_{i\in\mathcal{I}} w_{c,i} \vee \max_{j\in\mathcal{J}} w_{s,j}\big)\Big[\int_0^{T_m} \mathbb{E}[\|m^{-1}Q^{m,*}(t) - q^*\|]dt + \delta_m(T - T_m)\Big].$$

Hence, dividing both sides by T and then letting $T \to \infty$, we obtain that

$$\frac{J^m(\psi^{m,*})}{m} \leq \bar{J}^* + \big(\max_{i\in\mathcal{I}} w_{c,i} \vee \max_{j\in\mathcal{J}} w_{s,j}\big)\delta_m.$$

As a result, $\limsup_{m\to\infty} J^m(\psi^{m,*})/m \leq \bar{J}^*$. This, together with the inequality established in part (i), concludes the desired equality. □

Proof of Lemma EC.1. The proof adopts the idea in Appendix A.2 of Atar et al. (2011). We will show that for each $i \in \mathcal{I}$ and $j \in \mathcal{J}$, $X^m_{c,i}(t)/t \to 0$ and $X^m_{s,j}(t)/t \to 0$ as $t \to \infty$ almost surely.

By (5), we have

$$X^m_{c,i}(t) = X^m_{c,i}(0) + A_{c,i}(\lambda^m_{c,i}t) - S_{c,i}\left(\mu_{c,i}\int_0^t Z^m_{c,i}(s)ds\right) - G_{c,i}\left(\theta_{c,i}\int_0^t Q^m_{c,i}(s)ds\right).$$

Note that $X^m_{c,i}(t) = Z^m_{c,i}(t) + Q^m_{c,i}(t)$. Hence, by standard coupling argument (see e.g., the proof of Lemma 3 in Dong et al. (2015)), we can show that the process $X^m_{c,i}$ is stochastically dominated (upper bounded) by the number-in-system (i.e., the total number of customer in the system) process of an $M/M/\infty$ system, in which the arrival rate is $\lambda^m_{c,i}$, the service rate is $\mu_{c,i} \wedge \theta_{c,i}$, and initial state is $X^m_{c,i}(0)$.

Similarly, by (10), we have

$$X^m_{s,j}(t) = X^m_{s,j}(0) + \sum_{i\in\mathcal{I}} E^m_{ij}(t) - S_{s,j}\left(\mu_{s,j}\int_0^t Z^m_{s,j}(s)ds\right) - G_{s,j}\left(\theta_{s,j}\int_0^t Q^m_{s,j}(s)ds\right).$$

Note that $\sum_{i\in\mathcal{I}} E^m_{ij}(t) \leq \sum_{i\in\mathcal{I}} \big(X^m_{c,i}(0) + A_{c,i}(\lambda^m_{c,i}t)\big)$. Hence, again by using coupling method, we can show that the process $X^m_{s,j}$ is stochastically dominated by the number-in-system process of an $M/M/\infty$ system, in which the arrival rate is $\sum_{i\in\mathcal{I}} \lambda^m_{c,i}$, the service rate of each server is $\mu_{s,j} \wedge \theta_{s,j}$, and initial state is $X^m_{s,j}(0)$.

According to the above two upper bound results, the desired result follows if we can show that for any $M/M/\infty$ system, the corresponding number-in-system process, denoted by Y, satisfies $Y(t)/t \to 0$ almost surely as $t \to \infty$. In fact, this claim has been shown in Appendix A.2 of Atar et al. (2011), for a more general $G/M/\infty$ system. Their argument is based on considering the system's embedded Markov chain and using a second moment estimate of Y.

For the sake of completeness, below we provide an alternative proof. Suppose the $M/M/\infty$ system has an arrival rate λ, service rate μ, and an initial state $Y(0)$. Because we deal with almost sure convergence, we can assume the value of $Y(0)$ is fixed. Note that $Y(t)$ has two independent sources: one is from the newly arriving customers and the other from those initially present in the system. Then, for all $t \geq 0$, $Y(t) - Y(0)$ is bounded by $\tilde{Y}(t)$, the new arrivals who are still in the systems, which is a Poisson distributed random variable with parameter $\lambda(1-e^{-\mu t})/\mu$ (Eick et al. 1993). For any given $\Delta > 0$, let $\Gamma_k := \max_{t\in[k\Delta,(k+1)\Delta)} Y(t)/t$, $k \in \mathbb{Z}_+$. In order to establish that $Y(t)/t \to 0$ almost surely as $t \to \infty$, it suffices to show that $\lim_{k\to\infty} \Gamma_k = 0$ almost surely. Note that for any given $\epsilon > 0$, we have

$$\begin{aligned}\sum_{k=1}^{\ell} \mathbb{P}(\Gamma_k > \epsilon) &\leq \sum_{k=1}^{\ell} \mathbb{P}\Big(\max_{t\in[k\Delta,(k+1)\Delta)} Y(t) - Y(0) > \epsilon\cdot k\Delta - Y(0)\Big) \\ &\leq \sum_{k=1}^{\ell} \mathbb{P}\big(\tilde{Y}(k\Delta) + Z_k > (\epsilon\cdot k\Delta - Y(0))^+\big),\end{aligned} \tag{EC.5}$$

for sufficiently large ℓ, where Z_k denotes the number of arrivals during time interval $(k\Delta,(k+1)\Delta)$, which is Poisson distributed with parameter $\lambda\Delta$. Moreover, since $\tilde{Y}(k\Delta)$ relies on the system evolution before period k while Z_k encodes information about customer arrivals in period k, they are independent from each other. Hence, $\tilde{Y}(k\Delta) + Z_k$ is also Poisson distributed with parameter $\lambda(1-e^{-\mu k\Delta})/\mu + \lambda\Delta$. By the Chebychef's inequality, for any $\epsilon > 0$ and sufficiently small Δ,

$$\begin{aligned}&\sum_{k=0}^{\infty} \mathbb{P}\left(\tilde{Y}(k\Delta) + Z_k > (\epsilon\cdot k\Delta - Y(0))^+\right) \\ &= \sum_{k=1}^{\lfloor Y(0)/(\epsilon\cdot\Delta)\rfloor} \mathbb{P}\left(\tilde{Y}(k\Delta) + Z_k > 0\right) + \sum_{k=\lceil Y(0)/(\epsilon\cdot\Delta)\rceil}^{\infty} \mathbb{P}\left(\tilde{Y}(k\Delta) + Z_k > \epsilon\cdot k\Delta - Y(0)\right) \\ &\leq \left\lfloor \frac{Y(0)}{\epsilon\cdot\Delta} \right\rfloor + \sum_{k=\lceil Y(0)/(\epsilon\cdot\Delta)\rceil}^{\infty} \frac{(\lambda(1-e^{-\mu k\Delta})/\mu + \lambda\Delta)(\lambda(1-e^{-\mu k\Delta})/\mu + \lambda\Delta + 1)}{(\epsilon\cdot k\Delta - Y(0))^2} \\ &\leq \left\lfloor \frac{Y(0)}{\epsilon\cdot\Delta} \right\rfloor + \left(\frac{\lambda}{\mu} + \lambda\Delta\right)\left(\frac{\lambda}{\mu} + \lambda\Delta + 1\right)\sum_{k=1}^{\infty} \frac{C}{k^2} < \infty,\end{aligned}$$

for an appropriate constant $C > 0$. Hence, by the Borel–Cantelli lemma, we have $Y(t)/t \to 0$ (a.s.) as $t \to \infty$. This completes the proof. □

EC.2. Proof of the Results in Section 4

EC.2.1. Proof of Lemma 4

(i) From the expressions of $\xi_{i,\phi}$ and $\xi_{i,\phi+1}$, one can verify that

$$\xi_{i,\phi} - \xi_{i,\phi+1} = -\left(\frac{w_{s,\phi+1} r_{s,\phi+1} \mu_{s,\phi+1}}{\theta_{s,\phi+1}} - \frac{w_{s,\phi+2} r_{s,\phi+2} \mu_{s,\phi+2}}{\theta_{s,\phi+2}}\right) \chi_{i,\phi+1} \leq 0, \tag{EC.6}$$

where the equality uses $\chi_{i,\phi+1} - \chi_{i,\phi} = \frac{p_{i,\phi+1}}{r_{s,\phi+1}\mu_{s,\phi+1}}$, and the inequality is due to $\frac{w_{s,\phi+1}r_{s,\phi+1}\mu_{s,\phi+1}}{\theta_{s,\phi+1}} \geq \frac{w_{s,\phi+2}r_{s,\phi+2}\mu_{s,\phi+2}}{\theta_{s,\phi+2}}$ according to the ordering in (36).

(ii) If $i \in \mathcal{I}_+(\phi)$, then from part (i), $\xi_{i,\phi+1} \geq \xi_{i,\phi} \geq 0$, hence $i \in \mathcal{I}_+(\phi+1)$. □

EC.2.2. Solutions to Sub-problems (43)

Here we establish Proposition EC.1, which solves the sub-problems (43). Recall that for each $\phi \in [\mathcal{J}]$, we separate $\mathcal{I}$ into two disjoint subsets: $\mathcal{I}_+(\phi) = \{i \in \mathcal{I} : \xi_{i,\phi} \geq 0\}$ and $\mathcal{I}_-(\phi) = \{i \in \mathcal{I} : \xi_{i,\phi} < 0\}$, and denote by $i_+(\phi) = |\mathcal{I}_+(\phi)|$. To facilitate the presentation, we arrange the indices in $\mathcal{I}$ such that

$$\frac{\xi_{1,\phi}}{\chi_{1,\phi}} \geq \cdots \geq \frac{\xi_{i_+(\phi),\phi}}{\chi_{i_+(\phi),\phi}} \geq 0 > \frac{\xi_{i_+(\phi)+1,\phi}}{\chi_{i_+(\phi)+1,\phi+1}} \geq \cdots \geq \frac{\xi_{I,\phi}}{\chi_{I,\phi+1}} \tag{EC.7}$$

In particular, for the case when $\phi = J$, because $\xi_{i,J} \geq 0$ for all i, we have $\mathcal{I}_+(J) = \mathcal{I}$ and thus arrange the indices in $\mathcal{I}$ such that $\xi_{1,J}/\chi_{1,J} \geq \cdots \geq \xi_{I,J}/\chi_{I,J}$.

PROPOSITION EC.1. *Recall the sets $\mathcal{S}_i$ with $i = 0,1,2,3$ defined in (44), and fix $\phi \in [\mathcal{J}]$. If $\phi \in \mathcal{S}_0$, there is no feasible solution to problem (43). Otherwise, the optimal solution to Problem (43) exists, and is characterized as follows.*

(i) *If $\phi \in \mathcal{S}_1$, then there exists $i_1 := \min\{i' \in \mathcal{I} : \sum_{i \leq i'} \lambda_{c,i}\chi_{i,\phi+1} \geq 1\} > i_+(\phi)$, such that $\varphi^*_{c,i}(\phi) = \lambda_{c,i}$ for $i < i_1$ and $\varphi^*_{c,i}(\phi) = 0$ for $i > i_1$, with $\varphi^*_{c,i_1}(\phi)$ being set so that the second constraint in the set $\bar{\Pi}_\phi$ is binding, i.e., $\sum_{i \in \mathcal{I}} \chi_{i,\phi+1}\varphi^*_{c,i}(\phi) = 1$, or equivalently, $\varphi^*_{c,i_1}(\phi) = (1 - \sum_{i<i_1} \lambda_{c,i}\chi_{i,\phi+1})/\chi_{i_1,\phi+1}$.*

(ii) *If $\phi \in \mathcal{S}_2$, then $\varphi^*_{c,i}(\phi) = \lambda_{c,i}$ for $i \in \mathcal{I}_+(\phi)$, and $\varphi^*_{c,i}(\phi) = 0$ for $i \in \mathcal{I}_-(\phi)$.*

(iii) *If $\phi \in \mathcal{S}_3$, then there exists $i_2 := \min\{i' \in \mathcal{I}_+(\phi) : \sum_{i \leq i'} \lambda_{c,i}\chi_{i,\phi} \geq 1\} \leq i_+(\phi)$, such that $\varphi^*_{c,i}(\phi) = \lambda_{c,i}$ for $i < i_2$, $\varphi^*_{c,i}(\phi) = 0$ for $i > i_2$, and with $\varphi^*_{c,i_2}(\phi)$ being set so that the first constraint in the set $\bar{\Pi}_\phi$ is binding, i.e., $\sum_{i \in \mathcal{I}} \chi_{i,\phi}\varphi^*_{c,i}(\phi) = 1$, or equivalently, $\varphi^*_{c,i_2}(\phi) = (1 - \sum_{i<i_2} \lambda_{c,i}\chi_{i,\phi})/\chi_{i_2,\phi}$.*

Proof of Proposition EC.1. First, if $\phi \in \mathcal{S}_0$, then $\phi < J$ and the second constraint in $\bar{\Pi}_\phi$ cannot hold, thus problem (43) has no feasible solution. Hence, in the following, we consider $\phi \in \mathcal{S}_0^c$, that is, $\sum_{i \in \mathcal{I}} \lambda_{c,i}\chi_{i,\phi+1} \geq 1$. We first consider $\phi < J$.

1. When $\phi \in \mathcal{S}_1$, that is, $\sum_{i \in \mathcal{I}_+(\phi)} \lambda_{c,i}\chi_{i,\phi+1} < 1$, then one must have $\varphi^*_{c,i}(\phi) > 0$ for some $i \in \mathcal{I}_-(\phi)$ to ensure that the second constraint in $\bar{\Pi}_\phi$ holds. Moreover, the second constraint in $\bar{\Pi}_\phi$ must be binding; otherwise, one can always decrease $\varphi^*_{c,i}(\phi) > 0$ for $i \in \mathcal{I}_-(\phi)$ to get a larger

objective value. Also note that $\chi_{i,\phi} \le \chi_{i,\phi+1}$ and recall that $J \in \mathcal{S}_3$ which means $\phi < J$. Hence, problem (43) is reduced to

$$\max_{\varphi_c} \sum_{i\in\mathcal{I}} \xi_{i,\phi}\varphi_{c,i} + \frac{w_{s,\phi+1}r_{s,\phi+1}\mu_{s,\phi+1}}{\theta_{s,\phi+1}},$$
$$\text{s.t.} \sum_{i\in\mathcal{I}} \chi_{i,\phi+1}\varphi_{c,i} = 1, \text{ and } 0 \le \varphi_{c,i} \le \lambda_{c,i}, \text{ for } i \in \mathcal{I}.$$

Therefore, we arrange indices in $\mathcal{I}$ such that $\xi_{1,\phi}/\chi_{1,\phi+1} \ge \cdots \ge \xi_{I,\phi}/\chi_{I,\phi+1}$. Then with $i_1 := \min\{i' \in \mathcal{I} : \sum_{i\le i'} \lambda_{c,i}\chi_{i,\phi+1} \ge 1\} > i_+(\phi)$, the optimal solution is given as $\varphi^*_{c,i}(\phi) = \lambda_{c,i}$ for $i \le i_1 - 1$, $\varphi^*_{c,i}(\phi) = 0$ for $i > i_1$, with $\varphi^*_{c,i_1}(\phi)$ being set so that the constraint of the above problem being binding. This gives Case (i) of Proposition EC.1.

2. When $\phi \notin \mathcal{S}_1$, that is, $\sum_{i\in\mathcal{I}_+(\phi)} \lambda_{c,i}\chi_{i,\phi+1} \ge 1$, then we have the following two cases:

   (i) If $\phi \in \mathcal{S}_2$, that is, $\sum_{i\in\mathcal{I}_+(\phi)} \lambda_{c,i}\chi_{i,\phi} \le 1$, then the proposed solution is a feasible solution because $\sum_{i\in\mathcal{I}} \chi_{i,\phi}\varphi^*_{c,i}(\phi) = \sum_{i\in\mathcal{I}_+(\phi)} \chi_{i,\phi}\lambda_{c,i} \le 1$ and $\sum_{i\in\mathcal{I}} \chi_{i,\phi+1}\varphi^*_{c,i}(\phi) = \sum_{i\in\mathcal{I}_+(\phi)} \chi_{i,\phi+1}\lambda_{c,i} \ge 1$. By the definition of $\Pi_+(\phi)$ and $\Pi_-(\phi)$, one can also verify that the value of any other feasible solution cannot be larger than that of the proposed solution. This gives Case (ii) of Proposition EC.1.

   (ii) If $\phi \in \mathcal{S}_3$, that is, $\sum_{i\in\mathcal{I}_+(\phi)} \lambda_{c,i}\chi_{i,\phi} > 1$, then when $\phi < J$ one must have $\varphi^*_{c,i}(\phi) = 0$ for $i \in \mathcal{I}_-(\phi)$. This is because, if $\varphi^*_{c,i}(\phi) > 0$ for some $i \in \mathcal{I}_-(\phi)$, then one can always decrease such $\varphi^*_{c,i}(\phi)$ to get a larger objective value (if violating the second constraint, then one can increase $\varphi^*_{c,i}(\phi)$ for some $i \in \mathcal{I}_+(\phi)$). Moreover, the first constraint in $\bar{\Pi}_\phi$ must be binding; otherwise, if $\sum_{i\in\mathcal{I}_+(\phi)} \chi_{i,\phi}\varphi^*_{c,i}(\phi) < 1$, then there must be $\varphi^*_{c,i}(\phi) < \lambda_{c,i}$ for some $i \in \mathcal{I}_+(\phi)$. Increasing such $\varphi^*_{c,i}(\phi)$ will lead to a larger objective value. Then the second constraint in $\bar{\Pi}_\phi$ holds because $\chi_{i,\phi+1} \ge \chi_{i,\phi}$. Therefore, problem (43) is equivalent to

$$\max_{\varphi_c} \sum_{i\in\mathcal{I}_+(\phi)} \xi_{i,\phi}\varphi_{c,i} + \frac{w_{s,\phi+1}r_{s,\phi+1}\mu_{s,\phi+1}}{\theta_{s,\phi+1}},$$
$$\text{s.t.} \sum_{i\in\mathcal{I}_+(\phi)} \chi_{i,\phi}\varphi_{c,i} = 1, \text{ and } 0 \le \varphi_{c,i} \le \lambda_{c,i}, \text{ for } i \in \mathcal{I}_+(\phi).$$

   Therefore, we arrange indices in $\mathcal{I}_+(\phi)$ such that $\xi_{1,\phi}/\chi_{1,\phi} \ge \cdots \ge \xi_{i_+(\phi),\phi}/\chi_{i_+(\phi),\phi}$. Then with $i_2 := \min\{i' \in \mathcal{I}_+(\phi) : \sum_{i\le i'} \lambda_{c,i}\chi_{i,\phi} \ge 1\} \le i_+(\phi)$, the optimal solution is given as $\varphi^*_{c,i}(\phi) = \lambda_{c,i}$ for $i \le i_2 - 1$, $\varphi^*_{c,i}(\phi) = 0$ for $i > i_2$, with $\varphi^*_{c,i_2}(\phi)$ being set so that the constraint of the above problem being binding. This gives Case (iii) of Proposition EC.1.

Finally, when $\phi = J \in \mathcal{S}_3$, $\mathcal{I}_+(J) = [\mathcal{J}]$ and $\mathcal{I}_-(J) = \emptyset$. Then problem (43) becomes

$$\max_{\varphi_c \in \bar{\Pi}_J} \sum_{i\in\mathcal{I}} \varphi_{c,i}\xi_{i,J}. \tag{EC.8}$$

This is a bin packing problem. Therefore, with $i_2 := \min\{i' \in [\mathcal{J}] : \sum_{i \le i'} \lambda_{c,i}\chi_{i,\phi} \ge 1\}$, the optimal solution is given by

$$\varphi_c^*(\phi) = \left(\lambda_{c,1}, \ldots, \lambda_{c,i_2-1}, \frac{1}{\chi_{i_2,J}}\left(1 - \sum_{i<i_2} \lambda_{c,i}\chi_{i,J}\right), 0, \ldots, 0\right).$$

This is consistent with Case (iii) of Proposition EC.1. This completes the proof. □

EC.2.3. Proof of Proposition 1

(i) This can be proved by noting that $\chi_{i,\phi}$ is nondecreasing in ϕ. We prove $\phi_0 < \phi_1$ for illustration and other inequalities can be proved similarly. Suppose, to the contrary that, $\phi_0 \ge \phi_1$, then we have $\chi_{i,\phi_1+1} \le \chi_{i,\phi_0+1}$, hence $\sum_{i\in\mathcal{I}} \lambda_{c,i}\chi_{i,\phi_1+1} \le \sum_{i\in\mathcal{I}} \lambda_{c,i}\chi_{i,\phi_0+1} < 1$. This contradicts with $\sum_{i\in\mathcal{I}} \lambda_{c,i}\chi_{i,\phi_1+1} \ge 1$ as $\phi_1 \in \mathcal{S}_0^c$.

(ii) If $\phi_1 \in \mathcal{S}_1$, then from Case 1 of Proposition EC.1, the second constraint in $\bar{\Pi}_{\phi_1}$ must be binding. That is, $\sum_{i\in\mathcal{I}} \chi_{i,\phi_1+1}\varphi_{c,i}^*(\phi_1) = 1$. Then because $\chi_{i,\phi}$ is nondecreasing in ϕ, we have $\chi_{i,\phi_1+1} \le \chi_{i,\phi_1+2}$ and $\sum_{i\in\mathcal{I}} \chi_{i,\phi_1+2}\varphi_{c,i}^*(\phi_1) \ge 1$. Hence, $\varphi_c^*(\phi_1) \in \bar{\Pi}_{\phi_1+1}$; that is, $\varphi_c^*(\phi_1)$ is a feasible solution to sub-problem $\phi_1 + 1$. Moreover, we have

$$\begin{aligned}
\pi^*(\phi_1) &= \sum_{i\in\mathcal{I}} \xi_{i,\phi_1}\varphi_{c,i}^*(\phi_1) + \frac{w_{s,\phi_1+1}r_{s,\phi_1+1}\mu_{s,\phi_1+1}}{\theta_{s,\phi_1+1}} \\
&= \sum_{i\in\mathcal{I}} \xi_{i,\phi_1+1}\varphi_{c,i}^*(\phi_1) + \frac{w_{s,\phi_1+1}r_{s,\phi_1+1}\mu_{s,\phi_1+1}}{\theta_{s,\phi_1+1}} \\
&\quad - \left(\frac{w_{s,\phi_1+1}r_{s,\phi_1+1}\mu_{s,\phi_1+1}}{\theta_{s,\phi_1+1}} - \frac{w_{s,\phi_1+2}r_{s,\phi_1+2}\mu_{s,\phi_1+2}}{\theta_{s,\phi_1+2}}\right)\sum_{i\in\mathcal{I}} \chi_{i,\phi_1+1}\varphi_{c,i}^*(\phi_1) \\
&= \sum_{i\in\mathcal{I}} \xi_{i,\phi_1+1}\varphi_{c,i}^*(\phi_1) + \frac{w_{s,\phi_1+1}r_{s,\phi_1+1}\mu_{s,\phi_1+1}}{\theta_{s,\phi_1+1}} \\
&\quad - \left(\frac{w_{s,\phi_1+1}r_{s,\phi_1+1}\mu_{s,\phi_1+1}}{\theta_{s,\phi_1+1}} - \frac{w_{s,\phi_1+2}r_{s,\phi_1+2}\mu_{s,\phi_1+2}}{\theta_{s,\phi_1+2}}\right) \\
&= \sum_{i\in\mathcal{I}} \xi_{i,\phi_1+1}\varphi_{c,i}^*(\phi_1) + \frac{w_{s,\phi_1+2}r_{s,\phi_1+2}\mu_{s,\phi_1+2}}{\theta_{s,\phi_1+2}} \le \pi^*(\phi_1+1),
\end{aligned}$$

where the second equality follows from (EC.6), and the third equality follows from $\sum_{i\in\mathcal{I}} \chi_{i,\phi_1+1}\varphi_{c,i}^*(\phi_1) = 1$.

(iii) First, we show that if $\mathcal{S}_2 \ne \emptyset$, then all elements in $\mathcal{S}_2$ have the same $\mathcal{I}_+(\phi)$. Assume, without loss of generality, that $\phi_2 < \phi_2'$ are in $\mathcal{S}_2$ and $\mathcal{I}_+(\phi_2) \ne \mathcal{I}_+(\phi_2')$. Then

$$1 \le \sum_{i\in\mathcal{I}_+(\phi_2)} \lambda_{c,i}\chi_{i,\phi_2+1} \le \sum_{i\in\mathcal{I}_+(\phi_2)} \lambda_{c,i}\chi_{i,\phi_2'} < \sum_{i\in\mathcal{I}_+(\phi_2')} \lambda_{c,i}\chi_{i,\phi_2'}.$$

The last strict inequality follows from Part (ii) of Lemma 4. Hence $\phi_2' \notin \mathcal{S}_2$, which is a contradiction. By Case 2 of Proposition EC.1, the solutions for sub-problems ϕ_2 and ϕ_2' are the same. As a result, $\pi^*(\phi_2) = \pi^*(\phi_2')$, for $\phi_2, \phi_2' \in \mathcal{S}_2$.

(iv) If $\phi_3 \in \mathcal{S}_3$, then from Case 3 of Proposition EC.1, the first constraint in $\bar{\Pi}_{\phi_3}$ must be binding. That is, $\sum_{i\in\mathcal{I}} \chi_{i,\phi_3}\varphi^*_{c,i}(\phi_3) = 1$. Then because $\chi_{i,\phi}$ is nondecreasing in ϕ, we have $\chi_{i,\phi_3-1} \le \chi_{i,\phi_3}$, and $\sum_{i\in\mathcal{I}} \chi_{i,\phi_3-1}\varphi^*_{c,i}(\phi_3) \le 1$. Hence, $\varphi^*_c(\phi_3) \in \bar{\Pi}_{\phi_3-1}$; that is, $\varphi^*_c(\phi_3)$ is a feasible solution to sub-problem $\phi_3 - 1$. Moreover, by (EC.6), it holds that $\xi_{i,\phi_3-1} = \xi_{i,\phi_3} - \left(\frac{w_{s,\phi_3} r_{s,\phi_3}\mu_{s,\phi_3}}{\theta_{s,\phi_3}} - \frac{w_{s,\phi_3+1} r_{s,\phi_3+1}\mu_{s,\phi_3+1}}{\theta_{s,\phi_3+1}}\right)\chi_{i,\phi_3}$. Hence, we have

$$\xi_{i,\phi_3} < \left(\frac{w_{s,\phi_3} r_{s,\phi_3}\mu_{s,\phi_3}}{\theta_{s,\phi_3}} - \frac{w_{s,\phi_3+1} r_{s,\phi_3+1}\mu_{s,\phi_3+1}}{\theta_{s,\phi_3+1}}\right)\chi_{i,\phi_3}, \tag{EC.9}$$

for $i \notin \mathcal{I}_+(\phi_3 - 1)$. Then, we have

$$\begin{aligned}
\pi^*(\phi_3) &= \sum_{i\in\mathcal{I}} \xi_{i,\phi_3}\varphi^*_{c,i}(\phi_3) + \frac{w_{s,\phi_3+1} r_{s,\phi_3+1}\mu_{s,\phi_3+1}}{\theta_{s,\phi_3+1}} \\
&\le \sum_{i\in\mathcal{I}_+(\phi_3-1)} \xi_{i,\phi_3}\varphi^*_{c,i}(\phi_3) + \frac{w_{s,\phi_3+1} r_{s,\phi_3+1}\mu_{s,\phi_3+1}}{\theta_{s,\phi_3+1}} \\
&\quad + \left(\frac{w_{s,\phi_3} r_{s,\phi_3}\mu_{s,\phi_3}}{\theta_{s,\phi_3}} - \frac{w_{s,\phi_3+1} r_{s,\phi_3+1}\mu_{s,\phi_3+1}}{\theta_{s,\phi_3+1}}\right)\left(1 - \sum_{i\in\mathcal{I}_+(\phi_3-1)} \chi_{i,\phi_3}\varphi^*_{c,i}(\phi_3)\right) \\
&= \sum_{i\in\mathcal{I}_+(\phi_3-1)} \xi_{i,\phi_3-1}\varphi^*_{c,i}(\phi_3) + \frac{w_{s,\phi_3} r_{s,\phi_3}\mu_{s,\phi_3}}{\theta_{s,\phi_3}} \le \pi^*(\phi_3 - 1),
\end{aligned}$$

where the inequality follows from $\sum_{i\in\mathcal{I}} \chi_{i,\phi_3}\varphi^*_{c,i}(\phi_3) = 1$ and (EC.9), and the second equality is due to (EC.6). This completes the proof. □

EC.2.4. Proof of Theorem 2

Recall that we can choose ϕ^* as the smallest index in $\mathcal{S}_2 \cup \mathcal{S}_3$. We have the following two cases.

(i) If $\mathcal{S}_2 \neq \emptyset$, $\phi^* \in \mathcal{S}_2$. Then φ^*_c is derived by Case (ii) of Proposition EC.1, and φ^*_s is derived by Lemma 2.

(ii) If $\mathcal{S}_2 = \emptyset$, $\phi^* \in \mathcal{S}_3$. Then φ^*_c is derived by Case (iii) of Proposition EC.1, and φ^*_s is derived by Lemma 2. □

EC.2.5. Proof of Theorem 3

Introduce $\mathcal{Z}_c(\cdot) := (\mathcal{Z}_{c,i}(\cdot))_{i\in\mathcal{I}}$ and $\mathcal{Q}_c(\cdot) := (\mathcal{Q}_{c,i}(\cdot))_{i\in\mathcal{I}}$, which are mappings from $\mathbb{N}_+^I$ to $\mathbb{N}_+^I$ and defined as

$$\mathcal{Z}_{c,i}(x) = x_i \wedge \left(N_c^{m,*} - \sum_{k=1}^{i-1} x_k\right)^+, \quad \mathcal{Q}_{c,i}(x) = x_i - \mathcal{Z}_{c,i}(x), \text{ for } x = (x_i)_{i\in\mathcal{I}}.$$

Also introduce $\mathcal{Z}_s(\cdot) := (\mathcal{Z}_{s,j}(\cdot))_{j\in\mathcal{J}}$ and $\mathcal{Q}_s(\cdot) := (\mathcal{Q}_{s,j}(\cdot))_{j\in\mathcal{J}}$, which are mappings from $\mathbb{N}_+^J$ to $\mathbb{N}_+^J$ and defined as

$$\mathcal{Z}_{s,j}(y) = y_j \wedge N_{s,j}^{m,*}, \quad \mathcal{Q}_{c,i}(y) = y_j - \mathcal{Z}_{s,j}(y), \text{ for } y = (y_j)_{j\in\mathcal{J}}.$$

Let $x_c^* = (x_{c,i}^*)_{i\in\mathcal{I}}$ and $x_s^* = (x_{s,j}^*)_{j\in\mathcal{J}}$, in which $x_{c,i}^* = z_{c,i}^* + q_{c,i}^*$ and $x_{s,j}^* = z_{s,j}^* + q_{s,j}^*$. Here, $(z_{c,i}^*, z_{s,j}^*) = (\frac{\varphi_{c,i}^*}{\mu_{c,i}}, \frac{\varphi_{s,j}^*}{\mu_{s,j}})$ according to the definition of φ, and $(q_{c,i}^*, q_{s,j}^*) = (\frac{\lambda_{c,i}-\varphi_{c,i}^*}{\theta_{c,i}}, \frac{\sum_{i\in\mathcal{I}}\varphi_{c,i}^* p_{ij} - \varphi_{s,j}^*}{\theta_{s,j}})$ by (23) and (24). The following lemma can be verified directly, hence its proof is omitted.

LEMMA EC.2. (i) *For any $x \in \mathbb{N}_+^I$, we have $\mathcal{Z}_c(x) + \mathcal{Q}_c(x) = x$. Moreover, $\mathcal{Z}_c(mx_c^*) = mz_c^*$, $\mathcal{Q}_c(mx_c^*) = mq_c^*$.*

(ii) *For any $x, x' \in \mathbb{N}_+^I$ and $i \in \mathcal{I}$, there exists a value $\rho_i(x,x') \in [0,1]$ such that $\mathcal{Q}_{c,i}(x) - \mathcal{Q}_{c,i}(x') = \rho_i(x,x')\cdot(x_i - x_i')$. As a result, $\mathcal{Z}_{c,i}(x) - \mathcal{Z}_{c,i}(x') = (1-\rho_i(x,x'))\cdot(x_i - x_i')$.*

(iii) *For any $y \in \mathbb{N}_+^J$, we have $\mathcal{Z}_s(y) + \mathcal{Q}_s(y) = y$. Moreover, $\mathcal{Z}_s(mx_s^*) = mz_s^*$, $\mathcal{Q}_s(mx_s^*) = mq_s^*$.*

(iv) *For any $y, y' \in \mathbb{N}_+^J$ and $j \in \mathcal{J}$, there exists a value $\varrho_j(y,y') \in [0,1]$ such that $\mathcal{Q}_{s,j}(y) - \mathcal{Q}_{s,j}(y') = \varrho_j(y,y')\cdot(y_i - y_i')$. As a result, $\mathcal{Z}_{s,j}(y) - \mathcal{Z}_{s,j}(y') = (1-\varrho_j(y,y'))\cdot(y_i - y_i')$.*

With the above mappings, for the mth system under policy $\psi^{m,*}$, the vectors $Z^m(t) = (Z_c^m(t), Z_s^m(t))$ and $Q^m(t) = (Q_c^m(t), Q_s^m(t))$ can be represented as functions of $X^m(t) = (X_c^m(t), X_s^m(t))$ as follows:

$$Z_c^m(t) = \mathcal{Z}_c(X_c^m(t)), \quad Q_c^m(t) = \mathcal{Q}_c(X_c^m(t)), \quad Z_s^m(t) = \mathcal{Z}_s(X_s^m(t)), \quad Q_s^m(t) = \mathcal{Q}_s(X_s^m(t)).$$

We first assume that there exists δ_m (satisfying $\lim_{m\to\infty}\delta_m = 0$) such that

$$\limsup_{t\to\infty} \mathbb{E}\left[\|m^{-1}X^m(t) - x^*\|\right] \le \delta_m. \tag{EC.10}$$

Let $\mathcal{Q}(\cdot) := (\mathcal{Q}_c(\cdot), \mathcal{Q}_s(\cdot))$. From Lemma EC.2 (ii) and (iv), we have $\|\mathcal{Q}(x) - \mathcal{Q}(x')\| \le \|x - x'\|$ for $x, x' \in \mathbb{N}_+^{I+J}$. This implies

$$\|m^{-1}Q^m(t) - q^*\| = m^{-1}\|\mathcal{Q}(X^m(t)) - \mathcal{Q}(mx^*)\| \le m^{-1}\|X^m(t) - mx^*\| = \|m^{-1}X^m(t) - x^*\|.$$

From (EC.10), we then have (32).

The remaining part of this section is devoted to proving (EC.10). Note that $X^m = (X_c^m, X_s^m)$ is a continuous-time Markov chain with generator

$$\begin{aligned}(\mathcal{L}f)(x,y) = &\sum_{i\in\mathcal{I}} \lambda_{c,i}^m\big(f(x+e_{i,I},y) - f(x,y)\big) + \sum_{i\in\mathcal{I}}\mu_{c,i}\mathcal{Z}_{c,i}(x)\Big[\sum_{j\in\mathcal{J}} p_{ij}\big(f(x-e_{i,I},y+e_{j,J}) - f(x,y)\big)\\ &+ \big(1-\sum_{j\in\mathcal{J}}p_{ij}\big)\big(f(x-e_{i,I},y) - f(x,y)\big)\Big] + \sum_{i\in\mathcal{I}}\theta_{c,i}\mathcal{Q}_{c,i}(x)\big(f(x-e_{i,I},y) - f(x,y)\big)\\ &+ \sum_{j\in\mathcal{J}}^{J}\big(\mu_{s,j}\mathcal{Z}_{s,j}(y) + \theta_{s,j}\mathcal{Q}_{s,j}(y)\big)\cdot\big(f(x,y-e_{j,J}) - f(x,y)\big), \quad (x,y)\in\mathbb{N}_+^I\times\mathbb{N}_+^J,\end{aligned} \tag{EC.11}$$

for any function f defined on $\mathbb{N}_+^I \times \mathbb{N}_+^J$. Here, we use $e_{a,n}$ to denote an n-dimensional column vector with all entries being 0 but its ath entry being 1. The generator $\mathcal{L}$ is well defined, as for

some $(x,y)\in\mathbb{N}_+^I\times\mathbb{N}_+^J$, if $x-e_{i,I}\notin\mathbb{N}_+^I$ (resp., $y-e_{j,J}\notin\mathbb{N}_+^J$), then $\mathcal{Z}_{c,i}(x)=\mathcal{Q}_{c,i}(x)=0$ (resp., $\mathcal{Z}_{s,j}(y)=\mathcal{Q}_{s,j}(y)=0$).

To proceed, we will use the following Lyapunov function defined on $\mathbb{N}_+^I\times\mathbb{N}_+^J$:

$$f^m(x,y)=\sum_{i\in\mathcal{I}}\beta_{c,i}(x_i-mx_{c,i}^*)^2+\sum_{j\in\mathcal{J}}\beta_{s,j}(y_j-mx_{s,j}^*)^2,$$

where $\beta_{c,i}$ and $\beta_{s,j}$ are strictly positive constants, whose values are to be determined. The following proposition establishes the Foster–Lyapunov drift condition for Markovian process $X^m=(X_c^m,X_s^m)$ regarding the function f^m.

PROPOSITION EC.2. *There exist constants $\beta_{c,i}>0$, $i\in\mathcal{I}$ and $\beta_{s,j}>0$, $j\in\mathcal{J}$ such that*

$$(\mathcal{L}f^m)(x,y)\le -a_1f^m(x,y)+a_2\|(x,y)\|+\epsilon_m m^2,\quad (x,y)\in\mathbb{N}_+^I\times\mathbb{N}_+^J,\quad m\ge m_0, \tag{EC.12}$$

where $a_1>0$, $a_2\ge 0$ and m_0 are constants not depending on (x,y) or m, and $\{\epsilon_m\}$ is a sequence of positive numbers that is independent of (x,y) and converges to zero.

We will also use the following growth property for $\mathbb{E}[\|X^m(t)\|]$.

LEMMA EC.3. *There exist constants $c_i>0$, $i=1,2,3$, such that*

$$\mathbb{E}[\|X^m(t)\|]\le c_1e^{-c_2t}\mathbb{E}[\|X^m(0)\|]+c_3m$$

for all $t\ge 0$ and sufficiently large m.

The proofs of Proposition EC.2 and Lemma EC.3 will be deferred to the end of this section.

Note that from $\mathbb{E}[\|X^m(0)\|^2]<\infty$ and $\mathbb{E}[\|E_c^m(t)\|^2]<\infty$ for all $t\ge 0$, and $\|X^m(t)\|\le\|X_c^m(0)\|+\|E_c^m(t)\|$, one can get $\mathbb{E}[\|X^m(t)\|^2]\le\mathbb{E}[(\|X_c^m(0)\|+\|E_c^m(t)\|)^2]<\infty$, and

$$\begin{aligned}\mathbb{E}[f^m(X_c^m(t),X_s^m(t))]=&\sum_{i\in\mathcal{I}}\beta_{c,i}\mathbb{E}[(X_{c,i}^m(t)-mx_{c,i}^*)^2]+\sum_{j\in\mathcal{J}}\beta_{s,j}\mathbb{E}[(X_{s,j}^m(t)-mx_{s,j}^*)^2)]\\
=&\sum_{i\in\mathcal{I}}\beta_{c,i}\mathbb{E}[(X_{c,i}^m(t))^2]+\sum_{j\in\mathcal{J}}\beta_{s,j}\mathbb{E}[(X_{s,j}^m(t))^2]-2m\sum_{i\in\mathcal{I}}\beta_{c,i}x_{c,i}^*\mathbb{E}[X_{c,i}^m(t)]\\
&-2m\sum_{j\in\mathcal{J}}\beta_{s,j}x_{s,j}^*\mathbb{E}[X_{s,j}^m(t)]+m^2\left(\sum_{i\in\mathcal{I}}\beta_{c,i}(x_{c,i}^*)^2+\sum_{j\in\mathcal{J}}\beta_{s,j}(x_{s,j}^*)^2\right)\\
\le&\beta_{\max}\mathbb{E}[\|X^m(t)\|^2]+(I+J)m^2\beta_{\max}(x_{\max}^*)^2<\infty,\end{aligned}$$

for all $t\ge 0$, where $\beta_{\max}=(\max_{i\in\mathcal{I}}\beta_{c,i})\vee(\max_{j\in\mathcal{J}}\beta_{s,j})$ and $x_{\max}^*=(\max_{i\in\mathcal{I}}x_{c,i}^*)\vee(\max_{j\in\mathcal{J}}x_{s,j}^*)$, and the last inequality follows from

$$\sum_{i\in\mathcal{I}}\mathbb{E}[(X_{c,i}^m(t))^2]+\sum_{j\in\mathcal{J}}\mathbb{E}[(X_{s,j}^m(t))^2]\le\mathbb{E}\Big[\Big(\sum_{i\in\mathcal{I}}\big|X_{c,i}^m(t)\big|+\sum_{j\in\mathcal{J}}\big|X_{s,j}^m(t)\big|\Big)^2\Big]=\mathbb{E}[\|X^m(t)\|^2].$$

As a result, the process

$$M(\cdot) := f^m(X_c^m(\cdot), X_s^m(\cdot)) - \int_0^{\cdot} (\mathcal{L}f^m)(X_c^m(s), X_s^m(s))ds$$

is a martingale (see, e.g., Theorem 8.3.1. of Øksendal 1998), and we have that for $m \geq m_0$,

$$\begin{aligned}\mathbb{E}[f^m(X_c^m(t), X_s^m(t))] &= \mathbb{E}[f^m(X_c^m(0), X_s^m(0))] + \mathbb{E}\left[\int_0^t (\mathcal{L}f^m)(X_c^m(s), X_s^m(s))ds\right] \\ &\leq \mathbb{E}[f^m(X_c^m(0), X_s^m(0))] \\ &\quad + \mathbb{E}\left[\int_0^t (-a_1 f^m(X_c^m(s), X_s^m(s)) + a_2\|(X_c^m(s), X_s^m(s))\| + \epsilon_m m^2)ds\right],\end{aligned}$$

where the inequality follows from Proposition EC.2. This, combining with Lemma EC.3, yields that for any sufficiently large m,

$$\begin{aligned}&\mathbb{E}[f^m(X_c^m(t), X_s^m(t))] \\ &\leq \mathbb{E}[f^m(X_c^m(0), X_s^m(0))] + \int_0^t \left(a_2\left(c_1 e^{-c_2 s}\mathbb{E}[\|X^m(0)\|] + c_3 m\right) + \epsilon_m m^2\right) ds \\ &\quad - a_1 \int_0^t \mathbb{E}[f^m(X_c^m(s), X_s^m(s))]ds \\ &\leq \mathbb{E}[f^m(X_c^m(0), X_s^m(0))] + \frac{a_2 c_1}{c_2}\mathbb{E}[\|X^m(0)\|] + (a_2 c_3 m + \epsilon_m m^2)t - a_1 \int_0^t \mathbb{E}[f^m(X_c^m(s), X_s^m(s))]ds.\end{aligned}$$

Then, using the generalized Gronwall's inequality (see, e.g., Viorel 1974), we have

$$\begin{aligned}\mathbb{E}[f^m(X_c^m(t), X_s^m(t))] \leq & e^{-a_1 t}\left(\mathbb{E}[f^m(X_c^m(0), X_s^m(0))] + \frac{a_2 c_1}{c_2}\mathbb{E}[\|X^m(0)\|]\right) \\ & + \frac{1}{a_1}(a_2 c_3 m + \epsilon_m m^2)(1 - e^{-a_1 t}).\end{aligned}$$

Note that there exists a positive constant ϑ that is sufficiently small such that

$$f^m(x, y) = \sum_{i\in\mathcal{I}} \beta_{c,i} x_i^2 + \sum_{j\in\mathcal{J}} \beta_{s,j} y_j^2 \geq \vartheta\left(\sum_{i\in\mathcal{I}} x_i + \sum_{j\in\mathcal{J}} y_j\right)^2 = \vartheta\|(x, y)\|^2,$$

for any $(x, y) \in \mathbb{R}_+^I \times \mathbb{R}_+^J$. Let $\delta_m := \sqrt{\frac{1}{a_1 m^2 \vartheta}(a_2 c_3 + \epsilon_m m)}$. Then we have

$$\limsup_{t\to\infty} \mathbb{E}\left[\|m^{-1}X^m(t) - x^*\|^2\right] \leq \limsup_{t\to\infty} \frac{1}{m^2\vartheta}\mathbb{E}[f^m(X_c^m(t), X_s^m(t))] \leq \delta_m^2.$$

This implies (EC.10). □

Proof of Proposition EC.2. The proof follows a similar argument as that for Lemma 3.1 in Atar et al. (2011). Fix $(x,y)\in\mathbb{N}_+^I\times\mathbb{N}_+^J$. For notational simplicity, we write $\zeta_i := x_i - mx^*_{c,i}$, $\eta_j := y_j - mx^*_{s,j}$, $\delta_{c,i} := \theta_{c,i}\wedge\mu_{c,i}$, and $\delta_{s,j} := \theta_{s,j}\wedge\mu_{s,j}$.

We first prove that

$$\begin{aligned}(\mathcal{L}f^m)(x,y) \le& \sum_{i\in\mathcal{I}}\beta_{c,i}\left(\lambda^m_{c,i} + (\mu_{c,i}+\theta_{c,i})x_i + 2\zeta_i\left(\lambda^m_{c,i} - m\lambda_{c,i}\right)\right) + \sum_{j\in\mathcal{J}}\beta_{s,j}\left(\sum_{i\in\mathcal{I}}\mu_{c,i}x_ip_{ij} + (\mu_{s,j}+\theta_{s,j})y_j\right)\\ &- 2\sum_{i\in\mathcal{I}}\beta_{c,i}\delta_{c,i}\zeta_i^2 + 2\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}}\mu_{c,i}\beta_{s,j}p_{ij}\zeta_i\eta_j - 2\sum_{j\in\mathcal{J}}\beta_{s,j}\delta_{s,j}\eta_j^2.\end{aligned} \tag{EC.13}$$

Note that

$$\begin{aligned}f(x+e_{i,I},y) - f(x,y) &= \beta_{c,i} + 2\beta_{c,i}\zeta_i,\\ f(x-e_{i,I},y+e_{j,J}) - f(x,y) &= \beta_{c,i} + \beta_{s,j} - 2\beta_{c,i}\zeta_i + 2\beta_{s,j}\eta_j,\\ f(x-e_{i,I},y) - f(x,y) &= \beta_{c,i} - 2\beta_{c,i}\zeta_i,\\ f(x,y-e_{j,J}) - f(x,y) &= \beta_{s,j} - 2\beta_{s,j}\eta_j.\end{aligned}$$

Thus, from (EC.11), we have

$$\begin{aligned}(\mathcal{L}f^m)(x,y) =& \sum_{i\in\mathcal{I}}\beta_{c,i}\big(\lambda^m_{c,i} + \mu_{c,i}\mathcal{Z}_{c,i}(x) + \theta_{c,i}\mathcal{Q}_{c,i}(x)\big) + 2\sum_{i\in\mathcal{I}}\beta_{c,i}\zeta_i\big(\lambda^m_{c,i} - \mu_{c,i}\mathcal{Z}_{c,i}(x) - \theta_{c,i}\mathcal{Q}_{c,i}(x)\big)\\ &+ \sum_{j\in\mathcal{J}}\beta_{s,j}\Big(\sum_{i\in\mathcal{I}}\mu_{c,i}\mathcal{Z}_{c,i}(x)p_{ij} + \mu_{s,j}\mathcal{Z}_{s,j}(y) + \theta_{s,j}\mathcal{Q}_{s,j}(y)\Big)\\ &+ 2\sum_{j\in\mathcal{J}}\beta_{s,j}\eta_j\Big(\sum_{i\in\mathcal{I}}\mu_{c,i}\mathcal{Z}_{c,i}(x)p_{ij} - \mu_{s,j}\mathcal{Z}_{s,j}(y) - \theta_{s,j}\mathcal{Q}_{s,j}(y)\Big).\end{aligned} \tag{EC.14}$$

Note that $0\le\mathcal{Z}_c(x)\le x$ and $0\le\mathcal{Q}_c(x)\le x$ for any $x\in\mathbb{N}_+^I$. (Here for any two vectors x, $x'\in\mathbb{N}_+^I$, we say that $x\le x'$ if $x_i\le x'_i$ for any $i\in\mathcal{I}$.) Hence, for the first term on the right-hand side of (EC.14), we have

$$\lambda^m_{c,i} + \mu_{c,i}\mathcal{Z}_{c,i}(x) + \theta_{c,i}\mathcal{Q}_{c,i}(x) \le \lambda^m_{c,i} + (\mu_{c,i}+\theta_{c,i})x_i. \tag{EC.15}$$

Similarly, because $0\le\mathcal{Z}_s(y)\le y$ and $0\le\mathcal{Q}_s(y)\le y$ for any $y\in\mathbb{N}_+^J$, for the third term on the right-hand side of (EC.14), we have

$$\sum_{i\in\mathcal{I}}\mu_{c,i}\mathcal{Z}_{c,i}(x)p_{ij} + \mu_{s,j}\mathcal{Z}_{s,j}(y) + \theta_{s,j}\mathcal{Q}_{s,j}(y) \le \sum_{i\in\mathcal{I}}\mu_{c,i}x_ip_{ij} + (\mu_{s,j}+\theta_{s,j})y_j. \tag{EC.16}$$

For the second term on the right-hand side of (EC.14), using Lemma EC.2 (i) and (ii), we have

$$\begin{aligned}&\lambda^m_{c,i} - \mu_{c,i}\mathcal{Z}_{c,i}(x) - \theta_{c,i}\mathcal{Q}_{c,i}(x)\\ =&\lambda^m_{c,i} - \mu_{c,i}\mathcal{Z}_{c,i}(mx^*_c) - \theta_{c,i}\mathcal{Q}_{c,i}(mx^*_c) - \left[\mu_{c,i}(1-\rho_i(x,mx^*_c)) + \theta_{c,i}\rho_i(x,mx^*_c)\right]\zeta_i\\ =&\lambda^m_{c,i} - m\lambda_{c,i} - \left[\mu_{c,i}(1-\rho_i(x,mx^*_c)) + \theta_{c,i}\rho_i(x,mx^*_c)\right]\zeta_i\\ \le&\lambda^m_{c,i} - m\lambda_{c,i} - \delta_{c,i}\zeta_i,\end{aligned} \tag{EC.17}$$

where the second equality holds because (q_c^*, z_c^*) satisfies (23), and the inequality holds because $\rho_i(x, mx_c^*) \in [0,1]$.

Similarly, for the fourth term on the right-hand side of (EC.14), we use Lemma EC.2 to obtain

$$\begin{aligned}
&\sum_{i\in\mathcal{I}} \mu_{c,i}\mathcal{Z}_{c,i}(x)p_{ij} - \mu_{s,j}\mathcal{Z}_{s,j}(y) - \theta_{s,j}\mathcal{Q}_{s,j}(y)\\
=&\sum_{i\in\mathcal{I}} \mu_{c,i}\mathcal{Z}_{c,i}(mx_c^*)p_{ij} - \mu_{s,j}\mathcal{Z}_{s,j}(mx_s^*) - \theta_{s,j}\mathcal{Q}_{s,j}(mx_s^*) + \sum_{i\in\mathcal{I}} \mu_{c,i}\rho_i(x, mx_c^*)\zeta_i p_{ij}\\
&- \Big[\mu_{s,j}(1-\varrho_j(y, mx_s^*)) + \theta_{s,j}\varrho_j(y, mx_s^*)\Big]\eta_j\\
\le&\sum_{i\in\mathcal{I}} \mu_{c,i}\zeta_i p_{ij} - \delta_{s,j}\eta_j, \qquad \text{(EC.18)}
\end{aligned}$$

where the inequality is from (24) and the fact that both $\rho_i(x, mx_c^*)$ and $\varrho_j(y, mx_s^*)$ are in $[0,1]$.

Combining the upper bound results in (EC.15)–(EC.18), and substituting them back to (EC.14), we then have (EC.13).

Next, we bound the terms on the right-hand of (EC.13) to get (EC.12). Fix $a_1 > 0$ such that $a_1 < 2\big(\min_{i\in\mathcal{I}}\delta_{c,i} \wedge \min_{j\in\mathcal{J}}\delta_{s,j}\big)$. Then there exist $\beta_{c,i}$'s that are appropriately chosen such that

$$2\beta_{c,i}\delta_{c,i} - \sum_{j\in\mathcal{J}} p_{ij}\mu_{c,i}^2 - a_1\beta_{c,i} \ge 0, \qquad \text{for } i\in\mathcal{I},$$

Also, there exist $\beta_{s,j}$'s that are appropriately chosen such that

$$2\beta_{s,j}\delta_{s,j} - \sum_{i\in\mathcal{I}} p_{ij}\beta_{s,j}^2 - a_1\beta_{s,j} \ge 0, \qquad \text{for } j\in\mathcal{J}.$$

These then imply

$$\sum_{i\in\mathcal{I}}\Big(2\beta_{c,i}\delta_{c,i} - \sum_{j\in\mathcal{J}} p_{ij}\mu_{c,i}^2 - a_1\beta_{c,i}\Big)\zeta_i^2 + \sum_{j\in\mathcal{J}}\Big(2\beta_{s,j}\delta_{s,j} - \sum_{i\in\mathcal{I}} p_{ij}\beta_{s,j}^2 - a_1\beta_{s,j}\Big)\eta_j^2 \ge 0.$$

This, together with $2\mu_{c,i}\beta_{s,j}\zeta_i\eta_j \le \mu_{c,i}^2\zeta_i^2 + \beta_{s,j}^2\eta_j^2$, gives

$$\begin{aligned}
&2\sum_{i\in\mathcal{I}}\beta_{c,i}\delta_{c,i}\zeta_i^2 - 2\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} p_{ij}\mu_{c,i}\beta_{s,j}\zeta_i\eta_j + 2\sum_{j\in\mathcal{J}}\beta_{s,j}\delta_{s,j}\eta_j^2\\
&\ge a_1\Big(\sum_{i\in\mathcal{I}}\beta_{c,i}\zeta_i^2 + \sum_{j\in\mathcal{J}}\beta_{s,j}\eta_j^2\Big) = a_1 f^m(x,y).
\end{aligned} \qquad \text{(EC.19)}$$

Note that by (17), for any $m_0 > 1$, there exists a constant $a_2 > 0$, and a sequence $\{\epsilon_m; m\in\mathbb{N}\}$ with $\lim_{m\to\infty}\epsilon_m = 0$ such that

$$\begin{aligned}
&\sum_{i\in\mathcal{I}}\beta_{c,i}\big(\lambda_{c,i}^m + (\mu_{c,i}+\theta_{c,i})x_i + 2\zeta_i\left(\lambda_{c,i}^m - m\lambda_{c,i}\right)\big) + \sum_{j\in\mathcal{J}}\beta_{s,j}\Big(\sum_{i\in\mathcal{I}}\mu_{c,i}x_i p_{ij} + (\mu_{s,j}+\theta_{s,j})y_j\Big)\\
&= \sum_{i\in\mathcal{I}}\beta_{c,i}\Big(\mu_{c,i}\big(1+\sum_{j\in\mathcal{J}}\beta_{s,j}p_{ij}\big) + \theta_{c,i}\Big)x_i + \sum_{j\in\mathcal{J}}\beta_{s,j}\left(\mu_{s,j}+\theta_{s,j}\right)y_j + 2m^2\sum_{i\in\mathcal{I}}\beta_{c,i}\left(\frac{x_i}{m} - x_{c,i}^*\right)\left(\frac{\lambda_{c,i}^m}{m} - \lambda_{c,i}\right)\\
&\le a_2\|(x,y)\| + \epsilon_m m^2, \qquad \text{(EC.20)}
\end{aligned}$$

for any $(x,y)\in\mathbb{N}_+^I\times\mathbb{N}_+^J$ and $m\geq m_0$.

Combining (EC.13), (EC.19) and (EC.20), we have (EC.2). The proof is complete. □

Proof of Lemma EC.3. We first establish a bound for $\mathbb{E}\left[X_{c,i}^m(t)\right]$, $i\in\mathcal{I}$. For any $i\in\mathcal{I}$, from (1), (3)–(5), we have (let $\delta_{c,i}:=\theta_{c,i}\wedge\mu_{c,i}$)

$$\begin{aligned}\mathbb{E}\left[X_{c,i}^m(t)\right]&=\mathbb{E}\left[X_{c,i}^m(0)\right]+\lambda_{c,i}^m t-\theta_{c,i}\int_0^t\mathbb{E}\left[Q_{c,i}^m(s)\right]dt-\mu_{c,i}\int_0^t\mathbb{E}\left[Z_{c,i}^m(s)\right]ds\\&\leq\mathbb{E}\left[X_{c,i}^m(0)\right]+\lambda_{c,i}^m t-\delta_{c,i}\int_0^t\mathbb{E}\left[X_{c,i}^m(s)\right]ds,\quad i\in\mathcal{I}\end{aligned}$$

for any $t\geq 0$, where the inequality is from $X_{c,i}^m(\cdot)=Q_{c,i}^m(\cdot)+Z_{c,i}^m(\cdot)$. Hence, from the generalized Gronwall's inequality (see, e.g., Viorel 1974), for $i\in\mathcal{I}$ and $t\geq 0$,

$$\begin{aligned}\mathbb{E}\left[X_{c,i}^m(t)\right]&\leq\mathbb{E}\left[X_{c,i}^m(0)\right]+\lambda_{c,i}^m t-\delta_{c,i}\int_0^t\left(\mathbb{E}\left[X_{c,i}^m(0)\right]+\lambda_{c,i}^m s\right)e^{-\delta_{c,i}(t-s)}ds\\&=e^{-\delta_{c,i}t}\mathbb{E}[X_{c,i}^m(0)]+\frac{\lambda_{c,i}^m}{\delta_{c,i}}\left(1-e^{-\delta_{c,i}t}\right).\end{aligned}\tag{EC.21}$$

Next we establish a bound for $\sum_{j\in\mathcal{J}}\mathbb{E}\left[X_{s,j}^m(t)\right]$, $j\in\mathcal{J}$. Fix $j\in\mathcal{J}$. From (7) and (8) to (10), we have (let $\delta_{s,j}:=\theta_{s,j}\wedge\mu_{s,j}$)

$$\begin{aligned}\mathbb{E}\left[X_{s,j}^m(t)\right]=&\mathbb{E}\left[X_{s,j}^m(0)\right]+\sum_{i\in\mathcal{I}}p_{ij}\int_0^t\mu_{c,i}\mathbb{E}\left[Z_{c,i}^m(s)\right]ds-\theta_{s,j}\int_0^t\mathbb{E}\left[Q_{s,j}^m(s)\right]ds-\mu_{s,j}\int_0^t\mathbb{E}\left[Z_{s,j}^m(s)\right]ds\\\leq&\mathbb{E}\left[X_{s,j}^m(0)\right]+\sum_{i\in\mathcal{I}}p_{ij}\mu_{c,i}\int_0^t\mathbb{E}\left[X_{c,i}^m(s)\right]ds-\delta_{s,j}\int_0^t\mathbb{E}\left[X_{s,j}^m(s)\right]ds\\\leq&\mathbb{E}\left[X_{s,j}^m(0)\right]+\sum_{i\in\mathcal{I}}\frac{p_{ij}\mu_{c,i}}{\delta_{c,i}}\left[\left(\mathbb{E}[X_{c,i}^m(0)]-\frac{\lambda_{c,i}^m}{\delta_{c,i}}\right)\cdot\left(1-e^{-\delta_{c,i}t}\right)+\lambda_{c,i}^m t\right]-\delta_{s,j}\int_0^t\mathbb{E}\left[X_{s,j}^m(s)\right]ds\\\leq&\mathbb{E}\left[X_{s,j}^m(0)\right]+\sum_{i\in\mathcal{I}}\frac{p_{ij}\mu_{c,i}}{\delta_{c,i}}\left(\mathbb{E}[X_{c,i}^m(0)]+\lambda_{c,i}^m t\right)-\delta_{s,j}\int_0^t\mathbb{E}\left[X_{s,j}^m(s)\right]ds,\end{aligned}$$

where the second inequality follows from (EC.21). Hence, by invoking the generalized Gronwall's inequality for $\mathbb{E}[X_{s,j}(t)]$, we have

$$\mathbb{E}\left[X_{s,j}^m(t)\right]\leq e^{-\delta_{s,j}t}\left(\mathbb{E}[X_{s,j}^m(0)]+\sum_{i\in\mathcal{I}}\frac{p_{ij}\mu_{c,i}}{\delta_{c,i}}\mathbb{E}[X_{c,i}^m(0)]\right)+\sum_{i\in\mathcal{I}}\frac{p_{ij}\mu_{c,i}\lambda_{c,i}^m}{\delta_{c,i}\delta_{s,j}}(1-e^{-\delta_{s,j}t}),\tag{EC.22}$$

for all $j\in\mathcal{J}$ and $t\geq 0$. Finally, by combining (EC.21) and (EC.22), we have

$$\begin{aligned}\mathbb{E}\left[\|X^m(t)\|\right]=&\sum_{i\in\mathcal{I}}\mathbb{E}\left[X_{c,i}^m(t)\right]+\sum_{j\in\mathcal{J}}\mathbb{E}\left[X_{s,j}^m(t)\right]\\\leq&\sum_{i\in\mathcal{I}}\left(e^{-\delta_{c,i}t}+\sum_{j\in\mathcal{J}}e^{-\delta_{s,j}t}\frac{p_{ij}\mu_{c,i}}{\delta_{c,i}}\right)\mathbb{E}[X_{c,i}^m(0)]+\sum_{j\in\mathcal{J}}e^{-\delta_{s,j}t}\mathbb{E}[X_{s,j}^m(0)]\end{aligned}$$

$$+\sum_{i\in\mathcal{I}}\frac{\lambda^m_{c,i}}{\delta_{c,i}}\left(1-e^{-\delta_{c,i}t}\right)+\sum_{j\in\mathcal{J}}\sum_{i\in\mathcal{I}}\frac{p_{ij}\mu_{c,i}\lambda^m_{c,i}}{\delta_{c,i}\delta_{s,j}}\left(1-e^{-\delta_{s,j}t}\right)$$
$$\leq\left(I+J+\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}}\frac{p_{ij}\mu_{c,i}}{\delta_{c,i}}\right)e^{-\delta_{\min}t}\mathbb{E}[\|X^m(0)\|]+\sum_{i\in\mathcal{I}}\frac{\lambda^m_{c,i}}{\delta_{c,i}}\left(1+\sum_{j\in\mathcal{J}}\frac{p_{ij}\mu_{c,i}}{\delta_{s,j}}\right),$$

where $\delta_{\min}:=(\min_{i\in\mathcal{I}}\delta_{c,i})\wedge(\min_{j\in\mathcal{J}}\delta_{s,j})$. This, together with (17), yields the desired result. □

References

Atar, R., C. Giat, and N. Shimkim (2011). On the asymptotic optimality of the $c\mu/\theta$ rule under ergodic cost. *Queueing Systems 67*(2), 127–144.

Dong, J., P. Feldman, and G. B. Yom-Tov (2015). Service systems with slowdowns: Potential failures and proposed solutions. *Operations Research 63*(2), 305–324.

Eick, S. G., W. A. Massey, and W. Whitt (1993). The physics of the $M_t/G/\infty$ queue. *Operations Research 41*(4), 731–742.

Øksendal, B. K. (1998). *Stochastic Differential Equations: An Introduction with Applications.* Berlin: Springer.

Viorel, B. (1974). Principles of differential and integral equations. *Journal of the Franklin Institute 298*(3), 246–247.