Mathematical Methods of Operations Research (2019) 90:399–451
https://doi.org/10.1007/s00186-019-00683-9

ORIGINAL ARTICLE

Virtual allocation policies for many-server queues with abandonment

Zhenghua Long[1,2] · **Jiheng Zhang**[1]

Received: 7 April 2018 / Revised: 21 July 2019 / Published online: 8 October 2019

Abstract
We study a multiclass many-server queueing system with renewal arrivals and generally distributed service and patience times under a nonpreemptive allocation policy. The status of the system is described by a pair of measure-valued processes to track the residual service and patience times of customers in each class. We establish fluid approximations and study the long-term behavior of the fluid model. The equilibrium state of the fluid model leads to a nonlinear program, which enables us to identify a lower bound for the long-run expected total holding and abandonment costs and design an allocation policy to achieve this lower bound. The optimality of the proposed policy is also demonstrated via numerical experiments.

Keywords Multiclass queue · Customer abandonment · Fluid limits · Measure-valued processes

1 Introduction

Multiclass many-server queueing models have been extensively used to model service systems such as telephone call centers, e.g., Mandelbaum et al. (1998), Mandelbaum and Stolyar (2004). This paper considers such a model where I classes of customers are served by a pool of many homogeneous servers. Customers of class i arrives according to a renewal process with rate λ_i, for $i = 1, \ldots, I$. Each class has its own queue with infinite capacity to keep customers who cannot be served immediately upon arrival. Customers within each class are served based on the first-come-first-served (FCFS)

✉ Zhenghua Long
zlong@nju.edu.cn

Jiheng Zhang
jiheng@ust.hk

1 School of Management, Nanjing University, Nanjing 210093, China

2 Department of Industrial Engineering and Decision Analytics, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong

discipline. Each class-i customer abandons the system once the waiting time exceeds his patience time, which is modeled using a random variable following a general distribution F_i. Customers will not abandon once service has started. Service times of class-i customers follow distribution G_i with mean $1/\mu_i$.

In such a multiclass queueing model, deciding which class to serve when a server becomes available is a stochastic control problem. The objective is to minimize the long-run average holding cost, c_i per unit time for a class-i customer waiting in the queue, plus penalty cost, γ_i for each abandoned class-i customer. We study the stochastic control problem in the many-server heavy-traffic regime, where the arrival rate of each class and the number of servers tend to infinity proportionally. In the Markovian case both G_i and F_i are exponential distributions with rates μ_i and θ_i, respectively. Atar et al. (2010, 2011) showed that a priority policy, referred to as the $c\mu/\theta$ rule, is asymptotically optimal in the many-server heavy-traffic regime. The $c\mu/\theta$ rule is a priority policy that assigns priority to classes according to the value $c_i\mu_i/\theta_i$ (the higher the value, the higher the priority). A customer can start service only when there is no higher priority class customers waiting in the queue. Recently, Atar et al. (2014) established fluid limits for many-server systems with abandonment under a priority policy. Their work extended the optimality of the $c\mu/\theta$ rule to a more general setting which allows renewal arrivals and general service time distributions. But it remains an open problem whether the $c\mu/\theta$ rule is asymptotically optimal when the patience time distributions are general. If it is not, can we identify an asymptotically optimal policy?

A challenge caused by the general patience time distributions is that the queue length and abandonment count do not exhibit a simple linear relationship. In fact, the above described $c\mu/\theta$ rule only accounts for the holding cost of queues. To incorporate abandonment penalty, the cost coefficient c_i simply needs to be modified to $(c_i + \theta_i\gamma_i)$ thanks to the simple linear relationship (see Atar et al. 2010 for details). In other words, the total holding and abandonment costs can be expressed solely as a queue-length cost. Also when the patience time distributions are general, it is not necessary for them to have a finite mean. When the means are infinite, i.e., θ_i's are 0, one cannot identify a priority from the $c\mu/\theta$ rule. We demonstrate via numerical examples in Sect. 4.2.3 that the $c\mu/\theta$ rule (when θ_i's are non-zero) may not always be optimal. We propose a *virtual allocation* policy and show that it is asymptotically optimal when patience times have decreasing hazard rates.

Our proposed policy virtually allocates a fraction z_i of the servers to class i in the sense that servers allocated to class i will always try to find a class-i customer to serve. Only when there are no class-i customers waiting in the queue will they serve other classes of customers. We characterize the system dynamics under the virtual allocation policy and establish the fluid limit in the many-server heavy-traffic regime. We then show that the fluid limit converges to an equilibrium state as time goes to infinity.

The equilibrium state leads to a nonlinear optimization problem whose solution provides an asymptotic lower bound for achievable cost under any policy. The nonlinear part of the optimization results from the generality of the patience time distribution. It is not necessary for patience times to have decreasing hazard rates for the heavy-traffic analysis of the stochastic processes or the convergence of the fluid limit to the equilibrium. However, it becomes necessary if we want to establish the lower bound

on achievable performance. One reason is that we restrict our policies in the FCFS class, which is shown by Bassamboo and Randhawa (2013) to be suboptimal even in the single-class setting. Another reason is that when the patience time distribution is general, it is possible for a dynamic policy to do better as our virtual allocation is still a static policy [e.g., see Kim and Ward (2013) for dynamic policies of a multiclass system with only a single server].

1.1 Related literature

Our work relates to the growing literature on optimal control of multiclass queueing systems with many servers. Effective control policies have been devised through asymptotic analysis of the underlying stochastic processes in the heavy-traffic regime (see Mandelbaum et al. 1998; Whitt 2004). Focusing on diffusion approximations in the quality-and-efficiency-driven regime proposed by Halfin and Whitt (1981), Mandelbaum and Stolyar (2004) proved a generalized $c\mu$ rule is asymptotically optimal with convex delay costs; Atar et al. (2004) and Atar (2005) studied dynamic scheduling policies by formulating a Hamilton-Jacobi-Bellman equation based on the heavy-traffic limits; Dai and Tezcan (2008) developed robust control policies to minimize the sum of holding and reneging costs; Gurvich and Whitt (2009) proposed queue-and-idleness-ratio rules to solve staffing and control problems. Such asymptotic analysis has also been popular in studying systems with a single server in the conventional heavy-traffic regime. van Mieghem (1995) proposed the generalized $c\mu$ and proved its optimality. Harrison and López (1999) explicitly solved a dynamic control problem in the multiclass multi-server setting. Recently, Kim and Ward (2013) and Ata and Tongarlak (2013) considered dynamic index policies by solving Bellman equations based on the heavy-traffic limits.

It turns out that fluid models are useful in studying many-server queues particularly in the overloaded regime. Bassamboo et al. (2006) proposed joint staffing and control for a parallel server system with time-varying arrival rates based on fluid approximations. Perry and Whitt (2011) developed fluid approximations for threshold-based control policies to respond to unexpected overloads. A fluid model for many-server queues with generally distributed service and patience times was proposed by Whitt (2006), where approximation formulae for various performances were constructed based on the equilibrium state of the fluid model and simulations showed that the formulae are quite accurate. It has been rigorously proved in Bassamboo and Randhawa (2010) that the gap of fluid approximation remains bounded as the system size increases to infinity in the heavy-traffic regime.

The fluid model of Whitt (2006) was proved to be the fluid limit by Kang and Ramanan (2010) and Zhang (2013) using measure-valued processes. Liu and Whitt (2012a, b) extended the two-parameter fluid model to allow time-varying arrival rate and staffing. These fluid models have been shown to be uniquely determined by the same one-dimensional convolution equation in Long and Zhang (2019). Kang and Ramanan (2010) modeled the status of the system by keeping track of the "age" (the amount of time a customer has been in queue or in service) following Kaspi and Ramanan (2011) on many-server queues without abandonment. Recently, Atar et al.

(2014) developed a nice Skorohod map to extend Kang and Ramanan (2010) to the multiclass setting under priority policies. Moreover, Atar et al. (2014) showed that the $c\mu/\theta$ rule is asymptotically optimal in the case where patience time distributions are exponential, extending the optimality proved in the Markovian setting by Atar et al. (2010, 2011) to allow renewal arrivals and general service times. We aim to relax the assumption on exponential patience times by taking an alternative approach in Zhang (2013), which modeled the status of the system by tracking each customer's "residual life" (the remaining service and patience times). Using this alternative approach, Long and Zhang (2014) showed that the fluid model converges to the equilibrium state as time goes to infinity. The present work extends the heavy-traffic fluid analysis to the multiclass setting under the above mentioned virtual allocation policy.

1.2 Contributions

The main contributions of this paper can be summarized as follows.

- We use measure-valued processes and the corresponding measure-valued fluid models to analyze a multiclass $G/GI/n+GI$ queuing system by tracking each customer's "residual life" (the remaining service and patience times). Note that Atar et al. (2014) studied the same system under priority policies by tracking each customer's "age" (the amount of time a customer has been in queue or in service). Thus this paper provides an alternative approach to study the multiclass $G/GI/n+GI$ queue.
- When the multiclass many-server queue is underloaded, i.e., $\sum_{i=1}^{I}\lambda_i/\mu_i < 1$, we prove in Theorem 2 that any non-idling policy is asymptotically optimal for any general service and patience time distributions.
- When the multiclass many-server queue is critically loaded or overloaded, i.e., $\sum_{i=1}^{I}\lambda_i/\mu_i \geq 1$, we show in Theorem 3 that the virtual allocation policy is asymptotically optimal for any general patience time distributions but requiring the renewal functions of the service time distributions are either convex or concave. We also demonstrate the effectiveness of our policy through various numerical experiments in Sect. 4.
- This paper also expends our understanding on the convergence to equilibrium states for fluid models of many-server queues with general service and patience time distributions. Even for the single-class $G/GI/n+GI$ fluid model, this remains an open problem [see Theorem 2 in Long and Zhang (2014), where an additional assumption on the initial state of the server pool is needed for critically loaded and overloaded systems].
- For underloaded systems, we prove the convergence of the fluid model of multiclass many-server queues under any fluid non-idling policy for any general service and patience time distributions in Theorem 4.
- For critically loaded and overloaded systems, we relax the initial condition (3.7) in our previous work Long and Zhang (2014) to be the more general initial condition (50). We also prove in Theorem 5 the convergence of the fluid model under the fluid virtual allocation policy benefiting from the assumption of convexity or concavity of the renewal functions of the service time distributions. Our proposed policy

adapts automatically to the change of the allocation of the server pool due to the general setting of the initial state of the server pool in this paper. For example, as a result of the varities of the arrival rates in different time periods, the allocation of the server pool changes accordingly. Starting from the new allocation of the server pool, this paper ensures that the fluid model can still converge to equilibrium states. However, this cannot be guaranteed by the result in Long and Zhang (2014).

The rest of this paper is organized as follows. We introduce the control problem and the asymptotic framework as well as the stochastic and fluid models in Sect. 2, where we also show that the fluid-scaled stochastic processes converge to the fluid model in the heavy-traffic regime. In Sect. 3, a nonlinear program associated with equilibrium states of fluid models is proposed. Its solution is shown to be a lower bound for the long-run expected holding and abandonment costs for any control policy. Section 4 proposes control policies that can achieve the lower bound. We also demonstrate the effectiveness of our proposed virtual allocation policy using numerical experiments. Section 5 provides analysis of the fluid models and shows that the fluid models converge to equilibrium states. We state our conclusions in Sect. 6. "Appendix" contains some auxiliary lemmas.

We conclude this section by introducing the notation and definitions that will be used throughout this paper. For $a, b \in \mathbb{R}$, write a^+ for the positive part of a and $a \wedge b$ for the minimum. Denote by $\mathbb{R}$ the set of real numbers and $\mathbb{R}_+ = [0, \infty)$. For a sequence of random elements $\{X^n\}_{n\in\mathbb{N}}$, taking values in a metric space, we write $X^n \Rightarrow X$ to denote the weak convergence of X^n to X. Let $\mathbf{M}$ denote the set of all non-negative finite Borel measures on $\mathbb{R}$. For $\nu_1, \nu_2 \in \mathbf{M}$, the Prohorov metric is defined to be

$$\begin{aligned}\mathbf{d}[\nu_1, \nu_2] = \inf\{&\varepsilon > 0 : \nu_1(A) \le \nu_2(A^\varepsilon) + \varepsilon \\ &\text{and } \nu_2(A) \le \nu_1(A^\varepsilon) + \varepsilon \text{ for all closed Borel set } A \subset \mathbb{R}\},\end{aligned} \tag{1}$$

where $A^\varepsilon = \{b \in \mathbb{R} : \inf_{a\in A} |a - b| < \varepsilon\}$. This is the metric that induces the topology of weak convergence of finite Borel measure (see Billingsley 1999). Similarly, let $\mathbf{M}_+$ denote the set of all non-negative finite Borel measures on $(0, \infty)$. To simplify the notation, let us take the convention that for any Borel set $A \subset \mathbb{R}$, $\nu(A \cap (-\infty, 0]) = 0$ for any $\nu \in \mathbf{M}_+$. Also, by this convention, $\mathbf{M}_+$ is embedded as a subspace of $\mathbf{M}$. We consider the space $\mathbf{D}([0, \infty), \mathbf{M})$ ($\mathbf{D}([0, \infty), \mathbf{M}_+)$) of all right-continuous $\mathbf{M}$-valued ($\mathbf{M}_+$-valued) functions with finite left limits defined on the infinite interval $[0, \infty)$. Let $\mathbf{D}([0, \infty), \mathbb{R})$ be the space of right-continuous functions with left limits, defined on $[0, \infty)$ and taking real values. We equip these spaces with the Skorohod J_1-topology (see Ethier and Kurtz 1986).

2 Model and asymptotic framework

There is a single pool of n homogeneous servers serving I classes of customers (see Fig. 1). To emphasize its dependence on the number of servers, we append a superscript n to all the notations and refer to this system as the nth system. For $i \in \mathcal{I} := \{1, 2, \ldots, I\}$, let $X_i^n(t)$ denote the total number of class-i customers in the nth system

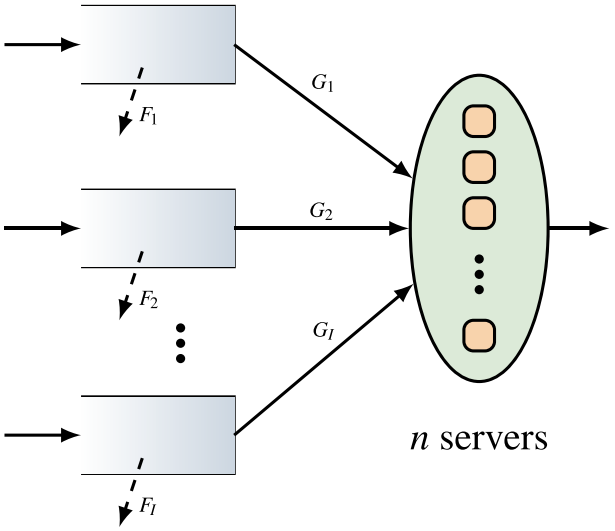


Fig. 1 The many-server queue with multiple customer classes

including $Z_i^n(t)$ customers being served in the server pool and $Q_i^n(t)$ customers waiting in the queue. Clearly, for all $t \geq 0$,

$$\sum_{i \in \mathscr{I}} Z_i^n(t) \leq n, \tag{2}$$

$$X_i^n(t) = Q_i^n(t) + Z_i^n(t), \quad i \in \mathscr{I}. \tag{3}$$

For each class i, there is an exogenous stream of arrivals denoted by $E_i^n(\cdot)$. Class-i customers' service times are i.i.d. random variables following distribution G_i. The patience times of class-i customers are also i.i.d. random variables following distribution F_i. Note that both G_i's and F_i's are independent of n. Assume the service and patience times for all classes are mutually independent. We use $A_i^n(t)$, $L_i^n(t)$ and $S_i^n(t)$ to denote the cumulative number of class-i customers who have entered service, abandoned queue and completed service by time t, respectively. The above processes are related through the following balance equations, for all $i \in \mathscr{I}$,

$$Q_i^n(t) = Q_i^n(0) + E_i^n(t) - L_i^n(t) - A_i^n(t), \tag{4}$$

$$Z_i^n(t) = Z_i^n(0) + A_i^n(t) - S_i^n(t). \tag{5}$$

A control policy decides how each class of customer is scheduled into service, based only on the observable information about the system status. We require that a customer can enter service either upon a service completion or upon arrival to prevent routing a batch of customers waiting in queue to the server pool. This means

$$\sum_{i \in \mathscr{I}} A_i^n(s,t) \leq \sum_{i \in \mathscr{I}} E_i^n(s,t) + \sum_{i \in \mathscr{I}} S_i^n(s,t), \quad 0 \leq s \leq t, \tag{6}$$

where $A_i^n(s,t) := A_i^n(t) - A_i^n(s)$ is the number of class-i customers who enter service during $(s,t]$ and $E_i^n(s,t)$, $S_i^n(s,t)$ are defined in a similar way. Following the discussion of Atar et al. (2010), a process

$$\pi^n = (A_i^n, L_i^n, S_i^n, X_i^n, Q_i^n, Z_i^n)_{i\in\mathscr{I}} \tag{7}$$

will be referred to as a control policy for the nth system, provided that (2)–(6) hold. The policy need not satisfy any nonidling constraint. Denote by Π^n the collection of all control policies satisfying (2)–(6) for the nth system. We focus on how to route different classes of customers assuming FCFS within each class. Readers are referred to Bassamboo and Randhawa (2016) for a study of non-FCFS policies.

Cost structure For any $i \in \mathscr{I}$, a holding cost $c_i \geq 0$ is incurred per unit time for each class-i customer waiting in the queue. For each class-i customer who has abandoned queue, there is a penalty cost $\gamma_i \geq 0$. So the average cost over the planning horizon $[0, T]$ under a control policy π^n is

$$C_T^n(\pi^n) = \frac{1}{T}\mathbb{E}\left[\sum_{i\in\mathscr{I}}\int_0^T c_i Q_i^n(t)dt + \sum_{i\in\mathscr{I}}\gamma_i L_i^n(T)\right]. \tag{8}$$

And set

$$\bar{C}_T^n(\pi^n) = \frac{1}{n}C_T^n(\pi^n),$$

which means the cost function is "rescaled" as the parameter n changes. When the patience time distributions are exponential, there is a simple relationship between the abandonment count and queue length processes. So the above cost function can be transformed into a cost function involving only the queue length process [see Remark 2.1 in Atar et al. (2010)]. Since we allow more general patience time distributions, we have to keep the abandonment process in the objective function, and deal with the associated difficulty it brings to the analysis.

The main purpose of this paper is to find a sequence of control policies such that it is asymptotically optimal among all the control policies.

Definition 1 A sequence of control policies $\{\pi_*^n : \pi_*^n \in \Pi^n\}$ is asymptotically optimal if

$$\limsup_{T\to\infty}\limsup_{n\to\infty}\bar{C}_T^n(\pi_*^n) \leq \liminf_{T\to\infty}\liminf_{n\to\infty}\bar{C}_T^n(\pi^n) \quad \text{for any } \pi^n \in \Pi^n.$$

To analyze the stochastic processes underlying this model with general service and patience times, focusing only on those introduced above does not suffice. In the following section, we introduce the measure-valued modeling. Note that the measure-valued modeling serves only the purpose of analysis. In reality, the control policy still only depends on the head-count processes introduced above.

2.1 Measure-valued modeling

Following Zhang (2013), we use a pair of measures $\left(\mathscr{R}_i^n(t), \mathscr{Z}_i^n(t)\right)$, $i \in \mathscr{I}$, to describe the status of all customers in class i at time t. Specifically, $\mathscr{R}_i^n(t)(C)$ is the number of class-i customers in the *virtual buffer* for class i with remaining patience time in the Borel set $C \subset \mathbb{R}$, and $\mathscr{Z}_i^n(t)(C)$ is the number of class-i customers in service with remaining service time in the Borel set $C \subset \mathbb{R}_+$. The virtual buffer holds all arriving customers who have not yet been scheduled for service, regardless of whether or not their patience times have expired. This means that in addition to customers waiting in the queue, the virtual buffer also holds some of the customers who have abandoned the system. So customers in the virtual buffer may have negative remaining patience time. Using this modeling approach, the number of class-i customers in the virtual buffer, in the physical buffer and in service can be determined as

$$R_i^n(t) = \mathscr{R}_i^n(t)(\mathbb{R}), \quad Q_i^n(t) = \mathscr{R}_i^n(t)((0,\infty)) \text{ and } Z_i^n(t) = \mathscr{Z}_i^n(t)((0,\infty)), \quad (9)$$

respectively. Initially, there are $R_i^n(0)$ customers in the virtual buffer of class i. Index them using $l = -R_i^n(0)+1, \ldots, 0$ according to their arrival time $a_{i,l}^n$, which is a negative number indicating how long the lth class-i customer had been there by time 0. Similarly, index the $Z_i^n(0)$ class-i customers initially in service by $l = -R_i^n(0)-Z_i^n(0)+1, \ldots, -R_i^n(0)$. Index arriving customers by $l = 1, 2, 3, \ldots, E_i^n(t)$, according to the order of arrival within class i, with $a_{i,l}^n$ being the lth arrival time. Let $u_{i,l}^n$ and $v_{i,l}^n$ be the patience and service times (remaining service times for $l \le -R_i^n(0)$) of the lth customer in class i, respectively. Define

$$B_i^n(t) = E_i^n(t) - R_i^n(t). \quad (10)$$

It is clear that the index of the first (earliest arrived) class-i customer in the virtual buffer at time t is $B_i^n(t)+1$. Actually, this is because that at time t the index of the last class-i customer in the virtual buffer is $E_i^n(t)$ and the number of class-i customers in the virtual buffer is $R_i^n(t)$. Moreover, $B_i^n(t) - B_i^n(s)$ can be viewed as the number of customers who would have started service during time interval $(s,t]$ had they been infinitely patient. However, only some of them actually obtained service since others have abandoned by the time they were scheduled to be served. Let $\tau_{i,l}^n$ denote the time when the lth customer of class i can start service. Note that only if $\tau_{i,l}^n - a_{i,l}^n$ is less than the patience time $u_{i,l}^n$, the lth customer will eventually be served. Let δ_x denote the Dirac point measure for any x in $\mathbb{R}$ or $\mathbb{R}\times\mathbb{R}$. For any Borel set $C \subset \mathbb{R}$, let $C + x = \{y + x : y \in C\}$. Denote $C_x = (x,\infty)$ for any $x \in \mathbb{R}$. Using the notations introduced above, we present measure-valued *stochastic dynamic equations* for class-i customers in the virtual buffer and in service.

$$\mathscr{R}_i^n(t)(C) = \sum_{l=B_i^n(t)+1}^{E_i^n(t)} \delta_{u_{i,l}^n}(C + t - a_{i,l}^n), \quad C \subset \mathbb{R}, \quad (11)$$

$$\mathscr{Z}_i^n(t)(C) = \sum_{l=-R_i^n(0)-Z_i^n(0)+1}^{-R_i^n(0)} \delta_{v_{i,l}^n}(C+t) + \sum_{l=-R_i^n(0)+1}^{B_i^n(t)} \delta_{(u_{i,l}^n, v_{i,l}^n)}(C_{\tau_{i,l}^n - a_{i,l}^n}) \times (C + t - \tau_{i,l}^n), \quad C \subset \mathbb{R}_+. \tag{12}$$

Note that the cumulative number of class-i customers who have entered service by time t can be written as

$$A_i^n(t) = \sum_{l=-R_i^n(0)+1}^{B_i^n(t)} \delta_{u_{i,l}^n}(C_{\tau_{i,l}^n - a_{i,l}^n}), \tag{13}$$

which only counts those customers with index up to $B_i^n(t)$, while the virtual buffer includes customers from $B_i^n(t)+1$ to $E_i^n(t)$. Furthermore, the expressions of $L_i^n(t)$ and $S_i^n(t)$ can be recovered from the balance Eqs. (4), (5) and (13).

2.2 Fluid model

We now introduce a fluid model for the system, which consists of a set of equations analogous to that of the stochastic model. For each class $i \in \mathscr{I}$, consider the fluid content entering the buffer at a constant rate λ_i, i.e., $\bar{E}_i(t) = \lambda_i t$. For any $t \geq 0$, let $\bar{\mathscr{R}}_i(t)(C_x)$ denote the amount of fluid content in the virtual buffer of class i with remaining patience time larger than $x \in \mathbb{R}$, and let $\bar{\mathscr{Z}}_i(t)(C_x)$ denote the amount of class-i fluid content in the server pool with remaining service time larger than $x \in \mathbb{R}_+$. Similar to the stochastic model, the total amounts of class-i fluid content in the virtual buffer, in the queue and in service are

$$\bar{R}_i(t) = \bar{\mathscr{R}}_i(\mathbb{R}), \quad \bar{Q}_i(t) = \bar{\mathscr{R}}_i(t)(C_0) \quad \text{and} \quad \bar{Z}_i(t) = \bar{\mathscr{Z}}_i(t)(C_0). \tag{14}$$

Let $\bar{X}_i(t)$ denote the total amount of class-i fluid content in the system. Then for any $t \geq 0$,

$$\sum_{i \in \mathscr{I}} \bar{Z}_i(t) \leq 1, \tag{15}$$

$$\bar{X}_i(t) = \bar{Q}_i(t) + \bar{Z}_i(t), \quad i \in \mathscr{I}. \tag{16}$$

Similar to the stochastic model, we introduce

$$\bar{B}_i(t) = \lambda_i t - \bar{R}_i(t). \tag{17}$$

We can think of $d\bar{B}_i(s)$ as the rate at which the fluid content in the virtual buffer is scheduled to receive service. Again, not all of it is actually served. In fact, for an infinitesimal amount of fluid ready to start service at time s, it has been waiting for

a period of $\frac{\bar{R}_i(s)}{\lambda_i}$. So only a fraction $F_i^c\left(\frac{\bar{R}_i(s)}{\lambda_i}\right)$ actually makes it to the server pool. Thus the fluid process of how class-i customers enter the server pool is

$$\bar{A}_i(t) = \int_0^t F_i^c\left(\frac{\bar{R}_i(s)}{\lambda_i}\right) d\bar{B}_i(s). \tag{18}$$

Replacing the Dirac point measure and summation in the stochastic equations (11) and (12) by the corresponding distribution function and integration, we have the following fluid dynamic equations for the virtual buffer and server pool,

$$\bar{\mathscr{R}}_i(t)(C_x) = \lambda_i \int_{t-\frac{\bar{R}_i(t)}{\lambda_i}}^t F_i^c(x+t-s)ds, \quad x \in \mathbb{R}, \tag{19}$$

$$\bar{\mathscr{Z}}_i(t)(C_x) = \bar{\mathscr{Z}}_i(0)(C_{x+t}) + \int_0^t F_i^c\left(\frac{\bar{R}_i(s)}{\lambda_i}\right) G_i^c(x+t-s)d\bar{B}_i(s), \quad x \in \mathbb{R}_+, \tag{20}$$

where $\bar{\mathscr{Z}}_i(0)(C_{x+t})$ is the amount of class-i fluid initially in service with remaining service time larger than $x+t$ at time 0. Denote by $\bar{L}_i(t)$ and $\bar{S}_i(t)$ the amounts of class-i fluid content that have abandoned queue and completed service by time t, respectively. We have the following balance equations for any $i \in \mathscr{I}$:

$$\bar{Q}_i(t) = \bar{Q}_i(0) + \lambda_i t - \bar{L}_i(t) - \bar{A}_i(t), \tag{21}$$

$$\bar{Z}_i(t) = \bar{Z}_i(0) + \bar{A}_i(t) - \bar{S}_i(t). \tag{22}$$

Corresponding to (6), we have

$$\sum_{i\in\mathscr{I}} \bar{A}_i(s,t) \le \sum_{i\in\mathscr{I}} \bar{E}_i(s,t) + \sum_{i\in\mathscr{I}} \bar{S}_i(s,t), \quad 0 \le s \le t, \tag{23}$$

where $\bar{A}_i(s,t) := \bar{A}_i(t) - \bar{A}_i(s)$ and $\bar{E}_i(s,t)$, $\bar{S}_i(s,t)$ are similarly defined.

Like the stochastic control policies introduced in (7), any fluid process

$$\bar{\pi} = (\bar{A}_i, \bar{L}_i, \bar{S}_i, \bar{X}_i, \bar{Q}_i, \bar{Z}_i)_{i\in\mathscr{I}}$$

will be referred to as a policy for the fluid model. Also, denote by $\bar{\Pi}$ the collection of all policies for the fluid model. For each $\bar{\pi} \in \bar{\Pi}$, we introduce the associated fluid cost

$$\bar{C}_T(\bar{\pi}) = \frac{1}{T}\sum_{i\in\mathscr{I}}\left[\int_0^T c_i\bar{Q}_i(s)ds + \gamma_i\bar{L}_i(T)\right]. \tag{24}$$

2.3 Fluid limits

We aim to asymptotically solve the control problem by considering a sequence of such systems in the many-server heavy-traffic regime, where the size of the system increases to infinity while an individual customer's service and patience time distributions remain fixed. Therefore, we make the following assumptions throughout this paper.

Assumption 1 (*On service and patience time distributions*) For each class $i \in \mathscr{I}$, the service time distribution G_i has a directly integrable density g_i and a finite mean $1/\mu_i$; and the patience time distribution F_i is absolutely continuous and strictly increasing.

The study of asymptotic optimal control policy relies on the analysis of the fluid-scaled process defined as for all $i \in \mathscr{I}$

$$\bar{\mathscr{R}}_i^n(t) = \frac{1}{n}\mathscr{R}_i^n(t), \quad \bar{\mathscr{Z}}_i^n(t) = \frac{1}{n}\mathscr{Z}_i^n(t),$$

for all $t \geq 0$. The fluid-scaled process of the external arrival process is defined in the same way, i.e.,

$$\bar{E}_i^n(t) = \frac{1}{n}E_i^n(t),$$

for all $t \geq 0$. The same scaling also applies to all the other processes R_i^n, B_i^n, A_i^n, L_i^n, S_i^n, X_i^n, Q_i^n and Z_i^n. The following assumption on the initial state and external arrival process is required throughout the paper.

Assumption 2 (*On initial state and arrival process*) The fluid-scaled initial condition satisfies that for all $i \in \mathscr{I}$, as $n \to \infty$,

$$(\bar{\mathscr{R}}_i^n(0), \bar{\mathscr{Z}}_i^n(0)) \Rightarrow (\bar{\mathscr{R}}_i(0), \bar{\mathscr{Z}}_i(0)), \tag{25}$$

$$\mathbb{E}(\bar{Q}_i^n(0)) \to \bar{Q}_i(0), \tag{26}$$

where $\Rightarrow$ denotes weak convergence in Skorohod-J_1 topology and the pair of the limiting initial measures $(\bar{\mathscr{R}}_i(0), \bar{\mathscr{Z}}_i(0))$ has no atoms. The external arrival process satisfies that for all $i \in \mathscr{I}$, as $n \to \infty$,

$$\mathbb{E}(\bar{E}_i^n(t)) \to \lambda_i t, \tag{27}$$

for all $t \geq 0$.

The following theorem shows that fluid models can be used to approximate the stochastic ones without specifying a specific family of policies. It also links the costs of stochastic and fluid models.

Theorem 1 *Given Assumptions 1 and 2, for any sequence of policies* $\{\pi^n : \pi^n \in \Pi^n\}$, *there exists a subsequence* $\{\pi^{n_k} : \pi^{n_k} \in \Pi^{n_k}\}$ *along which we have for all* $i \in \mathscr{I}$ *as* $k \to \infty$,

$$\begin{aligned}&(\bar{\mathscr{R}}_i^{n_k}, \bar{\mathscr{Z}}_i^{n_k}, \bar{R}_i^{n_k}, \bar{B}_i^{n_k}, \bar{A}_i^{n_k}, \bar{S}_i^{n_k}, \bar{L}_i^{n_k}, \bar{Q}_i^{n_k}, \bar{Z}_i^{n_k})\\ &\Rightarrow (\bar{\mathscr{R}}_i, \bar{\mathscr{Z}}_i, \bar{R}_i, \bar{B}_i, \bar{A}_i, \bar{S}_i, \bar{L}_i, \bar{Q}_i, \bar{Z}_i), \qquad (28)\\ &\bar{C}_T^{n_k}(\pi^{n_k}) \to \bar{C}_T(\bar{\pi}), \qquad (29)\end{aligned}$$

for some fluid policy $\bar{\pi}$ *and its associated fluid model satisfying* (14)–(23).

Proof For all $i \in \mathscr{I}$, we establish the tightness of $\bar{\mathscr{R}}_i^n, \bar{\mathscr{Z}}_i^n, \bar{R}_i^n, \bar{B}_i^n, \bar{A}_i^n, \bar{S}_i^n, \bar{L}_i^n, \bar{Q}_i^n, \bar{Z}_i^n$ in Lemmas B.4 and B.5 . Then according to an extended version of the Skorohod representation theorem [see Lemma C.1 of Zhang (2013)], we have that along any convergent subsequence, almost surely, as $k \to \infty$

$$\begin{aligned}&(\bar{\mathscr{R}}_i^{n_k}, \bar{\mathscr{Z}}_i^{n_k}, \bar{R}_i^{n_k}, \bar{B}_i^{n_k}, \bar{A}_i^{n_k}, \bar{S}_i^{n_k}, \bar{L}_i^{n_k}, \bar{Q}_i^{n_k}, \bar{Z}_i^{n_k})\\ &\to (\bar{\mathscr{R}}_i, \bar{\mathscr{Z}}_i, \bar{R}_i, \bar{B}_i, \bar{A}_i, \bar{S}_i, \bar{L}_i, \bar{Q}_i, \bar{Z}_i), i \in \mathscr{I},\end{aligned}$$

for some $\bar{\mathscr{R}}_i \in \mathbf{D}([0,\infty), \mathbf{M})$, $\bar{\mathscr{Z}}_i \in \mathbf{D}([0,\infty), \mathbf{M}_+)$, $\bar{R}_i, \bar{B}_i, \bar{A}_i, \bar{S}_i, \bar{L}_i, \bar{Q}_i, \bar{Z}_i \in \mathbf{D}([0,\infty), \mathbb{R})$. It remains to verify that the above limit satisfies (14)–(23). The fluid dynamic equations (14)–(17) and (21)–(23) can be verified by passing the corresponding stochastic equations to the heavy-traffic limit. On the other hand, (19) and (20) follow from exactly the same argument as Lemma 5.5 in Zhang (2013). Along with the convergent subsequence and letting $C = C_x$ in (12), the fluid-scaled process of the last term in (12) also converges to the last term in (20). The convergence is independent of the choice of the service time distribution G_i. Thus, (18) follows by choosing proper G_i, e.g., $G_i(\cdot) \equiv 0$. Till now we have proven (28).

Now we start to prove (29). For any fixed $T > 0$, it suffices to prove that the two convergent subsequences $\{\bar{Q}_i^{n_k}\}$ and $\{\bar{L}_i^{n_k}\}$ satisfy

$$\mathbb{E}\left(\int_0^T \bar{Q}_i^{n_k}(s)ds\right) \to \int_0^T \bar{Q}_i(s)ds \quad \text{and} \quad \mathbb{E}(\bar{L}_i^{n_k}(T)) \to \bar{L}_i(T) \quad \text{as } k \to \infty. \tag{30}$$

It is straightforward to see that the weak convergence of the queue length processes also implies $\int_0^T \bar{Q}^{n_k}(s)ds \Rightarrow \int_0^T \bar{Q}_i(s)ds$ as $k \to \infty$. From the balance equation (4) we see that

$$\begin{aligned}\int_0^T \bar{Q}_i^n(s)ds &\le \int_0^T [\bar{Q}_i^n(0) + \bar{E}_i^n(s)]ds,\\ \bar{L}_i^n(T) &\le \bar{Q}_i^n(0) + \bar{E}_i^n(T).\end{aligned}$$

According to (26) and (27), the right hand sides of the above inequalities are uniformly integrable. Therefore $\int_0^T \bar{Q}^{n_k}(s)ds$ and $\bar{L}^{n_k}(T)$ are also uniformly integrable. This together with the fact that the fluid limit is deterministic yields (30). This proves (29). □

3 A lower bound on performance

We first establish an asymptotic lower bound in Proposition 1. To present the lower bound, we need to introduce two new functions and an optimization problem. For each $i \in \mathscr{I}$, let

$$F_{i,d}(x) = \int_0^x F_i^c(y)dy, \tag{31}$$

$$H_i(x) = \begin{cases} F_i^c\left(F_{i,d}^{-1}\left(\frac{x}{\lambda_i}\right)\right) & \text{if } 0 \le x < \lambda_i N_{F_i}, \\ 0 & \text{if } x \ge \lambda_i N_{F_i}, \end{cases} \tag{32}$$

where N_{F_i} is the mean patience time, i.e., $N_{F_i} = \int_0^\infty F_i^c(y)dy$, which can be either finite or infinite.

Consider the following optimization problem:

$$\begin{aligned} \text{minimize} \quad & \sum_{i\in\mathscr{I}} [c_i q_i + \gamma_i(\lambda_i - z_i\mu_i)] \\ \text{subject to} \quad & \lambda_i H_i(q_i) = z_i\mu_i, \\ & \sum_{i\in\mathscr{I}} z_i \le 1, \\ & z_i, q_i \ge 0. \end{aligned} \tag{33}$$

Intuitively z_i's can be understood as the amount of service resource that is assigned to class-i customers in the long run. And q_i's are the corresponding queue lengths. First, note that $z_i \le \lambda_i/\mu_i$ for all i since H_i's are continuous and decreasing. This will be useful in proving Lemma 1 and Theorem 5. Second, because $H_i(0) = 1$, the first constraint implies $\lambda_i = z_i\mu_i$ if $q_i = 0$, i.e., no abandonment. If $q_i > 0$ then the first constraint implies that $\lambda_i \ge z_i\mu_i$ since H_i is continuous and decreasing. So $\lambda_i - z_i\mu_i \ge 0$ is interpreted as the abandonment rate of class-i customers. Thus, the objective is to minimize the unit time queue length and abandonment costs by choosing appropriate q_i's and z_i's. Actually, q_i and z_i are mutually determined through the first constraint. The second constraint states that z_i's must be chosen so that the amount of customers in service doesn't exceed the fluid-scaled service capacity 1 (see (15)). Denote by (q^*, z^*) an optimal solution to this optimization problem, and let V^* denote the minimal value. Here, $q^* = (q_1^*, q_2^*, \ldots, q_I^*)$ and $z^* = (z_1^*, z_2^*, \ldots, z_I^*)$. In the case where patience times are exponentially distributed, i.e., $F_i(x) = 1 - \exp(-\theta_i x)$, we have $H_i(x) = 1 - \frac{\theta_i}{\lambda_i}x$. Then the first constraint in (33) becomes $\lambda_i = q_i\theta_i + z_i\mu_i$. So the optimization problem is equivalent to the one in Atar et al. (2010) (see

(15) and Remark 2.1 there). We show that V^* serves as an asymptotic lower bound of all achievable costs in Proposition 1 below. Since F_i is absolutely continuous, it has a density function f_i. Let the hazard rate function associated with patience time distribution F_i be $h_i(x) = \frac{f_i(x)}{F_i^c(x)}$. We need to make the following assumption on the hazard rate function.

Assumption 3 The hazard rate functions h_i's of the patience time distributions are non-increasing.

This assumption implies that H_i is convex. To see this, take the derivative of (32)

$$\frac{d}{dx}H_i(x) = -\frac{f_i\left(F_{i,d}^{-1}\left(\frac{x}{\lambda_i}\right)\right)}{\lambda_i F_i^c\left(F_{i,d}^{-1}\left(\frac{x}{\lambda_i}\right)\right)} = -\frac{1}{\lambda_i}h_i\left(F_{i,d}^{-1}\left(\frac{x}{\lambda_i}\right)\right), \tag{34}$$

where the first equality is due to the fact that $\frac{d}{dx}F_{i,d}^{-1}(x) = \frac{1}{F_i^c(F_{i,d}^{-1}(x))}$, which follows from (31). Since $F_{i,d}^{-1}$ is non-decreasing, so is the derivative of H_i by Assumption 3. The convexity of H_i is required in the proof of Proposition 1.

The reason we need the convexity of H_i is that we want to show V^* is the asymptotic optimal value among all possible policies. If we restrict the policies to those that stabilize the system, i.e., those policies such that the long-term behavior of the system becomes a feasible solution of (33), then the assumption is not needed. In Sect. 5, we will show the long-term behavior the fluid models of the multiclass many-server queueing systems under the policies proposed in Sect. 4.

There is an interesting connection with Bassamboo and Randhawa (2016), who show that even in a single-class model, the FCFS discipline may not be optimal when Assumption 3 fails. Analysis of non-FCFS policies is beyond the scope of this paper. Interested readers may refer to Bassamboo and Randhawa (2016) for how to construct possibly dynamic policies for more general hazard rate functions.

Proposition 1 *Given Assumptions 1, 2 and 3, V^* is an asymptotic lower bound for any sequence of policies $\{\pi^n : \pi^n \in \Pi^n\}$, i.e.,*

$$V^* \le \liminf_{T\to\infty} \liminf_{n\to\infty} \bar{C}_T^n(\pi^n).$$

Proof To prove this result, it suffices to show that for any $\varepsilon > 0$, there exists $T_0 > 0$ such that $\liminf_{n\to\infty} \bar{C}_T^n(\pi^n) \ge V^* - \varepsilon$ for all $T > T_0$. It follows from Theorem 1 that there exists a $\bar{\pi} \in \bar{\Pi}$ such that $\liminf_{n\to\infty} \bar{C}_T^n(\pi^n) = \bar{C}_T(\bar{\pi})$. So it remains to show that

$$\bar{C}_T(\bar{\pi}) \ge V^* - \varepsilon \quad \text{for any } \bar{\pi} \in \bar{\Pi} \tag{35}$$

through the analysis of the fluid model. For any policy $\bar{\pi} \in \bar{\Pi}$, let

$$\bar{q}_i = \frac{1}{T}\int_0^T \bar{Q}_i(s)ds, \quad \bar{z}_i = \frac{1}{T}\int_0^T \bar{Z}_i(s)ds.$$

By (21) and (22), for every $i \in \mathscr{I}$, we have $\frac{1}{T}(\bar{X}_i(T) - \bar{X}_i(0)) = \lambda_i - \frac{1}{T}\bar{L}_i(T) - \frac{1}{T}\bar{S}_i(T)$. Let

$$\kappa_i = \frac{1}{T}(\bar{X}_i(T) - \bar{X}_i(0)) + \frac{1}{T}\bar{S}_i(T) - \bar{z}_i\mu_i = \lambda_i - \bar{z}_i\mu_i - \frac{1}{T}\bar{L}_i(T). \tag{36}$$

By Lemma A.2, there exists a $T_1 > 0$ such that for all $T \geq T_1$,

$$|\kappa_i| \leq \frac{\varepsilon}{2I}. \tag{37}$$

By (55) and the mean value theorem, there exists a $\bar{q}_i' \geq 0$ such that

$$\lambda_i - \frac{1}{T}\bar{L}_i(T) = \lambda_i \frac{1}{T}\int_0^T H_i(\bar{Q}_i(s))ds = \lambda_i H_i(\bar{q}_i').$$

It then follows from the last equation of (36) that $\lambda_i H_i(\bar{q}_i') = \bar{z}_i\mu_i + \kappa_i$. So the pair $(\bar{q}_i', \bar{z}_i)$ satisfies

$$\begin{cases} \lambda_i H_i(\bar{q}_i') = \bar{z}_i\mu_i + \kappa_i, \\ \sum\limits_{i\in\mathscr{I}} \bar{z}_i \leq 1, \\ \bar{z}_i,\ \bar{q}_i' \geq 0, \end{cases}$$

where κ_i satisfies (37). As a result of Lemma A.1 and the above constraints, there exists a T_2 such that for all $T \geq T_2$, we have

$$\sum_{i\in\mathscr{I}} c_i\bar{q}_i' + \gamma_i(\lambda_i - \bar{z}_i\mu_i) \geq V^* - \frac{\varepsilon}{2}. \tag{38}$$

Since $H_i(\cdot)$ is convex under Assumption 3, we can apply Jensen's inequality

$$H_i(\bar{q}_i') = \frac{1}{T}\int_0^T H_i(\bar{Q}_i(s))ds \geq H_i\left(\frac{1}{T}\int_0^T \bar{Q}_i(s)ds\right) = H_i(\bar{q}_i).$$

Note that $H_i(\cdot)$ is decreasing, so the above inequality implies $\bar{q}_i \geq \bar{q}_i'$. This, together with (36), (38) and the definition of $\bar{C}_T(\bar{\pi})$ in (24), yields

$$\bar{C}_T(\bar{\pi}) = \sum_{i\in\mathscr{I}}[c_i\bar{q}_i + \gamma_i(\lambda_i - \bar{z}_i\mu_i - \kappa_i)] \geq V^* - \varepsilon$$

for all large T. This proves (35), thus the result follows. □

The connection established above between the optimal solution of the optimization problem (33) and asymptotic performance provides clues for constructing an optimal policy. In what follows, we construct policies to make the system evolve to a state close to the optimal solution (q^*, z^*) and prove their asymptotic optimality.

4 Control policies

A control policy plays a key role in determining the dynamics of the model. In fact, even with the measure-valued description presented above, the dynamics of the queueing system is still not fully decided without a policy. We now present our main results in this paper to show how policies affect the asymptotic performance.

4.1 Non-idling policies

We present a result asserting the supplementary of all non-idling policies when the system is asymptotically underloaded $\sum_{i\in\mathscr{I}} \lambda_i/\mu_i < 1$. Let $\Pi_N^n \subset \Pi^n$ denote the subset of policies such that the following *non-idling constraint* is satisfied at any time $t \geq 0$,

$$\sum_{i\in\mathscr{I}} Q_i^n(t)\left(n - \sum_{i\in\mathscr{I}} Z_i^n(t)\right) = 0. \tag{39}$$

The non-idling constraint forbids the coexistence of queues (from any class) and idle servers. The following result shows that for underloaded systems the optimal performance can be achieved using *any* non-idling policy.

Theorem 2 *Given Assumptions 1, 2 and* $\sum_{i\in\mathscr{I}} \lambda_i/\mu_i < 1$*, for any sequence of non-idling policies* $\{\pi^n : \pi^n \in \Pi_N^n\}$*, there will be*

$$\liminf_{T\to\infty} \liminf_{n\to\infty} \bar{C}_T^n(\pi^n) = \limsup_{T\to\infty} \limsup_{n\to\infty} \bar{C}_T^n(\pi^n) = 0.$$

Proof It follows from Theorem 1 that for any sequence of non-idling policies $\{\pi^n : \pi^n \in \Pi_N^n\}$ we can always choose a convergent subsequence as the supremum. Therefore, there is a fluid policy $\bar{\pi} \in \bar{\Pi}$ such that $\limsup_{n\to\infty} \bar{C}_T^n(\pi^n) = \bar{C}_T(\bar{\pi})$. The non-idling constraints (39) and (59) imply that the fluid limit $\bar{\pi}$ is a non-idling control policy. Thus, the theorem immediately follows from the results in Theorem 4 in Sect. 5.1. □

4.2 The virtual allocation policy

The more interesting case is where the system is critically loaded or overloaded, i.e., $\sum_{i\in\mathscr{I}} \lambda_i/\mu_i \geq 1$. For the nth system, let $z^n = (z_1^n, \ldots, z_I^n)$ denote an allocation of the n servers to I service groups. The number of servers allocated to group i is a non-negative integer z_i^n and $\sum_{i\in\mathscr{I}} z_i^n = n$. We assume that as $n \to \infty$,

$$\frac{z_i^n}{n} \to z_i \quad \text{for all } i \in \mathscr{I}, \tag{40}$$

where z_i's can be any non-negative numbers as long as $\sum_{i\in\mathscr{I}} z_i = 1$. Recall that the system is (asymptotically) underloaded if $\sum_{i\in\mathscr{I}} \lambda_i/\mu_i < 1$, critically loaded if

$\sum_{i\in\mathscr{I}} \lambda_i/\mu_i = 1$, or overloaded if $\sum_{i\in\mathscr{I}} \lambda_i/\mu_i > 1$. By a slight abuse of terminology, queue i (together with group i) is referred to be (asymptotically) underloaded if $\lambda_i/\mu_i < z_i$, critically loaded if $\lambda_i/\mu_i = z_i$, or overloaded if $\lambda_i/\mu_i > z_i$.

To describe the control policy in detail, let $A^n_{ij}(t)$ and $S^n_{ij}(t)$ denote the cumulative number of class-i customers who have been routed to a group-j server and those who have completed service from a group-j server during the time interval $(0, t]$, respectively. Also, let $Z^n_{ij}(t)$ represent the number of class-i customers being served by group-j servers at time t. Clearly, we have

$$Z^n_i(t) = \sum_{j\in\mathscr{I}} Z^n_{ij}(t), \quad A^n_i(t) = \sum_{j\in\mathscr{I}} A^n_{ij}(t) \text{ and } S^n_i(t) = \sum_{j\in\mathscr{I}} S^n_{ij}(t), \tag{41}$$

and the balance equation for these head-count processes is

$$Z^n_{ij}(t) = Z^n_{ij}(0) + A^n_{ij}(t) - S^n_{ij}(t) \tag{42}$$

for all $i, j \in \mathscr{I}$. We index class-i customers who are routed to group-j servers after time 0 by $l = 1, 2, \ldots, A^n_{ij}(t)$, and index the $Z^n_{ij}(0)$ class-i customers initially being served by group-j servers using $l = -Z^n_{ij}(0) + 1, \ldots, 0$. Then, let $v^n_{ij,l}$ be the (remaining) service time of the lth class-i customer routed to group-j servers for all indices l and $\tau^n_{ij,l}$ be the time when the lth customer starts service for $l \geq 1$. Using these notations, the relations between $Z^n_{ij}(t)$ and $A^n_{ij}(t)$ can be written as

$$Z^n_{ij}(t) = \mathscr{Z}^n_{ij}(0)(C_t) + \sum_{l=1}^{A^n_{ij}(t)} \delta_{v^n_{ij,l}}(C_{t-\tau^n_{ij,l}}), \tag{43}$$

where $\mathscr{Z}^n_{ij}(0)(C_t) = \sum_{l=-Z^n_{ij}(0)+1}^{0} \delta_{v^n_{ij,l}}(C_t)$. And $\mathscr{Z}^n_i(0) = \sum_{j\in\mathscr{I}} \mathscr{Z}^n_{ij}(0)$. For convenience, we introduce

$$Z^n_{\cdot,i}(t) = \sum_{j\in\mathscr{I}} Z^n_{ji}(t), \quad A^n_{\cdot,i}(t) = \sum_{j\in\mathscr{I}} A^n_{ji}(t) \text{ and } S^n_{\cdot,i}(t) = \sum_{j\in\mathscr{I}} S^n_{ji}(t), \tag{44}$$

which represents the number of busy servers in group-i at time t, the number of customers who have been routed to group-i servers by time t, and the number of customers who have completed service from group-i servers by time t, respectively.

We now study the routing process $A^n_{ii}(t)$, $i \in \mathscr{I}$. Class-i customers are routed to group-i servers in two cases: i) upon arrival of a class-i customer if there is an idle group-i server; ii) upon service completion of any group-i server if there is a class-i customer waiting in the queue. Hence,

$$A^n_{ii}(t) = \int_0^t \mathbf{1}_{\{Z^n_{\cdot,i}(s-) < z^n_i\}} dE^n_i(s) + \int_0^t \mathbf{1}_{\{Q^n_i(s-)>0\}} dS^n_{\cdot,i}(s). \tag{45}$$

It remains to consider the process $A^n_{ij}(t)$, $i \neq j$. We should point out that the allocation of the server pool in (40) allows queue i to be underloaded, critically loaded,

or overloaded even if the whole system is overloaded. The process $A^n_{ij}(t)$, $i \neq j$, is devised to satisfy the following two rules (see Sect. 4.2.1 for more intuitive explanations).

- If queue i is underloaded or critically loaded, i.e., $\lambda_i/\mu_i \leq z_i$, then we allow group-i servers to serve other types of customers but don't allow class-i customers to be served by other types of servers.
- If queue i is overloaded, i.e., $\lambda_i/\mu_i > z_i$, then we don't allow group-i servers to serve other types of customers but allow class-i customers to be served by other types of servers.

Mathematically, let $\mathscr{I}_1 = \{i \in \mathscr{I} : \lambda_i/\mu_i > z_i\}$ and $\mathscr{I}_2 = \{i \in \mathscr{I} : \lambda_i/\mu_i \leq z_i\}$. Then following the same idea as in (16) of Atar et al. (2014), we can see from the above two rules that the process A^n_{ij}, $i \neq j$, should satisfy

$$A^n_{ij}(t) = \begin{cases} \int_0^t \mathbf{1}_{\{Z^n_{\cdot,i}(s)=z^n_i,\ Q^n_j(s)=0\}} dA^n_{ij}(s), & \text{when } i \in \mathscr{I}_1 \text{ and } j \in \mathscr{I}_2, \\ 0, & \text{otherwise.} \end{cases} \tag{46}$$

This relation imposes a necessary condition for a class-i customer, $i \in \mathscr{S}_1$, to be sent to service group j, $j \in \mathscr{S}_2$, at time s. Namely, that at time s all servers in group i are busy and no class-j customers are present in the queue.

In stead of (39), the non-idling constraint changes to be at any time $t \geq 0$

$$Q^n_i(t)(z^n_i - Z^n_{\cdot,i}(t)) = 0 \quad \text{for all } i \in \mathscr{I} \tag{47}$$

and

$$Q^n_i(t)(z^n_j - Z^n_{\cdot,j}(t)) = 0 \quad \text{for all } i \in \mathscr{I}_1,\ j \in \mathscr{I}_2. \tag{48}$$

Actually, the above two constraints correspond to (45) and (46). And now we do not have the global non-idling constraint (39). Let $\Pi^n_{vir} \subset \Pi^n$ be the subset of policies such that (40)–(48) are also satisfied. Denote by $\pi^n_{vir}(z^n) \in \Pi^n_{vir}$ the virtual allocation policy with an allocation vector of the server pool being z^n. As a special case of the policy π^n in (7), the virtual allocation policy $\pi^n_{vir}(z^n)$ can be expressed as

$$\pi^n_{vir}(z^n) = (A^n_{ij}, L^n_i, S^n_{ij}, X^n_i, Q^n_i, Z^n_{ij})_{i,j\in\mathscr{I}}. \tag{49}$$

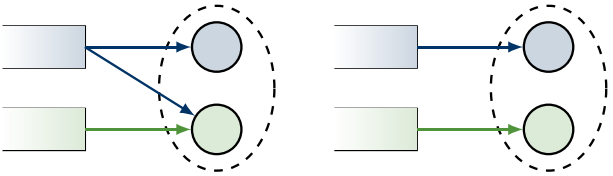
4.2.1 Intuitive explanation for our policy

In this subsection, we use numerical results to demonstrate that allowing proper cross sharing may have a significant impact on the performance of multi-class many-server queueing systems.

We consider an overloaded two-class many-server queue by setting $I = 2$ and $n = 100$. Obviously, there is at least an overloaded subqueue. To this end, we let $\mu^n = (1, 2)$ and $z^n = (20, 80)$. We fix $\lambda^n_1 = 120$ such that queue 1 is always

Table 1 Arrival rate and traffic intensity of queue 2

System	λ_2^n	Queue 2
I	120	Underloaded
II	160	Critically loaded
III	200	Overloaded



(a) N-model **(b)** Two Separate Queues

Fig. 2 The candidate routing networks of a two-class many-server queue under the virtual allocation policy

overloaded. We consider three systems with different values of λ_2^n presented in Table 1 such that queue 2 has all possible traffic intensities.

Customers' patience time distributions are assumed to be $F_i(x) = 1 - \frac{1}{x+1}$ for $i = 1, 2$. We set the holding costs to be $c = (2, 3)$ and the reneging penalties to be $\gamma = (0, 0)$. Assume that both inter-arrival and service times follow Erlang E_2 distributions.

Since queue 1 is fixed to be overloaded in our setting, the two-class many-server queue under the virtual allocation policy may reduce to an N-model [routing network (a)] shown in Fig. 2a or a model with two parallel queues [routing network (b)] shown in Fig. 2b depending on the traffic intensity of queue 2. In details, for Systems I and II, the virtual allocation policy corresponds to routing network (a). And it corresponds to routing network (b) for System III. For each value of λ_2^n, we simulate the system under the two routing networks and we run each simulation long enough by setting $T = 100$. Therefore, we consider 6 systems in total. In Table 2, we present the simulation approximations for $\mathbb{E}[Q_i^n]$'s, $\mathbb{E}[Z_i^n]$'s and the total long-run average cost C_T^n defined in (8) along with their 95% confidence intervals for five independent runs. In the last column titled "Improvement", we display the improvements in the total long-run average cost if the routing network (a) is used instead of the routing network (b).

It is worth noting that the routing networks have a significant effect on the performance of the systems. To illustrate this, we note that the total long-run average cost can be reduced by 38.29% when queue 2 is underloaded and 9.44% when queue 2 is critically loaded. Though we only allocate 20 servers to group 1, we observe that in Systems I and II there are on average 42.156 and 25.909 servers in serving class-1 customers, respectively. This implies that part of servers in group 2 also help to serve class-1 customers due to the effect of cross sharing in the routing network (a). This motivates us to define the first entry of (46). On the other hand, when queue 2 is overloaded, we can observe that the performance of the model (i.e., the values of C_T^n,

Table 2 Comparison of simulation results for two different routing networks

System	Perf. meas.	Routing network (a) N-model	Routing network (b) Two separate queues	Improvement (%)
	C_T^n	265.980 ± 0.602	430.987 ± 1.245	38.29
	$\mathbb{E}[Q_1^n]$	126.392 ± 0.301	215.461 ± 0.622	
I	$\mathbb{E}[Q_2^n]$	4.398 ± 0.015	0.022 ± 0.001	
	$\mathbb{E}[Z_1^n]$	42.156 ± 0.039	20.000 ± 0.000	
	$\mathbb{E}[Z_2^n]$	57.844 ± 0.039	60.009 ± 0.044	
	C_T^n	406.895 ± 0.831	449.307 ± 1.152	9.44
	$\mathbb{E}[Q_1^n]$	184.790 ± 0.416	215.461 ± 0.622	
II	$\mathbb{E}[Q_2^n]$	12.438 ± 0.058	6.128 ± 0.067	
	$\mathbb{E}[Z_1^n]$	25.909 ± 0.041	20.000 ± 0.000	
	$\mathbb{E}[Z_2^n]$	74.091 ± 0.042	77.089 ± 0.029	
	C_T^n	565.734 ± 1.276	566.531 ± 1.253	0.14
	$\mathbb{E}[Q_1^n]$	215.181 ± 0.646	215.461 ± 0.622	
III	$\mathbb{E}[Q_2^n]$	45.124 ± 0.132	45.203 ± 0.097	
	$\mathbb{E}[Z_1^n]$	20.042 ± 0.007	20.000 ± 0.000	
	$\mathbb{E}[Z_2^n]$	79.958 ± 0.006	79.997 ± 0.001	

$\mathbb{E}[Q_i^n]$'s and $\mathbb{E}[Z_i^n]$'s) with routing network (a) is very close to that of the model with routing network (b). This means the effect of cross sharing is indistinctive in this case. Thus, it suggests the second entry of (46).

4.2.2 Optimality of the virtual allocation policy

Since we are modeling the server pool at the group level, similar to Assumption 2, the following condition on the initial status of each group is assumed. Note that we require the fluid measure-valued initial state to be controlled by (50) in addition to that the initial measure converges.

Assumption 4 The fluid-scaled initial condition satisfies $\bar{\mathscr{Z}}_{ij}^n(0) \Rightarrow \bar{\mathscr{Z}}_{ij}(0)$ as $n \to \infty$, for all $i, j \in \mathscr{I}$, where the limit satisfies

$$\bar{\mathscr{Z}}_{ij}(0)'((0,t]) := \frac{d}{dt}\bar{\mathscr{Z}}_{ij}(0)((0,t]) \to 0 \quad \text{as } t \to \infty. \tag{50}$$

To prove the convergence of the system in critically and overloaded cases, we need the following key assumption on the service time distributions. Denote the renewal function of the service time distribution G_i by

$$M_{G_i}(t) = \sum_{n=1}^{\infty} G_i^{n*}(t), \tag{51}$$

where $G_i^{n*}(t)$ is the n-fold convolution of $G_i(t)$ with itself.

Assumption 5 For each $i \in \mathscr{I}$, the renewal function $M_{G_i}(t)$ of the service time distribution G_i is either convex or concave.

The additional assumption certainly creates more restriction but can still be satisfied by a wide range of distributions. It is clear that the exponential distribution is included. Brown (1980) proved that the renewal function is concave if the hazard rate of the distribution is decreasing. According to Problem 16 on page 231 in Karlin and Taylor (1975), the renewal function for the Erlang E_2 distribution is convex. The necessary and sufficient conditions for convexity/concavity of renewal functions are given in Shaked and Zhu (1992).

Theorem 3 *Given Assumptions 1, 2, 4, 5 and* $\sum_{i\in\mathscr{I}} \lambda_i/\mu_i \geq 1$, *if a sequence of allocations* z^n *satisfies* $z^n/n \to z^*$ *as* $n \to \infty$, *then for the sequence of virtual allocation policies* $\{\pi_{vir}^n(z^n)\}$, *there will be*

$$\liminf_{T\to\infty} \liminf_{n\to\infty} \bar{C}_T^n(\pi_{vir}^n(z^n)) = \limsup_{T\to\infty} \limsup_{n\to\infty} \bar{C}_T^n(\pi_{vir}^n(z^n)) = V^*.$$

Note that z^* *is one of the components of an optimal solution* (q^*, z^*) *to the optimization problem* (33).

Proof By Theorem 1, for the sequence of virtual allocation policies $\{\pi^n_{vir}(z^n)\}$ we can always choose a convergent subsequence as the supremum. Thus, there is a fluid policy $\bar{\pi} \in \bar{\Pi}$ such that $\limsup_{n\to\infty} \bar{C}^n_T(\pi^n_{vir}(z^n)) = \bar{C}_T(\bar{\pi})$. Due to the fact that $z^n/n \to z^*$ as $n \to \infty$, Proposition 2 in Sect. 5.2 reveals that the limit $\bar{\pi}$ is a fluid virtual allocation policy with the allocation z^*. Therefore, the definition of $\bar{C}_T(\bar{\pi})$ in (24) and the convergence of the fluid model in Theorem 5 in Sect. 5.2 imply

$$\limsup_{T\to\infty} \limsup_{n\to\infty} \bar{C}^n_T(\pi^n_{vir}(z^n)) = \lim_{T\to\infty} \bar{C}_T(\bar{\pi}) = V^*.$$

By the same reason, we also have $\liminf_{T\to\infty} \liminf_{n\to\infty} \bar{C}^n_T(\pi^n_{vir}(z^n)) = V^*$. Thus, the theorem follows. □

It's worth pointing out that the optimization problem (33) can have multiple solutions. Theorem 3 guarantees that, for any optimal solution to (33), one can find a virtual allocation policy that attains the optimal value V^* asymptotically. If Assumption 3 also holds, then combining Proposition 1 and Theorem 3 yields that

$$\limsup_{T\to\infty} \limsup_{n\to\infty} \bar{C}^n_T(\pi^n_{vir}(z^n)) \le \liminf_{T\to\infty} \liminf_{n\to\infty} \bar{C}^n_T(\pi^n)$$

for all $\pi^n \in \Pi^n$. By Definition 1, this implies that the virtual allocation policy is asymptotically optimal among all the policies Π^n.

4.2.3 Comparison to priority policies

In this subsection, we demonstrate the effectiveness of the proposed virtual allocation policy, which can be used to reduce the system cost as opposed to some baseline policies, e.g., the $c\mu/\theta$ rule and any priority control.

The $c\mu/\theta$ rule may be suboptimal We first demonstrate an example where the $c\mu/\theta$ rule assigns the wrong priority, thus becoming suboptimal. Consider a two-class many-server queue model with uniform patience time distributions $U[0, 2/\theta_i], i = 1, 2$. The expectations of patience times are $1/\theta_i, i = 1, 2$. Suppose that both inter-arrival and service times follow Erlang E_2 distributions. The arrival rate λ, service rate μ and cost c are specified in Table 3. The reneging penalties are assumed to be zero for easy comparison with the $c\mu/\theta$ rule. It is easy to see that class-2 customers should be served with priority according to the $c\mu/\theta$ rule since $c_2\mu_2/\theta_2 > c_1\mu_1/\theta_1$. Using the virtual allocation policy with parameter (z_1, z_2), and by Theorem 5, which states that $q_i = \lambda_i \int_0^{\omega_i} F_i^c(x)dx = \frac{\lambda_i}{\theta_i} - \frac{z_i^2\mu_i^2}{\theta_i\lambda_i}, i = 1, 2$, we can obtain the equilibrium state.

Therefore, the optimization problem (33) becomes

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{2} [c_i q_i + \gamma_i(\lambda_i - z_i \mu_i)] \\ \text{subject to} \quad & q_i = \frac{\lambda_i}{\theta_i} - \frac{z_i^2 \mu_i^2}{\theta_i \lambda_i}, \\ & \sum_{i=1}^{2} z_i \le 1, \\ & z_i, q_i \ge 0. \end{aligned}$$

With the parameters given in Table 3, the above optimization problem can be manually solved and the optimal solution is $z^* = (1, 0)$. Translating this optimal solution to the policy for a specific system implies that we should allocate all servers to group 1. We simulate a system with $n = 100$ and present the results in Table 3. All simulation experiments together with the 95% confidence intervals are based on five independent runs of length $T = 100$. As we can see from this table, the fluid approximation is quite accurate and allocating all servers to serve class 1 (based on the virtual allocation policy) yields a lower cost.

Any priority control may be suboptimal We show that under some circumstances any priority control may not be optimal. As an easy example, we choose patience time distributions with infinite expectations. However, this is not a necessity. Intuitively, priority policies may leave some classes of customers completely unserved, causing the queues of those classes of customers to decrease only due to abandonment. When the patience time distributions exhibits heavy tails, the queues will be extremely long. To illustrate this, let the patience time distribution be $F_i(x) = 1 - \frac{1}{x+1}$ for $i = 1, 2$. As in the previous example, the inter-arrival and service times follow the E_2 distribution. All parameters are specified in Table 4. Again, following from Theorem 5, we can obtain the equilibrium state under the virtual allocation policy with parameter (z_1, z_2),

$$q_i = \lambda_i \log\left(\frac{\lambda_i}{z_i \mu_i}\right), \quad i = 1, 2. \tag{52}$$

We simply need to replace the first constraint in (33) by the above equality and solve the optimization problem. With the parameters given in Table 4, the optimal solution is $z^* = (0.21, 0.79)$ (calculated numerically using Mathemetica). Thus, by Theorem 3, we use the virtual control policy $\pi^n_{vir}((21, 79))$ for the system with $n = 100$. As demonstrated in Table 4, the fluid approximation for the system under the virtual allocation policy exceedingly accurate. Moreover, both priority control policies yield a much larger cost than the virtual allocation policy $\pi^n_{vir}((21, 79))$. In fact, priority policies even fail to stabilize the system, since the fluid equilibrium queue length $q_i \to \infty$ as $z_i \to 0$ according to (52). We can see the instability of priority policies in Fig. 3 as their resulting cost rates exhibit an increasing trend after simulating it for a long time, while the virtual allocation policy is able to stabilize the cost rate at the state predicted by the fluid model.

Table 3 A comparison of fluid approximation with simulation estimates under the virtual allocation policy and the $c\mu/\theta$ rule in two-class $G/GI/100+GI$ queueing models

Two-class $G/GI/100+GI$ model with $c=(1,2)$, $\gamma=(0,0)$ and $T=100$				
$\lambda=(800,200)$	$\mu=(7,8)$		$\theta=(2,4)$	
Interarrival cdf=(E_2, E_2)	$G=(E_2, E_2)$		$F=(U(0,1), U(0,0.5))$	
Perf. meas.	Virtual allocation—$\pi^n((100,0))$		$c\mu/\theta$ rule—assign priority to class 2	
	Simulation	Fluid approximation	Simulation	Fluid approximation
C_T^n	193.378 ± 0.292	193.75	227.405 ± 0.292	l227.734
$\mathbb{E}[Q_1^n]$	93.374 ± 0.282	93.75	226.648 ± 0.291	227.734
$\mathbb{E}[Q_2^n]$	50.002 ± 0.020	50	0.378 ± 0.001	0
$\mathbb{E}[Z_1^n]$	99.997 ± 0.001	100	75.098 ± 0.016	75
$\mathbb{E}[Z_2^n]$	0.003 ± 0.001	0	24.902 ± 0.016	25

Table 4 A comparison of simulation estimates under the virtual allocation policy and two priority control policies in two-class $G/GI/100+GI$ queueing models

Two-class $G/GI/100+GI$ model with $c=(2,3)$, $\gamma=(0,0)$ and $T=100$				
$\lambda=(120,300)$	$\mu=(1,2)$		$\theta=(0,0)$	
Interarrival cdf$=(E_2,E_2)$	$G=(E_2,E_2)$		$F(x)=(1-\frac{1}{x+1},1-\frac{1}{x+1})$	
Perf. meas.	$\pi^n_{vir}((21,79))$		Priority on class 1	Priority on class 2
	Simulated	Approximated	Simulated	Simulated
C^n_T	997.781 ± 0.726	995.40	3761.639 ± 30.266	1354.281 ± 16.873
$\mathbb{E}[Q^n_1]$	209.791 ± 0.278	209.16	21.971 ± 1.997	494.853 ± 9.003
$\mathbb{E}[Q^n_2]$	192.733 ± 0.141	192.36	1239.231 ± 10.031	121.517 ± 4.296
$\mathbb{E}[Z^n_1]$	21.000 ± 0.000	21	99.937 ± 0.035	0.000 ± 0.000
$\mathbb{E}[Z^n_2]$	79.000 ± 0.000	79	0.053 ± 0.034	99.996 ± 0.003

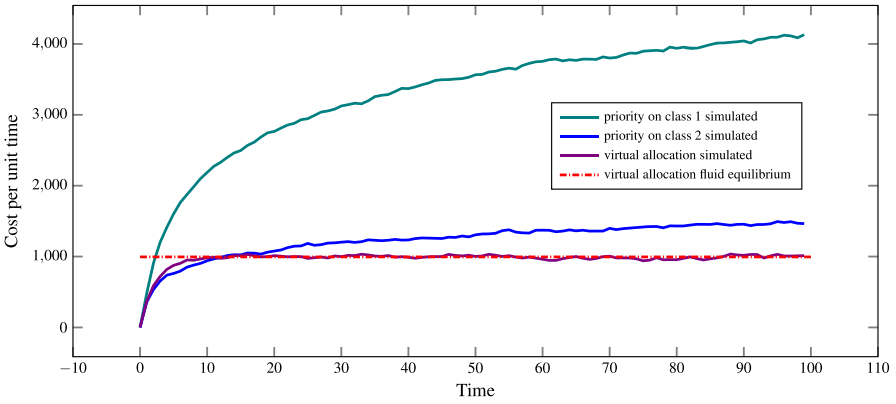


Fig. 3 Cost per unit time under the virtual allocation policy and two priority control policies

5 Convergence to equilibrium states of the fluid models

The idea behind proving the asymptotic optimality results (Theorems 2 and 3) is to analyze the long-term behavior of the corresponding fluid model (proposed in § 2.2). Our analysis of the fluid model is divided into two cases. For underloaded systems, we analyze it under any non-idling policy (see Theorem 4). For critically loaded and overloaded cases, we analyze it under the virtual allocation policy (see Theorem 5), where Assumption 5 is needed.

We first provide some preliminary analysis for each class $i \in \mathscr{I}$ in the fluid model. Taking $x = 0$ in (19), we obtain a relationship between the queue length and the size of virtual buffer,

$$\bar{Q}_i(t) = \lambda_i \int_{t-\frac{\bar{R}_i(t)}{\lambda_i}}^{t} F_i^c(t-s)ds = \lambda_i \int_0^{\frac{\bar{R}_i(t)}{\lambda_i}} F_i^c(s)ds. \tag{53}$$

Plugging (17) to (18) and then applying the definition of H_i (32) and the above equation, we obtain

$$\bar{A}_i(t) = \lambda_i \int_0^t H_i(\bar{Q}_i(s))ds - \bar{Q}_i(t) + \bar{Q}_i(0). \tag{54}$$

Using the balance equation (21), the above equation yields the following relation between the abandonment process $\bar{L}_i$ and the queue length $\bar{Q}_i$:

$$\bar{L}_i(t) = \lambda_i t - \lambda_i \int_0^t H_i(\bar{Q}_i(s))ds. \tag{55}$$

On the other hand, using (18), the fluid dynamic equation (20) can be written as

$$\bar{\mathscr{Z}}_i(t)(C_x) = \bar{\mathscr{Z}}_i(0)(C_{x+t}) + \int_0^t G_i^c(x+t-s)d\bar{A}_i(s). \tag{56}$$

Let $x = 0$ in the above equation. Then

$$\begin{aligned}\bar{Z}_i(t) &= \bar{\mathscr{Z}}_i(0)(C_t) + \int_0^t G_i^c(t-s)d\bar{A}_i(s)\\ &= \bar{\mathscr{Z}}_i(0)(C_t) + \bar{A}_i(t) - \int_0^t \bar{A}_i(t-s)dG_i(s).\end{aligned}$$

We now have a renewal equation, whose solution is

$$\bar{A}_i(t) = \big(\bar{Z}_i(t) - \bar{\mathscr{Z}}_i(0)(C_t)\big) + \big(\bar{Z}_i(t) - \bar{\mathscr{Z}}_i(0)(C_t)\big) * M_{G_i}(t). \tag{57}$$

It is interesting to see that the process $\bar{A}_i(t)$ connects to both $\bar{Q}_i(t)$ and $\bar{Z}_i(t)$ through (54) and (57), respectively. Moreover, if we convolve both sides of (54) and (57) with $G_i^c(t)$, then

$$\begin{aligned}&\bar{Z}_i(t) - \bar{\mathscr{Z}}_i(0)(C_t)\\ &\quad= \lambda_i \int_0^t H_i(\bar{Q}_i(s))ds * G_i^c(t) - \bar{Q}_i(t) + \int_0^t \bar{Q}_i(t-s)dG_i(s) + \bar{Q}_i(0)G_i^c(t).\end{aligned}$$

Performing integration by parts on the above equation, we obtain the following key equation

$$\begin{aligned}\bar{X}_i(t) = \bar{\mathscr{Z}}_i(0)(C_t) + \bar{Q}_i(0)G_i^c(t) + \frac{\lambda_i}{\mu_i}\int_0^t H_i(\bar{Q}_i(t-s))dG_{i,e}(s)\\ + \int_0^t \bar{Q}_i(t-s)dG_i(s),\end{aligned} \tag{58}$$

where $G_{i,e}(\cdot)$ is the associated equilibrium distribution of $G_i(\cdot)$ and is defined by $G_{i,e}(x) = \mu_i \int_0^x G_i^c(y)dy$ for all $x \geq 0$. The key equation (58) will play an important role in proving the convergence of the fluid model.

5.1 Fluid models under fluid non-idling policies

In view of the non-idling constraint (39), we have the following non-idling constraint for the fluid model,

$$\sum_{i\in\mathscr{I}} \bar{Q}_i(t)\big(1 - \sum_{i\in\mathscr{I}} \bar{Z}_i(t)\big) = 0. \tag{59}$$

The following theorem ensures that all queues vanish in the long run as long as we use non-idling policies when the system is underloaded. Let $\bar{\Pi}_N \subset \bar{\Pi}$ be the collection of fluid non-idling control policies given that the above equality holds.

Theorem 4 *Given Assumptions 1, 2 and* $\sum_{i\in\mathscr{I}} \lambda_i/\mu_i < 1$, *for any fluid non-idling policy* $\bar{\pi} \in \bar{\Pi}_N$ *the fluid model satisfies*

$$\lim_{t\to\infty} \bar{Z}_i(t) = \frac{\lambda_i}{\mu_i} \quad \textit{and} \quad \lim_{t\to\infty} \bar{Q}_i(t) = 0 \quad \textit{for all } i \in \mathscr{I}. \tag{60}$$

Consequently,

$$\lim_{T\to\infty} \bar{C}_T(\bar{\pi}) = 0. \tag{61}$$

Proof Let

$$\bar{K}(t) = -\sum_{i\in\mathscr{I}} \bar{Z}_i(t) + \sum_{i\in\mathscr{I}} \left[\bar{\mathscr{Z}}_i(0)(C_t) + \bar{Q}_i(0)G_i^c(t) + \frac{\lambda_i}{\mu_i} \int_0^t H_i(\bar{Q}_i(t-s))dG_{i,e}(s) \right]. \tag{62}$$

Then the key equation (58) can be written as

$$\sum_{i\in\mathscr{I}} \bar{Q}_i(t) = \bar{K}(t) + \sum_{i\in\mathscr{I}} \int_0^t \bar{Q}_i(t-s)dG_i(s). \tag{63}$$

If $\sum_{i\in\mathscr{I}} \bar{Q}_i(t) = 0$, then by (63), $\bar{K}(t) = 0 - \sum_{i\in\mathscr{I}} \int_0^t \bar{Q}_i(t-s)dG_i(s) \le 0$. If $\sum_{i\in\mathscr{I}} \bar{Q}_i(t) > 0$, then $\sum_{i\in\mathscr{I}} \bar{Z}_i(t) = 1$ due to the non-idling constraint (59). Since $\sum_{i\in\mathscr{I}} \lambda_i/\mu_i < 1$ and $H_i(\cdot) \le 1$, we can pick $\delta = (1 - \sum_{i\in\mathscr{I}} \lambda_i/\mu_i)/2$, which is positive, such that

$$\sum_{i\in\mathscr{I}} \left[\frac{\lambda_i}{\mu_i} \int_0^t H_i(\bar{Q}_i(t-s))dG_{i,e}(s) \right] \le 1 - 2\delta.$$

For this given $\delta > 0$, there exists a T_1 such that for all $t > T_1$, $\sum_{i\in\mathscr{I}} [\bar{\mathscr{Z}}_i(0)(C_t) + \bar{Q}_i(0)G_i^c(t)] \le \delta$. Applying the above estimates to (62), we have $\bar{K}(t) \le -1 + \delta + 1 - 2\delta = -\delta$ for all t satisfying $t > T_1$ and $\sum_{i\in\mathscr{I}} \bar{Q}_i(t) > 0$.

Denote by $\mathscr{S} = \{t \ge 0 : \sum_{i\in\mathscr{I}} \bar{Q}_i(t) > 0\}$ the collection of time epochs when the total fluid queue length is larger than 0. Following the discussion of the above two cases, we have that $\bar{K}(t) \le 0$ for any $t \in [T_1, +\infty)$ and $\bar{K}(t) \le -\delta$ for any $t \in \mathscr{S} \cap [T_1, +\infty)$. We show that $m(\mathscr{S}) < \infty$, where m is the Lebesgue measure of real numbers. Suppose the contradictory, i.e., $m(\mathscr{S}) = \infty$. Note that

$$\begin{aligned} \int_0^\infty e^{-yt} \bar{K}(t)dt &= \int_0^{T_1} e^{-yt} \bar{K}(t)dt + \int_{T_1}^\infty e^{-yt} \bar{K}(t)dt \\ &\le \int_0^{T_1} |\bar{K}(t)|dt - \int_{\mathscr{S}\cap[T_1,+\infty)} e^{-yt}\delta dt. \end{aligned} \tag{64}$$

Since we assume $m(\mathscr{S}) = \infty$, there exists a $T_2 > T_1$ such that $\int_{\mathscr{S}\cap[T_1,T_2]} \delta dt = 2 + 2\int_0^{T_1} |\bar{K}(t)|dt$. Choosing $y_0 = \frac{\ln 2}{T_2} > 0$ yields

$$\int_{\mathscr{S}\cap[T_1,+\infty)} e^{-y_0 t}\delta dt \geq e^{-y_0 T_2}\int_{\mathscr{S}\cap[T_1,T_2]} \delta dt = 1 + \int_0^{T_1} |\bar{K}(t)|dt.$$

So we have $\int_0^\infty e^{-y_0 t}\bar{K}(t)dt \leq -1$ due to (64). On the other hand, (63) implies that for all $y > 0$,

$$\int_0^\infty e^{-yt}\sum_{i\in\mathscr{I}}\bar{Q}_i(t)dt = \int_0^\infty e^{-yt}\bar{K}(t)dt \quad + \sum_{i\in\mathscr{I}}\left[\int_0^\infty e^{-yt}\bar{Q}_i(t)dt \cdot \int_0^\infty e^{-yt}dG_i(t)\right].$$

Due to the fact that $\int_0^\infty e^{-yt}dG_i(t) \leq 1$, we must have $\int_0^\infty e^{-yt}\bar{K}(t)dt \geq 0$ for all $y > 0$, which is a contradiction. Hence, we have shown by contradiction that $m(\mathscr{S}) < \infty$.

Since $m(\mathscr{S}) < \infty$, for any $\varepsilon \in (0, 1)$ there exists a $\tau \geq 1$ such that $m(\mathscr{S} \cap [\tau - 1, \infty)) < \varepsilon$. So for any $t \geq \tau$, there exists a $\xi \in [t-\varepsilon, t]$ such that $\sum_{i\in\mathscr{I}} \bar{Q}_i(\xi) = 0$. The balance equation (21) implies

$$\bar{Q}_i(t) \leq \bar{Q}_i(\xi) + \lambda_i\varepsilon = \lambda_i\varepsilon \quad \text{for all } t \geq \tau. \tag{65}$$

Due to the arbitrariness of ε, the above inequality yields $\lim_{t\to\infty}\bar{Q}_i(t) = 0$ for all $i \in \mathscr{I}$. Now by (16) and (58), $\bar{Q}_i(t)$ vanishing to 0 implies the convergence of $\bar{Z}_i(t)$ stated in (60). In view of (32) and (55), we have $\lim_{T\to\infty} \bar{L}_i(T)/T = 0$. Thus, (61) immediately follows from (24) and (60). □

5.2 Fluid models under the fluid virtual allocation policy

We now study the critically loaded and overloaded systems, which is more interesting since server pool keeps busy and control policies make a difference. In view of the virtual allocation policy in Sect. 4.2, we now propose the corresponding fluid version. By (40), $z = (z_1, z_2, \ldots, z_I)$ is actually an allocation of the server pool of the fluid model satisfying $\sum_{i\in\mathscr{I}} z_i = 1$. Similar to Sect. 4.2, for any $i, j \in \mathscr{I}$, denote by $\bar{A}_{ij}(t)$ the cumulative amount of class-i fluid having been routed to group j; and by $\bar{S}_{ij}(t)$ the cumulative amount of class-i fluid having completed service from group j by time t. At time t the amount of class-i fluid being served in group j is denoted by $\bar{Z}_{ij}(t)$. We have the following fluid version balance equations for the above quantities:

$$\bar{Z}_i(t) = \sum_{j\in\mathscr{I}}\bar{Z}_{ij}(t), \quad \bar{A}_i(t) = \sum_{j\in\mathscr{I}}\bar{A}_{ij}(t), \quad \bar{S}_i(t) = \sum_{j\in\mathscr{I}}\bar{S}_{ij}(t), \tag{66}$$

$$\bar{Z}_{ij}(t) = \bar{Z}_{ij}(0) + \bar{A}_{ij}(t) - \bar{S}_{ij}(t). \tag{67}$$

Let the amount of fluid in service group i at time t, the amount of fluid that have been routed to group i and the amount of fluid that has completed service from group i by time t be

$$\bar{Z}_{\cdot,i}(t) = \sum_{j\in\mathscr{I}} \bar{Z}_{ji}(t), \quad \bar{A}_{\cdot,i}(t) = \sum_{j\in\mathscr{I}} \bar{A}_{ji}(t) \text{ and } \bar{S}_{\cdot,i}(t) = \sum_{j\in\mathscr{I}} \bar{S}_{ji}(t), \tag{68}$$

respectively. Note that $\bar{Z}_{\cdot,i}(t) \le z_i$ due to the allocation of the server pool. Since we have expanded the space to look at the detailed status at the group level, we need the following additional equation to describe the fluid model,

$$\bar{Z}_{ij}(t) = \bar{\mathscr{Z}}_{ij}(0)(C_t) + \int_0^t G_i^c(t-s)d\bar{A}_{ij}(s), \tag{69}$$

where $\bar{\mathscr{Z}}_{ij}(0)(C_t)$ is the amount of initial class-i fluid served in group j with remaining service time larger than t.

Following the main idea of the virtual allocation policies, which is to match group-i servers with class-i customers as much as possible, we have the following fluid version of (45) and (46),

$$\begin{aligned}\bar{A}_{ii}(t) = &\int_0^t \mathbf{1}_{\{\bar{Z}_{\cdot,i}(s)<z_i\}}\lambda_i ds + \int_0^t \mathbf{1}_{\{\bar{Q}_i(s)>0\}} d\bar{S}_{\cdot,i}(s) \\ &+ \int_0^t \mathbf{1}_{\{\bar{Z}_{\cdot,i}(s)=z_i,\ \bar{Q}_i(s)=0\}}\big(\lambda_i \wedge \bar{S}'_{\cdot,i}(s)\big)ds\end{aligned} \tag{70}$$

for any $i \in \mathscr{I}$ and

$$\bar{A}_{ij}(t) = \begin{cases} \int_0^t \mathbf{1}_{\{\bar{Z}_{\cdot,i}(s)=z_i,\ \bar{Q}_j(s)=0\}} d\bar{A}_{ij}(s), & \text{when } i\in\mathscr{I}_1 \text{ and } j\in\mathscr{I}_2, \\ 0, & \text{otherwise.}\end{cases} \tag{71}$$

Note that (70) holds almost everywhere because $\bar{S}_{\cdot,i}(\cdot)$ is absolutely continuous as proved in Proposition 2, and hence, differentiable almost everywhere.

The fluid version of (47) and (48) becomes

$$\bar{Q}_i(t)(z_i - \bar{Z}_{\cdot,i}(t)) = 0 \quad \text{for all } i \in \mathscr{I} \tag{72}$$

and

$$\bar{Q}_i(t)(z_j - \bar{Z}_{\cdot,j}(t)) = 0 \quad \text{for all } i \in \mathscr{I}_1,\ j \in \mathscr{I}_2. \tag{73}$$

Let $\bar{\Pi}_{vir} \subset \bar{\Pi}$ be the collection of fluid virtual allocation policies given that (66)–(73) also hold. We call $\bar{\pi}_{vir}(z) \in \bar{\Pi}_{vir}$ the *fluid virtual allocation policy* associated with allocation z. Corresponding to (49), the fluid virtual allocation policy can be expressed

as

$$\bar{\pi}_{vir}(z) = (\bar{A}_{ij}, \bar{L}_i, \bar{S}_{ij}, \bar{X}_i, \bar{Q}_i, \bar{Z}_{ij})_{i,j\in\mathscr{I}}. \tag{74}$$

Example 1 To illustrate the application of the fluid virtual allocation policy, consider a system with two types of customers, i.e. $I = 2$, and satisfying the routing network Fig 2a. Then by the first entry of (71), one can find that

$$\bar{A}_{12}(t) = \int_0^t \mathbf{1}_{\{\bar{Z}_{\cdot,1}(s)=z_1,\ \bar{Q}_2(s)=0\}} d\bar{A}_{12}(s).$$

This example clarifies the fact that the "crossings" (defined by (71)) in the fluid model do not vanish for finite time t.

Proposition 2 *Given Assumptions 1, 2 and 4, if a sequence of allocations* $\{z^n\}$ *satisfy* $z^n/n \to z$ *as* $n \to \infty$, *then under the sequence of virtual allocation policies* $\{\pi^n_{vir}(z^n)\}$ *there exists a subsequence* $\{\pi^{n_k}_{vir}(z^{n_k})\}$ *such that*

$$(\bar{A}^{n_k}_{ij}, \bar{Z}^{n_k}_{ij}, \bar{S}^{n_k}_{ij}, \bar{A}^{n_k}_{\cdot,i}, \bar{Z}^{n_k}_{\cdot,i}, \bar{S}^{n_k}_{\cdot,i}) \Rightarrow (\bar{A}_{ij}, \bar{Z}_{ij}, \bar{S}_{ij}, \bar{A}_{\cdot,i}, \bar{Z}_{\cdot,i}, \bar{S}_{\cdot,i}),$$
$$i, j \in \mathscr{I} \quad as\ k \to \infty,$$

where $\bar{A}_{ij}$, $\bar{Z}_{ij}$, $\bar{S}_{ij}$, $\bar{A}_{\cdot,i}$, $\bar{Z}_{\cdot,i}$ *and* $\bar{S}_{\cdot,i}$ *satisfying* (66)–(73) *are absolutely continuous.*

Proof It follows from Lemma B.6 that $\{\bar{A}^n_{ij}(\cdot)\}$, $\{\bar{Z}^n_{ij}(\cdot)\}$ and $\{\bar{S}^n_{ij}(\cdot)\}$, $i, j \in \mathscr{I}$, are tight. According to an extended version of the Skorohod representation theorem [see Lemma C.1 of Zhang (2013)], we have that along any convergent subsequence, almost surely,

$$(\bar{A}^{n_k}_{ij}, \bar{Z}^{n_k}_{ij}, \bar{S}^{n_k}_{ij}, \bar{A}^{n_k}_{\cdot,i}, \bar{Z}^{n_k}_{\cdot,i}, \bar{S}^{n_k}_{\cdot,i}) \to (\bar{A}_{ij}, \bar{Z}_{ij}, \bar{S}_{ij}, \bar{A}_{\cdot,i}, \bar{Z}_{\cdot,i}, \bar{S}_{\cdot,i}),$$
$$i, j \in \mathscr{I}, \quad \text{as } k \to \infty, \tag{75}$$

for some $\bar{A}_{ij}$, $\bar{Z}_{ij}$, $\bar{S}_{ij}$, $\bar{A}_{\cdot,i}$, $\bar{Z}_{\cdot,i}$ and $\bar{S}_{\cdot,i} \in \mathbf{D}([0,\infty), \mathbb{R})$. It remains to verify that the above limit satisfies (66)–(73). In the following proof, we still use n to index the convergent subsequence for notational simplicity.

First, it is easy to see from the fluid-scaled version of (41), (42) and (44) that $\bar{A}_{ij}$, $\bar{Z}_{ij}$, $\bar{S}_{ij}$, $\bar{A}_{\cdot,i}$, $\bar{Z}_{\cdot,i}$ and $\bar{S}_{\cdot,i}$ satisfy (66), (67) and (68). It follows from (47) and (48) that the limit also satisfies fluid equations (72) and (73).

Let $I^n_i(t) = z^n_i - Z^n_{\cdot,i}(t)$, which is nonnegative and can be interpreted as the idle servers in group i at time t. Then the first entry of (46) is equivalent to $\int_0^t (I^n_i(s) + Q^n_j(s)) dA^n_{ij}(s) = 0$. This together with Lemma 2.4 in Dai and Williams (1996) yields $\int_0^t (\bar{I}_i(s) + \bar{Q}_j(s)) d\bar{A}_{ij}(s) = 0$, where $\bar{I}_i(s) := z_i - \bar{Z}_{\cdot,i}(s)$. Owing to the nonnegativity of $\bar{I}_i$ and $\bar{Q}_j$, it immediately follows that the first entry of (71) holds. This proves (71).

Next, we verify that the limit satisfies (69) and (70). We start to verify (69) and consider the difference

$$\begin{aligned} &|\bar{Z}_{ij}(t) - \bar{\mathscr{Z}}_{ij}(0)(C_t) - \int_0^t G_i^c(t-s)d\bar{A}_{ij}(t)| \\ &\quad \le |\bar{Z}_{ij}(t) - \bar{Z}_{ij}^n(t)| + |\bar{\mathscr{Z}}_{ij}(0)(C_t) - \bar{\mathscr{Z}}_{ij}^n(0)(C_t)| \\ &\quad + |\bar{Z}_{ij}^n(t) - \bar{\mathscr{Z}}_{ij}^n(0)(C_t) - \int_0^t G_i^c(t-s)d\bar{A}_{ij}(t)|. \end{aligned} \tag{76}$$

Let $\{t_k\}_{k=1}^K$ be a partition of the interval $[0, t]$ such that $0 = t_0 < t_1 < \cdots < t_K = t$ and $\max_k(t_{k+1} - t_k) < \delta$ for some $\delta > 0$. On the inverval $[t_k, t_{k+1}]$, for any $\varepsilon > 0$,

$$\begin{aligned} \frac{1}{n}\sum_{l=A_{ij}^n(t_k)+1}^{A_{ij}^n(t_{k+1})} \delta_{v_{ij,l}^n}(C_{t-\tau_{ij,l}^n}) &\le \frac{1}{n}\sum_{l=A_{ij}^n(t_k)+1}^{A_{ij}^n(t_{k+1})} \delta_{v_{ij,l}^n}(C_{t-t_{k+1}}) \\ &\le \bar{A}_{ij}(t_k, t_{k+1})G_i^c(t-t_{k+1}) + \varepsilon, \end{aligned}$$

for all large n, where the first inequality is due to $\tau_{ij,l}^n \le t_{k+1}$, and the second inequality is due to the component of $\bar{A}_{ij}^n$ in (75) and the Glivenko-Cantelli theorem [cf. Theorem 2.47 in Durrett (2010)]. Similarly, we have for all large n,

$$\frac{1}{n}\sum_{l=A_{ij}^n(t_k)+1}^{A_{ij}^n(t_{k+1})} \delta_{v_{ij,l}^n}(C_{t-\tau_{ij,l}^n}) \ge \bar{A}_{ij}(t_k, t_{k+1})G_i^c(t-t_k) - \varepsilon.$$

Note that $\sum_{k=0}^{K-1} G_i^c(t-t_{k+1})\bar{A}_{ij}(t_k, t_{k+1})$ and $\sum_{k=0}^{K-1} G_i^c(t-t_k)\bar{A}_{ij}(t_k, t_{k+1})$ serve as the upper and lower Reimann-Stieltjes sums of the integral $\int_0^t G_i^c(t-s)d\bar{A}_{ij}(s)$, and converge to the integration as the partition size $\delta \to 0$. Due to the above analysis, we have

$$\Big|\frac{1}{n}\sum_{l=1}^{A_{ij}^n(t)} \delta_{v_{ij,l}^n}(C_{t-\tau_{ij,l}^n}) - \int_0^t G_i^c(t-s)d\bar{A}_{ij}(t)\Big| \to 0 \quad \text{as } n \to \infty.$$

Combining this with (43), the component of $\bar{Z}_{ij}^n$ in (75) and the initial condition in Assumption 4, we can conclude that the left side of (76) vanishes as $n \to \infty$. Since the left hand side of (76) is independent of n, it must be equal to 0 implying (69).

Now we use (69) to prove the absolute continuity of the fluid limits. We can derive from (67) and (69) that

$$\begin{aligned} \bar{S}_{ij}(t) &= \bar{\mathscr{Z}}_{ij}(0)((0,t]) + \int_0^t G_i(t-s)d\bar{A}_{ij}(s) \\ &= \bar{\mathscr{Z}}_{ij}(0)((0,t]) + \int_0^t \int_0^x g_i(x-s)d\bar{A}_{ij}(s)dx, \end{aligned} \tag{77}$$

where the second equality follows from changing the order of integration. This together with Assumption 4 implies that $\bar{S}_{ij}(t)$ is absolutely continuous. By (68), $\bar{S}_{\cdot,i}(t)$ is also absolutely continuous. The absolute continuity of $\bar{A}_{ij}(t)$ and $\bar{A}_{\cdot,i}(t)$ follows from (23). Thus, $\bar{Z}_{ij}(t)$ and $\bar{Z}_{\cdot,i}(t)$ are also absolutely continuous by (67) and (68).

It remains to prove (70). To this end, we consider the corresponding fluid-scaled stochastic processes on a small interval $[t, t+\delta]$ in the following three cases. Note that $\bar{Z}_{\cdot,i}(t)$ and $\bar{Q}_i(t)$ are continuous following from Theorem 15.5 in Billingsley (1968) and Lemma B.4.

Case 1 $\bar{Z}_{\cdot,i}(t) < z_i$, then we have $Z^n_{\cdot,i}(s) < z^n_i$, $s \in [t, t+\delta]$, for all large n. Thus (45) implies $\bar{A}^n_{ii}(t, t+\delta) = \bar{E}^n_i(t, t+\delta)$. Therefore (75) yields $\bar{A}'_{ii}(t) = \lambda_i$.

Case 2 $\bar{Q}_i(t) > 0$, then we have $Q^n_i(s) > 0$, $s \in [t, t+\delta]$, for all large n. It again follows from (45) that $\bar{A}^n_{ii}(t, t+\delta) = \bar{S}^n_{\cdot,i}(t, t+\delta)$. Again, by (75), we have $\bar{A}'_{ii}(t) = \bar{S}'_{\cdot,i}(t)$.

Case 3 $\bar{Z}_{\cdot,i}(t) = z_i$ and $\bar{Q}_i(t) = 0$. We follow a similar argument in Theorem 3.2 of Atar et al. (2014) to analyze this case. By Theorem A.6.3 in Dupuis and Ellis (1997), $\bar{Z}'_{\cdot,i}(t) = 0$ almost everywhere (a.e.) on $A_1 := \{t : \bar{Z}_{\cdot,i}(t) = z_i\}$ and $\bar{Q}'_i(t) = 0$ a.e. on $A_2 := \{t : \bar{Q}_i(t) = 0\}$. From (67) and (68) we have $\bar{Z}'_{\cdot,i}(t) = \bar{A}'_{\cdot,i}(t) - \bar{S}'_{\cdot,i}(t)$, and by (21) one can see that $\bar{Q}'_i(t) = \lambda_i - \bar{A}'_i(t) - \bar{L}'_i(t)$. Note that by (55), $\bar{L}'_i(t) = 0$ a.e. on A_2. In view of (71), if $\sum_{j\neq i, j\in\mathscr{I}} \bar{A}_{ij}(t) = 0$, then by (66) $\bar{A}'_{ii}(t) = \bar{A}'_i(t)$. Thus a.e. on $A_1 \cap A_2$, we have $\bar{A}'_{ii}(t) = \bar{A}'_i(t) = \lambda_i$ and $\bar{A}'_{\cdot,i}(t) = \bar{S}'_{\cdot,i}(t)$, where $\bar{A}'_{ii}(t) \le \bar{S}'_{\cdot,i}(t)$. Hence a.e. on $A_1 \cap A_2$, $\bar{A}'_{ii}(t) = \lambda_i \wedge \bar{S}'_{\cdot,i}(t)$. Again by (71), if $\sum_{j\neq i, j\in\mathscr{I}} \bar{A}_{ji}(t) = 0$, then by (68) $\bar{A}'_{ii}(t) = \bar{A}'_{\cdot,i}(t)$. Thus a.e. on $A_1 \cap A_2$, we have $\bar{A}'_i(t) = \lambda_i$ and $\bar{A}'_{ii}(t) = \bar{A}'_{\cdot,i}(t) = \bar{S}'_{\cdot,i}(t)$, where $\bar{A}'_{ii}(t) \le \bar{A}'_i(t)$. Hence a.e. on $A_1 \cap A_2$, $\bar{A}'_{ii}(t) = \lambda_i \wedge \bar{S}'_{\cdot,i}(t)$. Combining the above three cases and the absolutely continuity, we can conclude that the limit satisfies (70). Till now we proved the result. □

Lemma 1 *Given Assumptions 1, 2, 4, 5 and* $\sum_{i\in\mathscr{I}} \lambda_i/\mu_i \ge 1$, *if the allocation of the server pool satisfies* $z_i \le \lambda_i/\mu_i$ *for all* $i \in \mathscr{I}$, *then under the fluid virtual allocation policy* $\bar{\pi}_{vir}(z)$ *we have for all* $i \in \mathscr{I}$,

$$\lim_{t\to\infty} \bar{Z}_{ii}(t) = z_i. \tag{78}$$

Proof Performing integration by parts in (69), we have the following renewal equation

$$\bar{A}_{ij}(t) = \bar{Z}_{ij}(t) - \bar{\mathscr{Z}}_{ij}(0)(C_t) + \int_0^t \bar{A}_{ij}(t-s)dG_i(s), \tag{79}$$

of which the solution is

$$\bar{A}_{ij}(t) = (\bar{Z}_{ij}(t) - \bar{\mathscr{Z}}_{ij}(0)(C_t)) + (\bar{Z}_{ij}(t) - \bar{\mathscr{Z}}_{ij}(0)(C_t)) * M_{G_i}(t), \tag{80}$$

where $M_{G_i}(t)$ is defined in (51). Combining the above with (67) yields

$$\bar{S}_{ij}(t) = \bar{\mathscr{Z}}_{ij}(0)((0,t]) + (\bar{Z}_{ij}(t) - \bar{\mathscr{Z}}_{ij}(0)(C_t)) * M_{G_i}(t). \tag{81}$$

Denote the second term on the right-hand side of the above equation by $\bar{Y}_{ij}(t)$, i.e.,

$$\begin{aligned}\bar{Y}_{ij}(t) &= (\bar{Z}_{ij}(t) - \bar{\mathscr{Z}}_{ij}(0)(C_t)) * M_{G_i}(t), \\ &= \int_0^t M'_{G_i}(t-s)(\bar{Z}_{ij}(s) - \bar{\mathscr{Z}}_{ij}(0)(C_s))ds. \end{aligned} \tag{82}$$

The process $\bar{Y}_{ij}(t)$ can be interpreted as the amount of newly arrived class-i fluid completing service from group j by time t. By (80) and (82),

$$\bar{Z}_{ij}(t) - \bar{\mathscr{Z}}_{ij}(0)(C_t) = \bar{A}_{ij}(t) - \bar{Y}_{ij}(t). \tag{83}$$

Since $M_{G_i}(t)$ is either convex or concave, we may take derivative of the equation (82) to obtain

$$\bar{Y}'_{ij}(t) = M'_{G_i}(0)(\bar{Z}_{ij}(t) - \bar{\mathscr{Z}}_{ij}(0)(C_t)) + \int_0^t \bar{Z}_{ij}(t-s) - \bar{\mathscr{Z}}_{ij}(0)(C_{t-s})dM'_{G_i}(s), \tag{84}$$

for all $i, j \in \mathscr{I}$. Combining (70) and (83) yields

$$(\bar{Z}_{ii}(t) - \bar{\mathscr{Z}}_{ii}(0)(C_t))' = \begin{cases} \lambda_i - \bar{Y}'_{ii}(t), & \bar{Z}_{\cdot,i}(t) < z_i, \\ \bar{S}'_{\cdot,i}(t) - \bar{Y}'_{ii}(t), & \bar{Q}_i(t) > 0, \\ \lambda_i \wedge \bar{S}'_{\cdot,i}(t) - \bar{Y}'_{ii}(t), & \bar{Z}_{\cdot,i}(t) = z_i \text{ and } \bar{Q}_i(t) = 0. \end{cases} \tag{85}$$

The following two claims will be used multiple times in this proof. □

Claim 1 Let a be a constant. If an absolutely continuous function $f : [0, \infty) \to \mathbb{R}$ satisfies 1) $f(0) \geq a$, and 2) $f'(t) \geq 0$ for all t such that $f(t) < a$, then $f(t) \geq a$ for all $t \geq 0$.

Proof Assume to the contrary that there exists a $T > 0$ such that $f(T) < a$. Let $\tau = \sup\{t < T : f(t) \geq a\}$ be the last time that f is larger than or equal to a before T. Then $f(\tau) = a$ and $f(t) < a$ for all $t \in (\tau, T]$. This implies $f'(t) \geq 0$ for all $t \in (\tau, T]$. Therefore $f(T) = f(\tau) + \int_\tau^T f'(t)dt \geq a$, which is a contradiction. Hence, $f(t) \geq a$ for all $t \geq 0$. □

Claim 2 If $\bar{Y}'_{ii}(t) \leq \lambda_i$ then $(\bar{Z}_{ii}(t) - \bar{\mathscr{Z}}_{ii}(0)(C_t))' \geq 0$. Moreover, if there exists a $T \geq 0$ such that $\bar{Y}'_{ii}(t) \leq \lambda_i$ for all $t \geq T$, then the fluid model satisfies

$$\lim_{t\to\infty} \bar{Z}_{ii}(t) = z_i.$$

Proof Note that for the second entry in (85), we always have $\bar{S}'_{\cdot,i}(t) - \bar{Y}'_{ii}(t) \geq 0$ by (81) and (82). The condition $\bar{Y}'_{ii}(t) \leq \lambda_i$ guarantees that the first and third entries are also non-negative. This proves the first part of this claim. And consequently, if $\bar{Y}'_{ii}(t) \leq \lambda_i$ for all $t \geq T$, we have $(\bar{Z}_{ii}(t) - \bar{\mathscr{Z}}_{ii}(0)(C_t))' \geq 0$ for all $t \geq T$. Since

$\bar{Z}_{ii}(t) - \bar{\mathscr{Z}}_{ii}(0)(C_t)$ is bounded by z_i, by the monotone convergence theorem, there exists a $\zeta \in [0, z_i]$ such that $\bar{Z}_{ii}(t) - \bar{\mathscr{Z}}_{ii}(0)(C_t)$ converges to ζ as $t \to \infty$. Since G_i has a directly integrable density g_i and a finite expectation $1/\mu_i$ as assumed in Assumption 1, we can conclude from Theorem 2 in Section XI.3 of Feller (1971) Page 367 that $\mu_i = \lim_{t\to\infty} M'_{G_i}(t)$. It then follows from (84) that

$$\lim_{t\to\infty} \bar{Y}'_{ii}(t) = \zeta\mu_i. \tag{86}$$

We will prove by contradiction that $\zeta = z_i$, which proves this claim. Assume to the contrary, then $\zeta\mu_i < z_i\mu_i \le \lambda_i$. Let $\mathscr{K} = \{t \ge 0 : \bar{A}'_{ii}(t) = \lambda_i\}$. It follows from (70) that

$$\bar{A}'_{ji}(t) = 0, \ \ j \ne i \quad \text{for all } t \notin \mathscr{K}. \tag{87}$$

On the other hand, (83) and (86) imply that $(\bar{Z}_{ii}(t) - \bar{\mathscr{Z}}_{ii}(0)(C_t))' > (\lambda_i - \zeta\mu_i)/2$ for all large $t \in \mathscr{K}$. So the Lebesgue measure of the set $\mathscr{K}$ must satisfy $m(\mathscr{K}) < \infty$ as otherwise $\bar{Z}_{ii}(t) - \bar{\mathscr{Z}}_{ii}(0)(C_t)$ will go to infinity. Consequently, for any $\varepsilon > 0$ there exists a τ such that $m(\mathscr{K} \cap [\tau, +\infty)) < \varepsilon$. So for any $t > \tau + \varepsilon$, there exists a $\xi \in [t - \varepsilon, t]$ such that $\bar{A}'_{ii}(\xi) \ne \lambda_i$. So by (70), we must have $\bar{Z}_{\cdot,i}(\xi) = z_i$. Since $\bar{Z}_{ii}(t)$ converges to ζ, by (68), we have

$$\limsup_{t\to\infty} \sum_{j\ne i, j\in\mathscr{I}} \bar{Z}_{ji}(t) = z_i - \zeta. \tag{88}$$

Let $\mathscr{K}_{\tau+t} := \{s : \tau + t - s \in \mathscr{K} \cap [\tau, \infty)\}$, then $m(\mathscr{K}_{\tau+t}) \le m(\mathscr{K} \cap [\tau, \infty)) < \varepsilon$. Introduce the time-shifted quantities $\bar{Z}_{ji,\tau}(t) := \bar{Z}_{ji}(t+\tau)$, $\bar{A}_{ji,\tau}(t) := \bar{A}_{ji}(t+\tau) - \bar{A}_{ji}(\tau)$. Then by (69) we have

$$\bar{Z}_{ji,\tau}(t) = \bar{\mathscr{Z}}_{ji}(0)(C_{\tau+t}) + \int_0^\tau G_j^c(\tau + t - s)d\bar{A}_{ji}(s) + \int_0^t G_j^c(t - s)d\bar{A}_{ji,\tau}(s), \tag{89}$$

where $j \ne i$. Note that the first two terms on the right-hand side of (89) vanishes as t goes to infinity. When $j \ne i$, the last term on the right-hand side of (89) can be written as $\frac{1}{\mu_j}\int_{\mathscr{K}_{\tau+t}} \bar{A}'_{ji,\tau}(t - s)dG_{j,e}(s)$ due to (87) and the definition of $\mathscr{K}_{\tau+t}$. It can be easily seen from (77) and (84) that $\bar{S}'_{\cdot,i}$ is bounded, and so is $\bar{A}'_{ji,\tau}$. Thus, by Theorem 12.34 in Hewitt and Stromberg (1975) we can choose an ε small enough and a corresponding τ such that the last term on the right-hand side of (89) is less than $\frac{z_i-\zeta}{2I}$. Thus we can see from (89) that $\limsup_{t\to\infty} \sum_{j\ne i, j\in\mathscr{I}} \bar{Z}_{ji}(t) \le (z_i - \zeta)/2$, which contradicts (88). So we must have $\zeta = z_i$, thus proving the claim. □

With the above preparation, we start to prove (78). Let us consider the case where $M_{G_i}(t)$ is convex. This means $M'_{G_i}(t)$ monotone increases to its limit μ_i. It follows

from (84) that for all $t \geq 0$,

$$\bar{Y}'_{ii}(t) \leq z_i M'_{G_i}(0) + z_i \int_0^t dM'_{G_i}(s) = z_i M'_{G_i}(t) \leq \lambda_i.$$

So we have (78) directly following from Claim 2. We now consider the case where $M_{G_i}(t)$ is concave. The concavity of $M_{G_i}(t)$ implies that $M'_{G_i}(t)$ is decreasing. Also we have $M'_{G_i}(0) = g_i(0) < \infty$ following from Assumption 1. It then follows from (84) that for all $t \geq 0$,

$$\bar{Y}'_{ii}(t) \leq M'_{G_i}(0)(\bar{Z}_{ii}(t) - \bar{\mathscr{Z}}_{ii}(0)(C_t)). \tag{90}$$

If $\lambda_i \geq z_i M'_{G_i}(0)$, then by equation (90), $\bar{Y}'_{ii}(t) \leq z_i M'_{G_i}(0) \leq \lambda_i$ for all $t \geq 0$. Again, by Claim 2 we have (78). If $\lambda_i < z_i M'_{G_i}(0)$, then the analysis is more complicated. We study it in the rest of this proof.

To this end, we first prove by induction that

$$\liminf_{t\to\infty}(\bar{Z}_{ii}(t) - \bar{\mathscr{Z}}_{ii}(0)(C_t)) \geq \frac{\lambda_i}{M'_{G_i}(0)} \sum_{i=0}^{N} \left(1 - \frac{\mu_i}{M'_{G_i}(0)}\right)^i, \tag{91}$$

for all $N \in \mathbb{N}$ such that $\frac{\lambda_i}{M'_{G_i}(0)} \sum_{i=0}^{N} \left(1 - \frac{\mu_i}{M'_{G_i}(0)}\right)^i < z_i$. To show that it holds for $N = 0$, we simply need to show there exists a $\tau_0 > 0$ such that

$$\bar{Z}_{ii}(t) - \bar{\mathscr{Z}}_{ii}(0)(C_t) \geq \frac{\lambda_i}{M'_{G_i}(0)} \tag{92}$$

for all $t \geq \tau_0$. Note that whenever (92) does not hold for any $t \geq \tau_0$ we have $\bar{Y}'_{ii}(t) < \lambda_i$ by (90), thus $(\bar{Z}_{ii}(t) - \bar{\mathscr{Z}}_{ii}(0)(C_t))' \geq 0$ by Claim 2. So, according to Claim 1 we only need to show that (92) holds for $t = \tau_0$. Suppose there does not exist such a τ_0, we must have

$$\bar{Z}_{ii}(t) - \bar{\mathscr{Z}}_{ii}(0)(C_t) < \frac{\lambda_i}{M'_{G_i}(0)} \quad \text{for all } t \geq 0. \tag{93}$$

This together with (90) implies $\bar{Y}'_{ii}(t) < \lambda_i$ for all $t \geq 0$. Then by Claim 2, $\bar{Z}_{ii}(t) - \bar{\mathscr{Z}}_{ii}(0)(C_t)$ converges to z_i, which contradicts (93). So (92) holds, and we have shown (91) holds for $N = 0$.

Suppose that (91) happens to be true for a particular value of N, say $N = k$. Then we have

$$\liminf_{t\to\infty}(\bar{Z}_{ii}(t) - \bar{\mathscr{Z}}_{ii}(0)(C_t)) \geq \frac{\lambda_i}{M'_{G_i}(0)} \sum_{i=0}^{k} \left(1 - \frac{\mu_i}{M'_{G_i}(0)}\right)^i.$$

From this, we need to show that the inequality continues to hold for $N = k+1$. Since $M'_{G_i}(t)$ is decreasing due to the concavity, the above inequality implies

$$\begin{aligned}
&\limsup_{t\to\infty} \int_0^t \bar{Z}_{ii}(t-s) - \bar{\mathscr{Z}}_{ii}(0)(C_{t-s}) dM'_{G_i}(s) \\
&= -\liminf_{t\to\infty} \int_0^\infty \bar{Z}_{ii}(t-s) - \bar{\mathscr{Z}}_{ii}(0)(C_{t-s}) d(-M'_{G_i}(s)) \\
&\le -\int_0^\infty \liminf_{t\to\infty} [\bar{Z}_{ii}(t-s) - \bar{\mathscr{Z}}_{ii}(0)(C_{t-s})] d(-M'_{G_i}(s)) \\
&\le -\frac{\lambda_i}{M'_{G_i}(0)} \sum_{i=0}^{k} \left(1 - \frac{\mu_i}{M'_{G_i}(0)}\right)^i (M'_{G_i}(0) - \mu_i) \\
&= -\lambda_i \sum_{i=1}^{k+1} \left(1 - \frac{\mu_i}{M'_{G_i}(0)}\right)^i,
\end{aligned}$$

where the first inequality follows from Fatou's Lemma and the second inequality uses the above inequality. Combining this with (84), we can see that for any $\varepsilon > 0$, there exists a T_1 such that for all $t > T_1$

$$\bar{Y}'_{ii}(t) \le M'_{G_i}(0)(\bar{Z}_{ii}(t) - \bar{\mathscr{Z}}_{ii}(0)(C_t)) - \lambda_i \sum_{i=1}^{k+1} \left(1 - \frac{\mu_i}{M'_{G_i}(0)}\right)^i + \varepsilon. \tag{94}$$

To prove (91) holds for $N = k+1$, it suffices to show that there exists a $\tau_1 \ge T_1$ such that

$$\bar{Z}_{ii}(t) - \bar{\mathscr{Z}}_{ii}(0)(C_t) \ge \frac{\lambda_i}{M'_{G_i}(0)} \sum_{i=0}^{k+1} \left(1 - \frac{\mu_i}{M'_{G_i}(0)}\right)^i - \frac{\varepsilon}{M'_{G_i}(0)} \tag{95}$$

for all $t \ge \tau_1$. By Claim 1 we only need to show that (95) holds for $t = \tau_1$, since whenever (95) does not hold for any $t \ge \tau_1$ we have $\bar{Y}'_{ii}(t) < \lambda_i$ by (94) and thus $(\bar{Z}_{ii}(t) - \bar{\mathscr{Z}}_{ii}(0)(C_t))' \ge 0$ by Claim 2. Suppose there does not exist such a τ_1, we must have

$$\bar{Z}_{ii}(t) - \bar{\mathscr{Z}}_{ii}(0)(C_t) < \frac{\lambda_i}{M'_{G_i}(0)} \sum_{i=0}^{k+1} \left(1 - \frac{\mu_i}{M'_{G_i}(0)}\right)^i - \frac{\varepsilon}{M'_{G_i}(0)}, \tag{96}$$

for all $t \ge T_1$. Substituting the above into (94) yields $\bar{Y}'_{ii}(t) < \lambda_i$ for all $t \ge T_1$. It then follows from Claim 2 that $\lim_{t\to\infty} \bar{Z}_{ii}(t) = z_i$. This contradicts (96). So (95) holds and we have shown that (91) holds for $N = k+1$. Thus the statement (91) is proved by induction.

Now we use (91) to analyze the convergence of (78) when $\lambda_i < z_i M'_{G_i}(0)$. Since $z_i \le \lambda_i/\mu_i$, we have $\frac{\mu_i}{M'_{G_i}(0)} \in (0, 1)$. It is easily seen that

$$\frac{\lambda_i}{M'_{G_i}(0)} \sum_{i=0}^{N} \left(1 - \frac{\mu_i}{M'_{G_i}(0)}\right)^i = \frac{\lambda_i}{\mu_i} \left(1 - \left(1 - \frac{\mu_i}{M'_{G_i}(0)}\right)^{N+1}\right). \tag{97}$$

If $\lambda_i = z_i \mu_i$, then by (91) and (97)

$$\lim_{t\to\infty} (\bar{Z}_{ii}(t) - \bar{\mathscr{Z}}_{ii}(0)(C_t)) = z_i.$$

If $\lambda_i > z_i \mu_i$, then equation (97) implies that there must exists an $N_0 \in \mathbb{N}$ such that

$$\frac{\lambda_i}{M'_{G_i}(0)} \sum_{i=0}^{N_0} \left(1 - \frac{\mu_i}{M'_{G_i}(0)}\right)^i < z_i \le \frac{\lambda_i}{M'_{G_i}(0)} \sum_{i=0}^{N_0+1} \left(1 - \frac{\mu_i}{M'_{G_i}(0)}\right)^i. \tag{98}$$

It follows from (84), (91) and (98) that for any $\varepsilon > 0$ there exists a $T > 0$ such that for all $t \ge T$,

$$\bar{Y}'_{ii}(t) \le M'_{G_i}(0)(\bar{Z}_{ii}(t) - \bar{\mathscr{Z}}_{ii}(0)(C_t)) - \lambda_i \sum_{i=1}^{N_0+1} \left(1 - \frac{\mu_i}{M'_{G_i}(0)}\right)^i + \varepsilon. \tag{99}$$

We show that there exists a $\tau \ge T$ such that

$$\bar{Z}_{ii}(t) - \bar{\mathscr{Z}}_{ii}(0)(C_t) \ge z_i - \frac{\varepsilon}{M'_{G_i}(0)} \tag{100}$$

for all $t \ge T$. Note that whenever (100) does not hold for any $t \ge \tau$, we have $\bar{Y}'_{ii}(t) < \lambda_i$ by (99) and the last inequality in (98). Again, we only need to show that (100) holds for $t = \tau$, by Claims 1 and 2. Suppose there does not exist such a τ, we must have

$$\begin{aligned} \bar{Z}_{ii}(t) - \bar{\mathscr{Z}}_{ii}(0)(C_t) &< z_i - \frac{\varepsilon}{M'_{G_i}(0)} \\ &\le \frac{\lambda_i}{M'_{G_i}(0)} \sum_{i=0}^{N_0+1} \left(1 - \frac{\mu_i}{M'_{G_i}(0)}\right)^i - \frac{\varepsilon}{M'_{G_i}(0)}, \end{aligned} \tag{101}$$

for all $t \ge T$, where the last inequality comes from the second inequality in (98). Plugging the above in (99) yields $\bar{Y}'_{ii}(t) < \lambda_i$ for all $t \ge T$. Then we have Claim 2, which contradicts (101). So (100) holds, implying (78). □

Theorem 5 *Given Assumptions 1, 2, 4, 5 and* $\sum_{i\in\mathscr{I}} \lambda_i/\mu_i \ge 1$, *if the allocation of the server pool satisfies* $z_i \le \lambda_i/\mu_i$ *for all* $i \in \mathscr{I}$, *then under the fluid virtual allocation*

policy $\bar{\pi}_{vir}(z)$ *we have*

$$\lim_{t\to\infty} \bar{Z}_i(t) = z_i \quad \textit{and} \quad \lim_{t\to\infty} \bar{Q}_i(t) = q_i = \lambda_i \int_0^{\omega_i} F_i^c(x)dx \quad \textit{for all } i \in \mathscr{I}, \tag{102}$$

where ω_i *is the solution to* $F_i(\omega_i) = \frac{\lambda_i - z_i\mu_i}{\lambda_i}$. *Consequently,*

$$\lim_{T\to\infty} \bar{C}_T(\bar{\pi}_{vir}(z)) = \sum_{i\in\mathscr{I}} \big(c_i q_i + \gamma_i(\lambda_i - z_i\mu_i)\big). \tag{103}$$

Proof As a consequence of Lemma 1, we have

$$\lim_{t\to\infty} \bar{Z}_{ij}(t) = 0, \quad i \neq j \quad \text{and} \quad \lim_{t\to\infty} \bar{Z}_i(t) = z_i. \tag{104}$$

Deduce from (81) and (84) that

$$\begin{aligned}\bar{S}'_{ij}(t) = \bar{\mathscr{Z}}_{ij}(0)'((0,t]) + M'_{G_i}(0)(\bar{Z}_{ij}(t) - \bar{\mathscr{Z}}_{ij}(0)(C_t)) \\ + \int_0^t \bar{Z}_{ij}(t-s) - \bar{\mathscr{Z}}_{ij}(0)(C_{t-s})dM'_{G_i}(s)\end{aligned}$$

for all $i, j \in \mathscr{I}$. Due to the fact $M_{G_i}(t)$ is either convex or concave, combining the above with (50), (104) and Lemma 1 yields

$$\lim_{t\to\infty} \bar{S}'_{ij}(t) = \begin{cases} z_i\mu_i, & i = j, \\ 0, & i \neq j. \end{cases}$$

From the last terms in (66) and (68), we obtain

$$\lim_{t\to\infty} \bar{S}'_i(t) = \lim_{t\to\infty} \bar{S}'_{\cdot,i}(t) = z_i\mu_i. \tag{105}$$

If $\lambda_i > z_i\mu_i$, then (67), (70) and (105) imply that $\bar{Z}_{\cdot,i}(t) = z_i$ for all sufficiently large t. Applying (105) to the routing process in (70) yields

$$\lim_{t\to\infty} \bar{A}'_{ii}(t) = z_i\mu_i. \tag{106}$$

Note that (106) also holds for $\lambda_i = z_i\mu_i$ following directly from (70) and (105). Now let's consider the rate $\bar{A}'_{ij}(t)$ for all $i, j \in \mathscr{I}$ with $i \neq j$. From (71), if $\bar{A}_{ij}(t) \equiv 0$, then obviously $\bar{A}'_{ij}(t) = 0$. Thus, we suppose $\bar{A}_{ij}(t) \not\equiv 0$. We deduce from (67) that $\bar{A}'_{\cdot,j}(t) \leq \bar{S}'_{\cdot,j}(t)$ whenever $\bar{Z}_{\cdot,j}(t) = z_j$. Consequently, $\bar{A}'_{ij}(t) \leq \bar{S}'_{\cdot,j}(t) - \bar{A}'_{jj}(t)$ due to (68). On the other hand, when $\bar{Z}_{\cdot,j}(t) < z_j$ we have $\bar{Q}_i(t) = 0$ from (73). By (54) this implies $\bar{A}'_i(t) \leq \lambda_i$. Combining this with (66) and (70) yields $\bar{A}'_{ij}(t) \leq$

$(\lambda_i - \bar{S}'_{\cdot,i}(t))^+$. So we have

$$\bar{A}'_{ij}(t) \le \begin{cases} \bar{S}'_{\cdot,j}(t) - \bar{A}'_{jj}(t), & \text{if } \bar{Z}_{\cdot,j}(t) = z_j, \\ (\lambda_i - \bar{S}'_{\cdot,i}(t))^+, & \text{if } \bar{Z}_{\cdot,j}(t) < z_j. \end{cases}$$

If $\lambda_i = z_i\mu_i$, then the above, (105) and (106) imply $\lim_{t\to\infty} \bar{A}'_{ij}(t) = 0$. If $\lambda_i > z_i\mu_i$, then due to the fact that $\lambda_j \ge z_j\mu_j$ and the non-idling constraint (73), similar to the previous analysis there is not only $\bar{Z}_{\cdot,i}(t) = z_i$ but also $\bar{Z}_{\cdot,j}(t) = z_j$ for all large t. Thus we still have $\lim_{t\to\infty} \bar{A}'_{ij}(t) = 0$ from the first entry of the above inequality. Now we can conclude that

$$\lim_{t\to\infty} \bar{A}'_{ij}(t) = \begin{cases} z_i\mu_i, & i = j, \\ 0, & i \ne j. \end{cases}$$

By (66), this implies $\lim_{t\to\infty} \bar{A}'_i(t) = z_i\mu_i$. Next, we prove the convergence of $\bar{Q}_i(t)$ in the following two cases:

Case 1 $\lambda_i = z_i\mu_i$. Assumption 1 implies that ω_i is the unique solution to $F_i(\omega_i) = \frac{\lambda_i - z_i\mu_i}{\lambda_i}$ ($\omega_i = 0$ in this case). It then follows from the definition of $H_i(\cdot)$ in (32) that for any $\varepsilon > 0$ there exists a $\delta > 0$ such that $\lambda_i H_i(\bar{Q}_i(t)) \le z_i\mu_i - \delta$ whenever $\bar{Q}_i(t) \ge \varepsilon$. Owing to the convergence of $\bar{A}'_i(\cdot)$, there exists a $T_0 > 0$ such that for all $t > T_0$, $\bar{A}'_i(t) \ge z_i\mu_i - \delta/2$. Let $\mathscr{L}'(t) = (\bar{Q}_i(t) - 0)^2$, then by (54) for all $t > T_0$

$$\mathscr{L}'(t) = 2(\bar{Q}_i(t) - 0)(\lambda_i H_i(\bar{Q}_i(t)) - \bar{A}'_i(t)) \le -\varepsilon\delta$$

whenever $\bar{Q}_i(t) \ge \varepsilon$. So there must be a $T_1 > 0$ such that $\bar{Q}_i(t) < \varepsilon$ for all $t > T_1$. Since ε can be arbitrarily close to 0, we have $\lim_{t\to\infty} \bar{Q}_i(t) = 0$.

Case 2 $\lambda_i > z_i\mu_i$. Let $\bar{Q}_i(\infty) = \lambda_i \int_0^{\omega_i} F_i^c(x)dx$. Then we have $\lambda_i H_i(\bar{Q}_i(\infty)) = z_i\mu_i$ by (32) and the definition of ω_i. Similar to the previous case, for any $\varepsilon > 0$, there exists a $\delta > 0$ such that

$$\begin{aligned} &\lambda_i H_i(\bar{Q}_i(t)) \le z_i\mu_i - \delta \quad \text{whenever } \bar{Q}_i(t) \ge \bar{Q}_i(\infty) + \varepsilon, \\ &\lambda_i H_i(\bar{Q}_i(t)) \ge z_i\mu_i + \delta \quad \text{whenever } \bar{Q}_i(t) \le \bar{Q}_i(\infty) - \varepsilon. \end{aligned}$$

One can see that there exists a $T_0 > 0$ such that for all $t > T_0$, $z_i\mu_i - \delta/2 \le \bar{A}'_i(t) \le z_i\mu_i + \delta/2$. Let $\mathscr{L}(t) = (\bar{Q}_i(t) - \bar{Q}_i(\infty))^2$, then by (54) for all $t > T_0$

$$\mathscr{L}'(t) = 2(\bar{Q}_i(t) - \bar{Q}_i(\infty))(\lambda_i H_i(\bar{Q}_i(t)) - \bar{A}'_i(t)) \le -\varepsilon\delta,$$

whenever $|\bar{Q}_i(t) - \bar{Q}_i(\infty)| \ge \varepsilon$. So there must be a $T_1 > 0$ such that $\bar{Q}_i(t) \in (\bar{Q}_i(\infty) - \varepsilon, \bar{Q}_i(\infty) + \varepsilon)$ for all $t > T_1$. Due to the arbitrariness of ε, we have $\lim_{t\to\infty} \bar{Q}_i(t) = \bar{Q}_i(\infty)$.

Due to the fact that $\lambda_i H_i(\bar{Q}_i(\infty)) = z_i\mu_i$, we have $\lim_{T\to\infty} \bar{L}_i(T)/T = \lambda_i - z_i\mu_i$ following from (55). This together with (24) and (102) immediately yields (103). This completes the proof. □

6 Conclusion

We have studied a multiclass many-server queueing system with general service and patience time distributions. We have identified a control policy based on a nonlinear program to minimize a combination of holding costs and abandonment penalties. Moreover, we have proven that any non-idling policy is asymptotically optimal when the system is underloaded since the queue and abandonment vanish in the heavy-traffic regime. Our method is based on analyzing the long-term behavior of the fluid model, which arises as the limit of the stochastic systems in the many-server heavy-traffic limit.

There are some directions for future research. (i) In order for the proposed virtual allocation policy to be asymptotically optimal, the patience times must have a decreasing hazard rate. Optimal policies for more general patience times have yet to be identified. Bassamboo and Randhawa (2016) shed light on this issue by considering a single-class model, and showed that it is quite a complicated problem even in the single-class setting. (ii) It remains to establish the convergence of the fluid model for more general service time distributions in which the condition on the convexity/concavity of the renewal functions imposed here are relaxed. (iii) Another interesting question is how to design an optimal policy for systems with time-varying or random arrival rates. For the time-varying system, one can derive a certain Bellman equation and solve it to identify the optimal policy. For system with random arrival rates, one can study robust and dynamic policies.

Acknowledgements The authors thank the editor and the anonymous reviewers for their careful reading of the paper, and for providing constructive feedback. Zhenghua Long's research is supported by NSFC No. 71871114 from the National Natural Science Foundation of China. Jiheng Zhang's research is supported in part by GRF Grants No. 16201417 and 16501015 from Hong Kong Research Grants Council.

Appendix

A Lemmas for the Proof of Proposition 1

Consider the following optimization problem where the constraint is perturbed by a small amount κ_i for each $i \in \mathscr{I}$,

$$\begin{aligned} \text{minimize} \quad & \sum_{i\in\mathscr{I}} [c_i q_i + \gamma_i(\lambda_i - z_i\mu_i)] \\ \text{subject to} \quad & \lambda_i H_i(q_i) = z_i\mu_i + \kappa_i, \ \sum_{i\in\mathscr{I}} z_i \le 1, \ z_i, q_i \ge 0. \end{aligned} \tag{107}$$

We use this problem to perform a sensitivity analysis for the optimization problem (33). Let $\kappa = (\kappa_1, \ldots, \kappa_I)$ and denote by V_κ^* the optimal value of (107).

Lemma A.1 *(Sensitivity Analysis) For any $\varepsilon > 0$ there exists a $\delta > 0$ such that $V_\kappa^* \geq V^* - \varepsilon$ for all κ satisfying* $\max_i |\kappa_i| \leq \delta$.

Proof Since the patience time distribution is strictly increasing following Assumption 1, the function $H_i(\cdot)$ defined by (32) is continuous and strictly decreasing on its support $[0, \lambda_i N_{F_i}]$ due to (34). So there exists a continuous inverse function $H_i^{-1}(\cdot)$. Therefore we can rewrite (33) and (107) as

$$\begin{aligned} \text{minimize} \quad & V(z) = \sum_{i\in\mathscr{I}} [c_i H_i^{-1}(\frac{z_i\mu_i}{\lambda_i}) + \gamma_i(\lambda_i - z_i\mu_i)] \\ \text{subject to} \quad & \sum_{i\in\mathscr{I}} z_i \leq 1,\ z_i\mu_i \leq \lambda_i,\ z_i \geq 0, \end{aligned} \tag{108}$$

and

$$\begin{aligned} \text{minimize} \quad & V_\kappa(z) = \sum_{i\in\mathscr{I}} [c_i H_i^{-1}(\frac{z_i\mu_i}{\lambda_i} + \frac{\kappa_i}{\lambda_i}) + \gamma_i(\lambda_i - z_i\mu_i)] \\ \text{subject to} \quad & \sum_{i\in\mathscr{I}} z_i \leq 1,\ 0 \leq z_i\mu_i + \kappa_i \leq \lambda_i,\ z_i \geq 0. \end{aligned} \tag{109}$$

Suppose $N_{F_i} < \infty$ for all $i \in \mathscr{I}$. By (32), $H_i^{-1}(\cdot)$ is continuous on $[0, 1]$ (thus also uniformly continuous) for all $i \in \mathscr{I}$. Denote by $z^*(\kappa)$ the optimal solution to (109). Then we can find a corresponding $z^\dagger$ in the feasible region of (108) such that $\max |z_i^*(\kappa) - z_i^\dagger| \leq \max |\kappa_i|/\mu_i$. So for any $\varepsilon > 0$ there exists a $\delta > 0$ such that

$$|V_\kappa(z^*(\kappa)) - V(z^\dagger)| \leq \varepsilon,$$

for all κ with $\max |\kappa_i| \leq \delta$. Since $V(z^\dagger) \geq V^*$, the above inequality immediately implies

$$V_\kappa^* \geq V^* - \varepsilon. \tag{110}$$

Suppose $N_{F_{i'}} = \infty$ for some $i' \in \mathscr{I}$. This implies that $H_{i'}^{-1}(0) = \infty$. Thus, $H_{i'}^{-1}(\cdot)$ is no longer continuous at the origin. The main idea in dealing with this case is to shrink the feasible region of (108) and (109) by pushing $z_{i'}$ and $z_{i'}\mu_{i'} + \kappa_{i'}$ away from the lower bound 0 without affecting the optimal value. To this end, we choose $z_i^\ddagger = \min\{\frac{\lambda_i}{2\mu_i}\} \wedge \frac{1}{I}$ for all $i \in \mathscr{I}$. It's easily seen that $z^\ddagger := (z_1^\ddagger, \ldots, z_I^\ddagger)$ is a feasible solution to (108). And there exists a $\delta' > 0$ such that $0 < z_i^\ddagger\mu_i + \kappa_i < \lambda_i$, $i \in \mathscr{I}$, for all κ with $\max |\kappa_i| \leq \delta'$. So $z^\ddagger$ is also a feasible solution to (109). Thus we can choose a large enough $M > 0$ such that $V(z^*) \leq V(z^\ddagger) \leq M$ and $V_\kappa(z^*(\kappa)) \leq V_\kappa(z^\ddagger) \leq M$ for all κ satisfying $\max |\kappa_i| \leq \delta'$. According to the fact that $H_{i'}^{-1}(\cdot)$ is decreasing we can conclude that there exists $\gamma > 0$ (depending on M) such that

$$z_{i'}^* \geq \gamma \quad \text{and} \quad z_{i'}^*(\kappa)\mu_{i'} + \kappa_{i'} \geq \gamma.$$

Now let $|\kappa_{i'}| \le \min\{\gamma/2, \delta'\}$, the last inequality implies

$$z^*_{i'}(\kappa) \ge \frac{\gamma}{2\mu_{i'}}.$$

Borrowing the idea from the cutting-plane method, we can update the corresponding constraints for the i'th class in (108) and (109) with $z_{i'} \ge \min\{\gamma, \frac{\gamma}{2\mu_{i'}}\}$ and $z_{i'}\mu_{i'} + \kappa_{i'} \ge \gamma$ for all κ satisfying $\max|\kappa_i| \le \delta'$ and $|\kappa_{i'}| \le \min\{\gamma/2, \delta'\}$. Then $H^{-1}_{i'}(\cdot)$ again is uniformly continuous on the updated feasible region. So we can apply the same argument as that for the case of $N_{F_i} < \infty$ for all $i \in \mathscr{I}$ to prove the result (110) still holds. □

Lemma A.2 *Given Assumptions 1 and 2, for any policy $\bar{\pi} \in \bar{\Pi}$, the following limits hold for all $i \in \mathscr{I}$*

$$\lim_{T\to\infty}\left|\frac{1}{T}\bar{A}_i(T) - \mu_i\frac{1}{T}\int_0^T \bar{Z}_i(s)ds\right| = 0, \tag{111}$$

$$\lim_{T\to\infty}\left|\frac{1}{T}\bar{S}_i(T) - \mu_i\frac{1}{T}\int_0^T \bar{Z}_i(s)ds\right| = 0, \tag{112}$$

$$\lim_{T\to\infty}\left|\frac{1}{T}\bar{L}_i(T) - \left(\lambda_i - \mu_i\frac{1}{T}\int_0^T \bar{Z}_i(s)ds\right)\right| = 0. \tag{113}$$

Proof Obviously, for any $\varepsilon > 0$ there exists a $T_0 > 0$ such that

$$|\bar{\mathscr{Z}}_i(0)(C_t)| \le \varepsilon \quad \text{for all } t \ge T_0.$$

Combining the above with the renewal theorem yields

$$\frac{1}{t}\int_0^{t-T_0} \bar{\mathscr{Z}}_i(0)(C_{t-s})dM_{G_i}(s) \le \varepsilon\mu_i \quad \text{as } t \to \infty. \tag{114}$$

Therefore

$$\begin{aligned}
&\limsup_{t\to\infty}\frac{1}{t}\int_0^t \bar{\mathscr{Z}}_i(0)(C_{t-s})dM_{G_i}(s)\\
&= \limsup_{t\to\infty}\frac{1}{t}\int_0^{t-T_0} \bar{\mathscr{Z}}_i(0)(C_{t-s})dM_{G_i}(s) + \limsup_{t\to\infty}\frac{1}{t}\int_{t-T_0}^t \bar{\mathscr{Z}}_i(0)(C_{t-s})dM_{G_i}(s)\\
&\le \varepsilon\mu_i + \limsup_{t\to\infty}\frac{1}{t}\int_{t-T_0}^t dM_{G_i}(s)\\
&= \varepsilon\mu_i,
\end{aligned}$$

where the inequality follows from (114) and the fact $\bar{\mathscr{Z}}_i(0)(C_t) \leq 1$. Since ε can be arbitrarily close to 0, we obtain

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t \bar{\mathscr{Z}}_i(0)(C_{t-s}) dM_{G_i}(s) = 0. \tag{115}$$

Since G_i has a directly integrable density g_i and a finite expectation $1/\mu_i$ as assumed in Assumption 1, Theorem 2 in Section XI.3 of Feller (1971) Page 367 shows that for any $\varepsilon > 0$ there exists a $T_1 > 0$ such that

$$|M'_{G_i}(t) - \mu_i| \leq \varepsilon \quad \text{for all } s > T_1. \tag{116}$$

With the help of the above analysis, next we consider $\bar{A}_i(\cdot)$, the process describing how fluid of class i enters service. It follows from (57) that

$$\begin{aligned}
&\limsup_{t\to\infty} \left| \frac{1}{t} \bar{A}_i(t) - \mu_i \frac{1}{t} \int_0^t \bar{Z}_i(s) ds \right| \\
&= \limsup_{t\to\infty} \left| \frac{1}{t} (\bar{Z}_i(t) - \bar{\mathscr{Z}}_i(0)(C_t)) + \frac{1}{t} \int_0^t \bar{Z}_i(t-s) dM_{G_i}(s) \right. \\
&\quad \left. - \frac{1}{t} \int_0^t \bar{\mathscr{Z}}_i(0)(C_{t-s}) dM_{G_i}(s) - \mu_i \frac{1}{t} \int_0^t \bar{Z}_i(s) ds \right| \\
&\overset{(a)}{=} \limsup_{t\to\infty} \left| \frac{1}{t} \int_0^t M'_{G_i}(t-s) \bar{Z}_i(s) ds - \mu_i \frac{1}{t} \int_0^t \bar{Z}_i(s) ds \right| \\
&\overset{(b)}{\leq} \limsup_{t\to\infty} \left[\frac{1}{t} \int_0^{t-T_1} |M'_{G_i}(t-s) - \mu_i| ds + \frac{1}{t} \int_{t-T_1}^t M'_{G_i}(t-s) ds + \frac{1}{t} \int_{t-T_1}^t \mu_i ds \right] \\
&\overset{(c)}{\leq} \varepsilon,
\end{aligned}$$

where (a) is due to (15) and (115); (b) is due to (15); (c) is due to (116). Since ε can be arbitrarily small, we obtain (111).

It follows from (17) that $\bar{R}_i(t) \leq \lambda_i t + \bar{R}_i(0)$. Then by (53) we have $\lim_{t\to\infty} \frac{1}{t} \bar{Q}_i(t) = 0$. Thus, as a consequence of (111), the limits (112) and (113) immediately hold by Eqs. (21) and (22). □

B Tightness of the fluid-scaled processes

B.1 Tightness under any control policy

The main result here is the tightness proved in Lemmas B.4 and B.5. We will present their proofs after proving the following auxiliary lemma.

Lemma B.3 *Given Assumptions 1 and 2, for each $\varepsilon, \eta > 0$ and $T > 0$ there exists an n_0 such that when $n > n_0$,*

$$\mathbb{P}^n\{\max_{i\in\mathscr{I}} \sup_{t\in[0,T]} \left|\bar{Q}_i^n(t) - \lambda_i \int_0^{\omega_i^n(t)} F_i^c(s)ds\right| \le \varepsilon\} \ge 1-\eta, \tag{117}$$

where $\omega_i^n(t)$ is the waiting time of the earliest arrived class-i customer in the virtual buffer at time t.

Proof It immediately follows from Assumption 2 that for each $\varepsilon, \eta > 0$ there exists an n_0 such that for all $n > n_0$,

$$\mathbb{P}^n(\max_{i\in\mathscr{I}} \sup_{0\le s<t\le T} |\bar{E}_i^n(s,t) - \lambda_i(t-s)| \le \frac{\varepsilon}{2}) \ge 1-\eta. \tag{118}$$

Denote the event in (118) by Ω_E^n.

Let $\{t_k\}_{k=0}^K$ be a partition of the interval $[\tau, t]$ such that $\tau = t_0 < t_1 < \cdots < t_K = t$ and $\max_k(t_{k+1} - t_k) < \delta$ for some $\delta > 0$. We can break the sum into K parts,

$$\frac{1}{n}\sum_{l=E_i^n(\tau)+1}^{E_i^n(t)} \delta_{u_{i,l}^n}(C_{t-a_{i,l}^n}) = \sum_{k=0}^{K-1}\frac{1}{n}\sum_{l=E_i^n(t_k)+1}^{E_i^n(t_{k+1})} \delta_{u_{i,l}^n}(C_{t-a_{i,l}^n}).$$

Note that $a_{i,l} \in [t_k, t_{k+1}]$ for all $l \in [E_i^n(t_k)+1, E_i^n(t_{k+1})]$. So on the event Ω_E^n we can apply the Glivenko-Cantelli theorem (cf. Theorem 2.47 in Durrett (2010), Lemma B.1 in Zhang (2013)) to obtain

$$\begin{aligned}
&\frac{1}{n}\sum_{l=E_i^n(t_k)+1}^{E_i^n(t_{k+1})} \delta_{u_{i,l}^n}(C_{t-a_{i,l}^n}) \\
&\le \frac{1}{n}\sum_{l=E_i^n(t_k)+1}^{E_i^n(t_{k+1})} \delta_{u_{i,l}^n}(C_{t-t_{k+1}}) \\
&\le \bar{E}_i^n(t_k, t_{k+1})\nu_{F_i}(C_{t-t_{k+1}}) + \varepsilon \\
&\quad \le \bar{E}_i(t_k, t_{k+1})\nu_{F_i}(C_{t-t_{k+1}}) + 2\varepsilon = \lambda_i \int_{t_k}^{t_{k+1}} F_i^c(t-t_{k+1})ds + 2\varepsilon
\end{aligned}$$

for all large n, where ν_{F_i} is the probability measure of the patience time distribution F_i. Similarly, we can obtain the corresponding inequality of the opposite direction

$$\frac{1}{n}\sum_{l=E_i^n(t_k)+1}^{E_i^n(t_{k+1})} \delta_{u_{i,l}^n}(C_{t-a_{i,l}^n}) \ge \lambda_i \int_{t_k}^{t_{k+1}} F_i^c(t-t_k)ds - 2\varepsilon.$$

Note that $\sum_{k=0}^{K-1} \lambda_i \int_{t_k}^{t_{k+1}} F_i^c(t - t_{k+1})ds$ and $\sum_{k=0}^{K-1} \lambda_i \int_{t_k}^{t_{k+1}} F_i^c(t - t_k)ds$ serve as the upper and lower Reimann sums of the integral $\lambda_i \int_\tau^t F_i^c(t-s)ds$. We can make ε arbitrarily small by making the partition finer. Thus, we can conclude that for any ε_0 there exists an n_0 such that for all $n > n_0$ there is

$$\left|\frac{1}{n} \sum_{l=E_i^n(\tau)+1}^{E_i^n(t)} \delta_{u_{i,l}^n}(C_{t-a_{i,l}^n}) - \lambda_i \int_\tau^t F_i^c(t-s)ds\right| < \varepsilon_0. \tag{119}$$

According to the definition of the virtual buffer we have $\bar{R}_i^n(t) = \bar{E}_i^n(t) - \bar{E}_i^n(t - \omega_i^n(t))$. From (10), this implies

$$\bar{B}_i^n(t) = \bar{E}_i^n(t - \omega_i^n(t)). \tag{120}$$

Plugging $\tau = t - \omega_i^n(t)$ into (119), and by (11) and (120) the result (117) holds. □

Lemma B.4 *Given Assumptions 1 and 2, for any control policy $\pi^n \in \Pi^n$ the sequences of fluid-scaled stochastic processes $\{\bar{A}_i^n\}$, $\{\bar{L}_i^n\}$, $\{\bar{S}_i^n\}$, $\{\bar{B}_i^n\}$, $\{\bar{X}_i^n\}$, $\{\bar{Q}_i^n\}$, $\{\bar{Z}_i^n\}$ and $\{\bar{R}_i^n\}$ for all $i \in \mathscr{I}$ are tight.*

Proof By the convergence of the initial condition (25), for any $\eta > 0$, there exists a compact set $\mathbf{K}_0 \subset \mathbf{M}$ such that

$$\liminf_{n\to\infty} \mathbb{P}^n\{\bar{\mathscr{R}}_i^n(0) \in \mathbf{K}_0 \text{ and } \bar{\mathscr{Z}}_i^n(0) \in \mathbf{K}_0 \text{ for all } i \in \mathscr{I}\} \geq 1 - \eta. \tag{121}$$

Denote the event in the above probability by Ω_0^n. On this event, by (9) and the definition of compact set in the space $\mathbf{M}$ (see Theorem 15.7.5 in Kallenberg (1986)), there exists an $M_0 > 0$ such that

$$\bar{R}_i^n(0) \leq M_0, \quad \bar{Q}_i^n(0) \leq M_0 \text{ and } \bar{Z}_i^n(0) \leq M_0.$$

Clearly, on the event Ω_0^n, we have $\bar{X}_i^n(0) \leq 2M_0$ and $|\bar{B}_i^n(0)| = \bar{R}_i^n(0) \leq M_0$ following from (3) and (10). Additionally, we have $\bar{A}_i^n(0) = \bar{L}_i^n(0) = \bar{S}_i^n(0) = 0$. Thus, condition (i) in Theorem 15.5 of Billingsley (1968) is satisfied by all the sequences of stochastic processes.

Now we turn to analyze the oscillation boundedness. We start by considering the oscillation bound of the sequence of service completion processes $\{\bar{S}_i^n\}$. In view of (5), (12) and (13), $\bar{S}_i^n$ can be recovered as

$$\bar{S}_i^n(t) = \frac{1}{n} \sum_{l=-R_i^n(0)-Z_i^n(0)+1}^{-R_i^n(0)} \delta_{v_{i,l}^n}((0,t]) + \frac{1}{n} \sum_{l=-R_i^n(0)+1}^{B_i^n(t)} \delta_{(u_{i,l}^n, v_{i,l}^n)}(C_{\tau_{i,l}^n - a_{i,l}^n}) \times ((0, t - \tau_{i,l}^n]).$$

It can then be seen from the above and (12) that for any $0 \le s \le t$,

$$\bar{S}_i^n(s,t) = \bar{\mathscr{Z}}_i^n(s)((0,t-s]) + \frac{1}{n}\sum_{l=B_i^n(s)+1}^{B_i^n(t)} \delta_{(u_{i,l}^n, v_{i,l}^n)}(C_{\tau_{i,l}^n - a_{i,l}^n}) \times ((0, t-\tau_{i,l}^n]).$$

In the above equation $l = B_i^n(s)+1, \ldots, B_i^n(t)$, which implies the start service times $\tau_{i,l}^n \in [s,t]$. Combining this with (10), for any $t \in [0,T]$ we have

$$\bar{S}_i^n(s,t) \le \bar{\mathscr{Z}}_i^n(s)((0,t-s]) + \frac{1}{n}\sum_{l=-R_i^n(0)+1}^{E_i^n(T)} \delta_{v_{i,l}^n}((0,t-s]). \tag{122}$$

Following the same argument in Lemma 5.3 of Zhang (2013), we can see that under Assumptions 1 and 2, for each $\varepsilon, \eta > 0$ and $T > 0$ there exists a $\kappa > 0$ (depending on ε and η) such that

$$\liminf_{n\to\infty} \mathbb{P}^n(\max_{i\in\mathscr{I}} \sup_{t\in[0,T]} \sup_{x\in\mathbb{R}_+} \bar{\mathscr{Z}}_i^n(t)([x,x+\kappa]) \le \varepsilon) \ge 1-\eta. \tag{123}$$

Denote the event in (123) by $\Omega_{\text{Reg}}^n(\kappa)$. The first term on the right side of (122) is always bounded by ε on $\Omega_{\text{Reg}}^n(\kappa)$ as long as $t-s < \kappa$. Denote the event in (117) by Ω_Q^n and let

$$\Omega^n = \Omega_0^n \cap \Omega_E^n \cap \Omega_{\text{Reg}}^n(\kappa) \cap \Omega_Q^n.$$

From (123), (118), (117) and (121)

$$\liminf_{n\to\infty} \mathbb{P}^n\{\Omega^n\} \ge 1-\eta.$$

In the remainder of the proof, all random objects are evaluated on a fixed sample path in Ω^n. On this event we have

$$\begin{aligned}\frac{1}{n}\sum_{l=-R_i^n(0)+1}^{E_i^n(T)} \delta_{v_{i,l}^n}((0,t-s]) &\le (\bar{E}_i^n(T) + \bar{R}_i^n(0))\nu_{G_i}((0,t-s]) + \frac{\varepsilon}{2} \\ &\le (M_0 + 2\lambda_i T)\nu_{G_i}((0,t-s]) + \frac{\varepsilon}{2}, \end{aligned} \tag{124}$$

where the first inequality follows from the Glivenko-Cantelli theorem with ν_{G_i} being the probability measure of the service time distribution G_i. Since G_i is absolutely continuous, we can choose $t-s$ small enough such that $\nu_{G_i}((0,t-s]) \le \frac{\varepsilon}{2(M_0+2\lambda_i T)}$. Then by (122) and (124), we have for all large n,

$$\bar{S}_i^n(s,t) \le 2\varepsilon. \tag{125}$$

By the definition of Ω_E^n, when $t - s \leq \frac{\varepsilon}{2\lambda}$ we have for all large n,

$$\bar{E}_i^n(s,t) \leq \varepsilon. \tag{126}$$

Combing the above two inequalities with (23) yields,

$$\bar{A}_i^n(s,t) \leq 3I\varepsilon, \tag{127}$$

as long as $t - s$ is small enough.

If the lth class-i customer abandons the queue in time interval $[s,t]$, the sum of his patience time $u_{i,l}^n$ and arrival time $a_{i,l}^n$ should be in the interval $[s,t]$, i.e., $u_{i,l}^n + a_{i,l}^n \in [s,t]$. Therefore,

$$\bar{L}_i^n(s,t) \leq \frac{1}{n} \sum_{l=-R_i^n(0)+1}^{E_i^n(t)} \delta_{u_{i,l}^n}([s,t] - a_{i,l}^n). \tag{128}$$

Let $\tau = t_0 < t_1 < \cdots < t_K = t$ be a partition of the interval $[\tau, t]$. Then

$$\begin{aligned} \frac{1}{n} \sum_{l=E_i^n(\tau)+1}^{E_i^n(t)} \delta_{u_{i,l}^n}([s,t] - a_{i,l}^n) &= \sum_{k=0}^{K-1} \frac{1}{n} \sum_{l=E_i^n(t_k)+1}^{E_i^n(t_{k+1})} \delta_{u_{i,l}^n}([s,t] - a_{i,l}^n) \\ &\leq \sum_{k=0}^{K-1} \frac{1}{n} \sum_{l=E_i^n(t_k)+1}^{E_i^n(t_{k+1})} \delta_{u_{i,l}^n}([s - t_{k+1}, t - t_k]), \end{aligned} \tag{129}$$

where the last inequality arises because on each sub-interval $[t_k, t_{k+1}]$ those l's to be summed must satisfy $t_k \leq a_{i,l}^n \leq t_{k+1}$. It follows from the Glivenko-Cantelli theorem that

$$\frac{1}{n} \sum_{l=E_i^n(t_k)+1}^{E_i^n(t_{k+1})} \delta_{u_{i,l}^n}([s - t_{k+1}, t - t_k]) \leq (\bar{E}_i^n(t_{k+1}) - \bar{E}_i^n(t_k)) \nu_{F_i}([s - t_{k+1}, t - t_k]) + \frac{\varepsilon}{2K}. \tag{130}$$

Since F_i is absolutely continuous, we can make $t - s$ small enough and the partition fine enough such that

$$\nu_{F_i}([s - t_{k+1}, t - t_k]) \leq \frac{\varepsilon}{2(M_0 + 2\lambda T)}. \tag{131}$$

It then follows from (129)–(131) that

$$\frac{1}{n} \sum_{l=E_i^n(\tau)+1}^{E_i^n(t)} \delta_{u_{i,l}^n}([s,t] - a_{i,l}^n) \leq \frac{\varepsilon}{2(M_0 + 2\lambda T)} [\bar{E}_i^n(t) - \bar{E}_i^n(\tau)] + \frac{\varepsilon}{2}. \tag{132}$$

Recall that $\omega_i^n(t)$ denoted in Lemma B.3 is the waiting time of the earliest arrived class-i customer in the virtual buffer at time t. Therefore $a^n_{i,-R^n(0)+1}$, the arrival time of the earliest arrived class-i customer in the virtual buffer at the initial time point, equals to $0-\omega_i^n(0)$. So (17) and (120) yield $B_i^n(0) = -R_i^n(0) = E_i^n(a^n_{i,-R^n(0)+1})$. Plugging $\tau = a^n_{i,-R^n(0)+1}$ into (132) and combining (128), we obtain

$$\bar{L}_i^n(s,t) \le \frac{\varepsilon}{2(M_0+2\lambda T)}[\bar{E}_i^n(t)+\bar{R}_i^n(0)]+\frac{\varepsilon}{2} \le \varepsilon \tag{133}$$

for all large n. Thus the oscillation bound of $\bar{L}_i^n(\cdot)$ is proved.

By balance equation (4)

$$|\bar{Q}_i^n(t)-\bar{Q}_i^n(s)| \le \bar{E}_i^n(s,t)+\bar{A}_i^n(s,t)+\bar{L}_i^n(s,t) \le 3(I+1)\varepsilon \tag{134}$$

for all large n and small enough $t-s$, where the last inequality is due to (126), (127) and (133). In view of (120), to prove the oscillation bound of $\bar{B}_i^n(\cdot)$ an essential step is to prove that of $\omega_i^n(\cdot)$. By Lemma B.3 and (134), when $t-s$ is small enough we have

$$|\int_0^{\omega_i^n(t)} F_i^c(x)dx - \int_0^{\omega_i^n(s)} F_i^c(x)dx| \le |\frac{\bar{Q}_i^n(t)}{\lambda_i}-\frac{\bar{Q}_i^n(s)}{\lambda_i}|+2\frac{\varepsilon}{\lambda_i} \le 3(I+2)\frac{\varepsilon}{\lambda_i} \tag{135}$$

for all large n. Denote $S_{F_i} = \inf\{x \ge 0 : F(x) = 1\}$. Observe that $F_{i,d}^{-1}(\cdot)$ is continuous on $[0,\infty)$ and $\omega_i^n(\cdot) < S_{F_i}$ since we can remove any customer whose waiting time exceeds S_{F_i} from the virtual buffer (without affecting the dynamics). So for any $\varepsilon_2 > 0$,

$$|\omega_i^n(t)-\omega_i^n(s)| = |F_{i,d}^{-1}\left(\int_0^{\omega_i^n(t)} F_i^c(x)dx\right) - F_{i,d}^{-1}\left(\int_0^{\omega_i^n(s)} F_i^c(x)dx\right)| \le \varepsilon_2$$

as long as ε in (135) is small enough. By (120) and the definition of Ω_E^n, the oscillation of $\bar{B}_i^n(\cdot)$ becomes

$$\begin{aligned}|\bar{B}_i^n(s,t)| &= |\bar{E}_i^n(t-\omega_i^n(t)) - \bar{E}_i^n(s-\omega_i^n(s))| \\ &\le \lambda|t-s-(\omega_i^n(t)-\omega_i^n(s))|+\varepsilon \\ &\le \lambda|t-s|+\lambda\varepsilon_2+\varepsilon,\end{aligned}$$

which can be made smaller than a multiple of ε by choosing $|t-s|$ and ε_2 small enough. Thus the oscillation bound of $\bar{B}_i^n(\cdot)$ is proved.

So we can conclude that the cumulative processes $\bar{A}_i^n$, $\bar{L}_i^n$, $\bar{S}_i^n$ and $\bar{B}_i^n$ all satisfy condition (ii) in Theorem 15.5 of Billingsley (1968). Thus, we have the desired tightness. The tightness of the head-count processes $\bar{X}_i^n$, $\bar{Q}_i^n$, $\bar{Z}_i^n$ and $\bar{R}_i^n$ directly follows from (4), (5) and (10). □

With the help of the tightness of the head-count processes, we now prove the following lemma.

Lemma B.5 *Given Assumptions 1 and 2, for any control policy $\pi^n \in \Pi^n$ the sequence of fluid-scaled measure-valued stochastic processes $\{(\bar{\mathscr{R}}_i^n, \bar{\mathscr{Z}}_i^n), n \in \mathbb{N}\}$, $i \in \mathscr{I}$, are tight.*

Proof When $t - s \leq \frac{\varepsilon}{2\lambda_i}$, by the definition of Ω_E^n we have $\bar{E}_i^n(s,t) \leq \varepsilon$. For any $s < t$ and any Borel set $C \subset \mathbb{R}$, consider the following two cases:
If $E_i^n(s) > B_i^n(t)$, then by (11),

$$\begin{aligned}
&\bar{\mathscr{R}}_i^n(t)(C) - \bar{\mathscr{R}}_i^n(s)(C^\varepsilon) \\
&= -\frac{1}{n} \sum_{l=B_i^n(s)+1}^{B_i^n(t)} \delta_{u_{i,l}^n}(C^\varepsilon + s - a_{i,l}^n) + \frac{1}{n} \sum_{l=E_i^n(s)+1}^{E_i^n(t)} \delta_{u_{i,l}^n}(C + t - a_{i,l}^n) \\
&\quad + \frac{1}{n} \sum_{l=B_i^n(t)+1}^{E_i^n(s)} \left[\delta_{u_{i,l}^n}(C + t - a_{i,l}^n) - \delta_{u_{i,l}^n}(C^\varepsilon + s - a_{i,l}^n) \right] \\
&\leq \bar{E}_i^n(s,t).
\end{aligned}$$

The first term on the right-hand side of the above equation is clearly non-positive since $B_i^n(t)$ is non-decreasing. Note that when $t - s \leq \varepsilon$, $C + t - a_{i,l}^n \subseteq C^\varepsilon + s - a_{i,l}^n$ for all $l \in \mathbb{Z}$, which implies that the third term in the above equation is less than zero. Therefore the inequality follows.
If $E_i^n(s) \leq B_i^n(t)$, then by the definition of $\bar{\mathscr{R}}_i^n(C)$ in (11) that

$$\bar{\mathscr{R}}_i^n(t)(C) - \bar{\mathscr{R}}_i^n(s)(C^\varepsilon) \leq \bar{\mathscr{R}}_i^n(t)(C) \leq \frac{1}{n} \sum_{l=E_i^n(s)+1}^{E_i^n(t)} \delta_{u_{i,l}^n}(C + t - a_{i,l}^n) \leq \bar{E}_i^n(s,t).$$

Therefore, for any case there will always be

$$\bar{\mathscr{R}}_i^n(t)(C) - \bar{\mathscr{R}}_i^n(s)(C^\varepsilon) \leq \bar{E}_i^n(s,t) \leq \varepsilon, \tag{136}$$

as long as $t - s$ is small enough.
On the other hand, when t and s are close enough combining (10) and (136) gives

$$\bar{\mathscr{R}}_i^n(s)(\mathbb{R} \setminus C^\varepsilon) - \bar{\mathscr{R}}_i^n(t)(\mathbb{R} \setminus C) \leq \bar{B}_i^n(s,t) \leq \varepsilon, \tag{137}$$

where the last inequality is owing to the oscillation bound of $\bar{B}_i^n$ shown in the proof of Lemma B.4. It follows from the fact $\mathbb{R} \setminus C \subseteq \{\mathbb{R} \setminus C^\varepsilon\}^{2\varepsilon}$ and C could be any Borel set in $\mathbb{R}$ that the above inequality yields

$$\bar{\mathscr{R}}_i^n(s)(C) - \bar{\mathscr{R}}_i^n(t)(C^{2\varepsilon}) \leq 2\varepsilon. \tag{138}$$

Combining (136) and (138), we have

$$\mathbf{d}[\bar{\mathscr{R}}_i^n(t), \bar{\mathscr{R}}_i^n(s)] \le \varepsilon, \tag{139}$$

where $\mathbf{d}$ is the Prohorov metric defined in (1).

Give a new index $l = A_i^n(s)+1, \ldots, A_i^n(t)$ to class-i customers who enter service in time interval $(s,t]$ according to the time $\tau_{i,l}^n$ at which they start service. It follows from (12) and (13) that

$$\bar{\mathscr{Z}}_i^n(t)(C) = \bar{\mathscr{Z}}_i^n(s)(C+t-s) + \frac{1}{n}\sum_{l=A_i^n(s)+1}^{A_i^n(t)} \delta_{v_{i,l}^n}(C+t-\tau_{i,l}^n).$$

Then

$$\bar{\mathscr{Z}}_i^n(t)(C) - \bar{\mathscr{Z}}_i^n(s)(C+t-s) \le \bar{A}_i^n(s,t) \le \varepsilon,$$

as long as $t-s$ is small enough, where the last inequality holds due to the oscillation bound of $\bar{A}_i^n$ shown in the proof of Lemma B.4. Note that when $t-s \le \varepsilon$, $C+t-s \subseteq C^\varepsilon$. Thus, we have

$$\bar{\mathscr{Z}}_i^n(t)(C) - \bar{\mathscr{Z}}_i^n(s)(C^\varepsilon) \le \bar{A}_i^n(s,t) \le \varepsilon. \tag{140}$$

Similar to (137), by (5) and the above we have

$$\bar{\mathscr{Z}}_i^n(s)(\mathbb{R}_+ \setminus C^\varepsilon) - \bar{\mathscr{Z}}_i^n(t)(\mathbb{R}_+ \setminus C) \le \bar{S}_i^n(s,t) \le \varepsilon,$$

for t and s close enough. Since the above inequality holds for any Borel set $C \subset \mathbb{R}_+$, we can use the same argument as that for (138) to obtain

$$\bar{\mathscr{Z}}_i^n(s)(C) - \bar{\mathscr{Z}}_i^n(t)(C^{2\varepsilon}) \le 2\varepsilon. \tag{141}$$

So (140) and (141) imply that

$$\mathbf{d}[\bar{\mathscr{Z}}_i^n(t), \bar{\mathscr{Z}}_i^n(s)] \le \varepsilon. \tag{142}$$

The oscillation bound condition in Theorem 3.7.2 of Ethier and Kurtz (1986) follows from (139) and (142). The compact containment property can be verified using the same argument in Lemma 5.1 of Zhang (2013). Thus $\bar{\mathscr{R}}_i^n$ and $\bar{\mathscr{Z}}_i^n$ are tight. □

B.2 Tightness under virtual allocation policies

Lemma B.6 *Given Assumptions 1, 2 and 4, for any sequence of virtual allocation policies* $\{\pi^n(z^n)\}$*, the sequences of fluid-scaled stochastic processes* $\{\bar{A}_{ij}^n(\cdot)\}$, $\{\bar{Z}_{ij}^n(\cdot)\}$ *and* $\{\bar{S}_{ij}^n(\cdot)\}$, $i, j \in \mathscr{I}$, *are tight.*

Proof As we can see, condition (i) in Theorem 15.5 of Billingsley (1968) holds due to the fact that $\bar{A}_{ij}^n(0) = \bar{S}_{ij}^n(0) = 0$. Each class-$i$ customer having completed service from service group j must be counted as one instance of service completions of class-i customers in the whole server pool. Therefore we have $\bar{S}_{ij}^n(s,t) \leq \bar{S}_i^n(s,t)$. Similarly, $\bar{A}_{ij}^n(s,t) \leq \bar{A}_i^n(s,t)$. Note that $\bar{S}_{ij}^n(\cdot)$ and $\bar{A}_{ij}^n(\cdot)$ are also nondecreasing. Thus the oscillation bounds of $\bar{S}_{ij}^n$ and $\bar{A}_{ij}^n$ are implied by those of $\bar{S}_i^n$ and $\bar{A}_i^n$, which have been proven in Lemma B.4. So $\bar{S}_{ij}^n$ and $\bar{A}_{ij}^n$ are tight. Due to the balance equation in (42), the tightness of $\{\bar{Z}_{ij}^n(t)\}$ immediately follows. □

References

Ata B, Tongarlak MH (2013) On scheduling a multiclass queue with abandonments under general delay costs. Queueing Syst 74(1):65–104

Atar R (2005) Scheduling control for queueing systems with many servers: asymptotic optimality in heavy traffic. Ann Appl Probab 15(4):2606–2650

Atar R, Mandelbaum A, Reiman MI (2004) Scheduling a multi class queue with many exponential servers: asymptotic optimality in heavy traffic. Ann Appl Probab 14(3):1084–1134

Atar R, Giat C, Shimkin N (2010) The $c\mu/\theta$ rule for many server queues with abandonment. Oper Res 58(5):1427–1439

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

Atar R, Kaspi H, Shimkin N (2014) Fluid limits for many-server systems with reneging under a priority policy. Math Oper Res 39(3):672–696

Bassamboo A, Randhawa RS (2010) On the accuracy of fluid models for capacity sizing in queueing systems with impatient customers. Oper Res 58(5):1398–1413

Bassamboo A, Randhawa RS (2013) Using estimated patience levels to optimally schedule customers. Technical report, Northwestern University and USC

Bassamboo A, Randhawa RS (2016) Scheduling homogeneous impatient customers. Manag Sci 62(7):2129–2147

Bassamboo A, Harrison JM, Zeevi A (2006) Design and control of a large call center: asymptotic analysis of an lp-based method. Oper Res 54(3):419–435

Billingsley P (1968) Convergence of probability measures. Wiley series in probability and statistics: probability and statistics. Wiley, New York

Billingsley P (1999) Convergence of probability measures, 2nd edn. Wiley series in probability and statistics. Wiley, New York

Brown M (1980) Bounds, inequalities, and monotonicity properties for some specialized renewal processes. Ann Probab 8:227–240

Dai JG, Tezcan T (2008) Optimal control of parallel server systems with many servers in heavy traffic. Queueing Syst 59(2):95–134

Dai JG, Williams RJ (1996) Existence and uniqueness of semimartingale reflecting brownian motions in convex polyhedrons. Theory Probab Appl 40(1):1–40

Dupuis P, Ellis RS (1997) A weak convergence approach to the theory of large deviations. Wiley series in probability and statistics. Wiley, New York

Durrett R (2010) Probability: theory and examples, 4th edn. Cambridge University Press, Cambridge

Ethier SN, Kurtz TG (1986) Markov processes. Wiley series in probability and mathematical statistics: probability and mathematical statistics. Wiley, New York

Feller W (1971) An introduction to probability theory and its applications, vol II, 2nd edn. Wiley, New York

Gurvich I, Whitt W (2009) Queue-and-idleness-ratio controls in many-server service systems. Math Oper Res 34(2):363–396

Halfin S, Whitt W (1981) Heavy-traffic limits for queues with many exponential servers. Oper Res 29(3):567–588

Harrison JM, López MJ (1999) Heavy traffic resource pooling in parallel-server systems. Queueing Syst 33(4):339–368
Hewitt E, Stromberg K (1975) Real and abstract analysis. Springer, New York
Kallenberg O (1986) Random measures, 4th edn. Akademie-Verlag, Berlin
Kang W, Ramanan K (2010) Fluid limits of many-server queues with reneging. Ann Appl Probab 20(6):2204–2260
Karlin S, Taylor HM (1975) A first course in stochastic processes, 2nd edn. Academic Press Inc., New York
Kaspi H, Ramanan K (2011) Law of large numbers limits for many-server queues. Ann Appl Probab 21(1):33–114
Kim J, Ward AR (2013) Dynamic scheduling of a $GI/GI/1+GI$ queue with multiple customer classes. Queueing Syst 75(2–4):339–384
Liu Y, Whitt W (2012a) The $G_t/GI/s_t+GI$ many-server fluid queue. Queueing Syst 71(4):405–444
Liu Y, Whitt W (2012b) A many-server fluid limit for the queueing model experiencing periods of overloading. Oper Res Lett 40(5):307–312
Long Z, Zhang J (2014) Convergence to equilibrium states for fluid models of many-server queues with abandonment. Oper Res Lett 42(6–7):388–393
Long Z, Zhang J (2019) A note on many-server fluid models with time-varying arrivals. Probab Eng Inf Sci 33(3):417–437
Mandelbaum A, Stolyar AL (2004) Scheduling flexible servers with convex delay costs: heavy-traffic optimality of the generalized $c\mu$-rule. Oper Res 52(6):836–855
Mandelbaum A, Massey WA, Reiman MI (1998) Strong approximations for markovian service networks. Queueing Syst 30(1/2):149–201
Perry O, Whitt W (2011) A fluid approximation for service systems responding to unexpected overloads. Oper Res 59(5):1159–1170
Shaked M, Zhu H (1992) Some results on block replacement policies and renewal theory. J Appl Probab 29(4):932–946
van Mieghem JA (1995) Dynamic scheduling with convex delay costs: the generalized $c\mu$ rule. Ann Appl Probab 5(3):809–833
Whitt W (2004) Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. Manag Sci 50(10):1449–1461
Whitt W (2006) Fluid models for multiserver queues with abandonments. Oper Res 54(1):37–54
Zhang J (2013) Fluid models of many-server queues with abandonment. Queueing Syst 73(2):147–193