Customer Service Chat Systems with General Service and Patience Times

Zhenghua Long
The Hong Kong University of Science and Technology
zlong@connect.ust.hk

Tolga Tezcan
London Business School
ttezcan@london.edu

Jiheng Zhang
The Hong Kong University of Science and Technology
jiheng@ust.hk

We study customer service chat (CSC) systems with generally distributed service and patience times by developing measure-valued processes to model and analyze the system dynamics. We first prove that these processes are tight in the many-server asymptotic regime and then show that their limits satisfy a set of fluid model equations. We then establish the invariant states of these limits and use these invariant states to obtain (non-asymptotic) approximations for various performance metrics of CSC systems in the steady state. We also demonstrate the accuracy of these approximations using extensive numerical experiments. These approximations allow us to establish the impact of service and patience time distributions on the system performance and to devise effective dynamic routing policies.

Key words: Customer service chat, measure-valued process, fluid model, invariant state, general distribution

1. Introduction

Customer service chat (CSC) systems have become an integral part of effective customer service. As much as 40% of customer contact centers now provide service through chat, see ICMI (2013) and International (2011). One of the reasons behind CSC systems' popularity is their efficiency: a chat agent can serve multiple customers simultaneously, whereas a call-center agent can only serve one customer at a time (see Tezcan and Zhang (2014) for more details). However, as an agent tries to chat simultaneously with more customers, the service speed will diminish along with the quality of service provided. This novel trade-off introduces new challenges in making operational decisions (which are well studied in other service systems), especially for staffing decisions (i.e., the number of agents scheduled to provide service) and the routing decisions (i.e., how to dynamically match arriving customers with available agents). Another distinct challenge in CSC systems concerns the maximum number of customers an agent should be allowed to handle at any one time. In this paper

we further explore how to make these decisions effectively in view of this trade-off by building on the earlier work in Tezcan and Zhang (2014).

Tezcan and Zhang (2014) studied the optimal routing and staffing in CSC systems to minimize the staffing costs while keeping the abandonment probability below a certain threshold under the assumption that service and abandonment times are exponential. Similar problems have been studied extensively for call center systems where customers are impatient and will abandon the system from the queue if their request is not handled in a timely manner, see Gans et al. (2003). However, unlike call center systems, abandonment during service should be taken into account in CSC systems when making operational decisions because with increased multitasking an agent becomes less and less responsive.

By explicitly modelling customer abandonment during service, Tezcan and Zhang (2014) established two important results. First, they showed that it may be optimal to avoid having agents at levels[1] that satisfy certain conditions, referred to as inefficient levels. In other words, surprisingly it is not always optimal to serve customers at the lowest possible level. (This is true, for example, when the total service rate is concave in a certain sense as a function of level, see equation (5) in Tezcan and Zhang (2014).) However agents' levels change dynamically when customers leave service or new customers are assigned to them. Hence they also provided routing policies that avoid having agents working at inefficient levels in an optimal way in the long run. However they assumed throughout their analysis that service and patience times follow exponential distributions, which is unlikely to hold in practice, for analytical tractability and our numerical simulations (see §7) show that the actual distributions of service and patience times (beyond their first two moments) have a significant impact on the system performance in steady state.

The main goal of this paper is to provide closed-form approximations for various performance measures, such as the steady-state abandonment probability and the expected time in system, for CSC systems with general service and patience time distributions. We assume that routing decisions are made using the policy Tezcan and Zhang (2014) proposed in cases where the arrival rate is not known precisely. The main goal of this policy is to avoid having agents at "inefficient" levels but it is not clear how the concepts of efficient and inefficient levels can even be extended to the case with general distributions. Also, a straightforward extension of the definitions in Tezcan and Zhang (2014) is not possible because their definitions use the exponential distribution assumption explicitly.

We begin our analysis by formulating an adaptation of the static planning linear programming (LP) in Tezcan and Zhang (2014), which they used to identify efficient and inefficient levels, to

[1] We use the term "level i" to refer to the activity (or task) of helping i customers at the same time, and an agent is said to be at level i if that agent is chatting with i customers.

the current case. However, to formulate this program we need to estimate the system performance and it is not clear how this can be done even if we were to ignore the routing problem. Therefore we propose a surprisingly simple and novel approximation method for system performance. We use the solution of this program to provide general definitions for efficiency and then apply the dynamic routing policy in Tezcan and Zhang (2014) with these definitions. We verify the accuracy of the proposed approximations and hence the validity of our definitions of efficient and inefficient levels in two ways: i) we show that our approximations are asymptotically accurate in certain situations by proving that the invariant state of a many-server fluid limit of a queueing system that is similar to the CSC model coincides with our approximations, and ii) we carry out extensive numerical experiments and show that our approximations are highly accurate (for example, our approximations for the abandonment probability are generally within 5% of the simulation results). We also demonstrate that the system performance can be improved significantly by applying the dynamic routing policy in Tezcan and Zhang (2014) (and also demonstrate that this policy avoids having agents at inefficient levels) relative to a commonly used policy that sends customers to one of the least-busy agents.

Technical contributions: A CSC system is essentially a many-server limited processor-sharing queue. Each agent in the server pool is a processor-sharing server who can serve up to I customers simultaneously and the service speed varies with the number of customers the agent is chatting with. Queues with a single processor-sharing server have been studied (Gromoll et al. 2002, Puha and Williams 2004, Zhang et al. 2009) using measure-valued processes and a similar modeling approach is also used in analyzing many-server queues with general distributions, see (Zhang 2013, Long and Zhang 2014). We take a similar modeling approach to these papers and use measure-valued processes to keep track of the system state.

Our analysis, however, is significantly different from these papers because of the inherent difference of CSC systems from processor-sharing and many-server systems. In a standard processer-sharing system, all customers are served by a single server or by a pool of servers whose service capacity can be divided equally among all customers. Hence all customers in service are served at the same rate. In many-server queues each agent can only serve one customer at a time. Therefore, there is no need to keep track of the number of agents at each level in these systems because one can identify the level of servers from the number of customers in service. On the other hand, a server can be operating at any level (up to a limit) in a CSC system and service completions and new arrivals will change the level of servers dynamically. Therefore a more extensive state descriptor than those used in (Gromoll et al. 2002, Puha and Williams 2004, Zhang et al. 2009) and (Zhang 2013, Long and Zhang 2014) is needed.

In our analysis of CSC systems, we first demonstrate that CSC systems can be approximated by a simpler system where the dependence among customers who are assigned to the same server is removed unequivocally. Then we define a measure-valued process that keeps track of the remaining service and patience times of customers in service. (The analysis of the buffer follows Zhang (2013) closely.) Then we show that as the number of servers and the arrival rate go to infinity, the fluid scaled version of this process is tight and every limit satisfies a set of fluid model equations. We then identify invariant states of these fluid models and show that our approximations are asymptotically accurate. Using these results we show that the steady-state behavior of the fluid model of our CSC system depends on the entire distributions of service and patience times and unlike $G/GI/N+GI$ queues, see Whitt (2006), even the steady-state abandonment probability depends on both distributions.

Summary of contributions: The main contributions of this paper can be summarized as follows. i) We extend the definition of efficient and inefficient levels to CSC systems with general distributions; ii) we provide closed-form approximations for various performance measures and show that the entire service and patience time distributions affect the performance of these systems; iii) we provide analytical support for the accuracy of the proposed approximations by showing that they match the invariant state of the measure-valued fluid limit of a similar system; and iv) we verify the accuracy of proposed approximations and support the concept of efficient and inefficient levels using simulation experiments for systems with various sizes and distributions.

1.1. Literature Review

The analysis of many-server and processor-sharing systems is challenging when the service/patience time distributions are general. The CSC system combines both, making the analysis even more difficult. For the processor-sharing (PS) systems, a sequence of works Gromoll et al. (2002), Gromoll (2004), Puha and Williams (2004) developed a framework of using measure-valued processes to obtain both the fluid and diffusion approximations. The framework was extended to the limited processor-sharing (LPS) systems by Zhang and Zwart (2008), Zhang et al. (2009, 2011). Gromoll et al. (2008) studied a PS model with abandonment during service, which is similar to each of the servers in our model. All of the above mentioned works are only for a single PS (or LPS) server. For many-server systems where servers do not multi-task, Whitt (2006) proposed an innovative way of modeling together with a fluid limit. The invariant state of the fluid limit provides fairly accurate approximations for various performance metrics when the system is overloaded. The fluid limit was rigorously proven to serve as the fluid approximation in the many-server heavy traffic regime by Zhang (2013) using measure-valued processes. Long and Zhang (2014) proved that the fluid limit (which is a deterministic dynamic system) converges to the invariant state as time goes

to infinite. It is clear in the literature that the invariant state of a fluid model provides an insightful approximation for the steady state of the original system. Bassamboo and Randhawa (2010) and Bassamboo et al. (2010) showed additional evidence that such fluid approximations yield accurate approximations to the underlying queueing system. Recently, Bassamboo and Randhawa (2016) used such an approximation to estimate patience levels and dynamically prioritize customers based on their time in the system in order to optimize any given system performance metric. Our work follows this line of research by proposing a fluid model to capture the system dynamics and study the invariant states of the fluid model.

This paper is part of our continuing effort to understand the CSC systems. In Luo and Zhang (2013), CSC systems without abandonment were analyzed under a fairly simple routing policy with new arrivals assigned to one of the least-busy agents. Later Tezcan and Zhang (2014) analyzed CSC systems with abandonment. A routing policy based on a linear programming was proposed and shown to be asymptotically optimal in terms of minimizing the abandonment probability. The routing policy, jointly with an LP-based staffing policy, was shown to minimize the required staffing level while keeping the abandonment probability below a desired level for high arrival rates. Both of the previous works heavily relied on the assumption of exponential distributions. This work aims to extend Tezcan and Zhang (2014) to generally distributed service and patience times.

The rest of this paper is organized as follows. §2 describes the system dynamics of the CSC system and presents the concept of efficient and inefficient levels based on a static planning problem and its optimal solution. We also present an effective routing policy that does not require knowledge of the exact external arrival rate. §3 proposes a framework involving measure-valued processes to model the system dynamics. The corresponding fluid model and the fluid approximation are presented in §4. The invariant state of the fluid model is analyzed in §5. We establish approximations for various performance metrics of the CSC system based on the invariant state of the fluid model in §6. We demonstrate the effectiveness of the approximations in §7 and conclude in §8. Finally, technical proofs and detailed simulation results are collected in the appendices.

2. Queueing Model and Preliminary Results

In this section we first introduce the queueing model and then present preliminary results that are fundamental to our analysis. We also review the results from our previous work Tezcan and Zhang (2014) that we need in the current context and highlight the additional complexity induced by non-exponential distributions.

2.1. System Dynamics

Consider a CSC model where customers arrive at the system according to a renewal process $\Lambda(t)$ with rate λ to seek service from a pool of N agents. Agents provide service by chatting with the

customers who are in the system and each agent may serve up to I customers simultaneously [2]. If all agents are busy serving I customers, arriving customers will join the queue and be served according to the first-come-first-served (FCFS) discipline. We assume that agents work in a "non-idling" fashion: if they finish serving a customer and the queue is nonempty, they will start serving the next customer at the head of the queue. Therefore customers wait in queue only when all agents are entirely busy, i.e., at level I.

The operation of a chat system is quite complex. It involves sending messages back and forth between an agent and the customers assigned to that agent and it takes a random number of messages to complete the service of a customer. However, modeling the details of how a chat session actually proceeds is challenging and provides very little insight on how these systems should be managed. Instead, following the models in (Luo and Zhang 2013, Tezcan and Zhang 2014), we assume that an agent serves all customers assigned to him or her simultaneously at a rate that depends on the number of assigned customers. Let μ_i denote the rate at which each customer receives service from an agent serving i customers simultaneously, $i = 1, 2, \ldots, I$, and $l(s)$ denote the level a specific customer is served at time s. Then the cumulative amount of service this customer receives from τ to $\tau + t$ is

$$\int_{\tau}^{\tau+t} \mu_{l(s)} ds. \tag{1}$$

The service of a customer is completed once the cumulative amount of service that the customer receives exceeds his or her service time V, which is assumed to be a random variable with distribution G.

Customers may abandon CSC systems while waiting in queue or during service. The abandonment in queue is modeled in the same way as in call center applications; each customer has a limited patience time following distribution F_q for waiting in queue, and abandons the queue once the time the customer has been waiting exceeds his or her patience time (see e.g., Garnett et al. (2002), Gans et al. (2003), Akşin et al. (2007), Reed and Ward (2008), Tezcan and Behzad (2012) for similar models). In a similar manner, we assume that customers have a limited patience for their service to be completed and we use F to denote its distribution. To illustrate how customers abandon during service, suppose a customer starts receiving service at time τ and is willing to wait U amount of time (i.e., his or her patience time during service) for service to be completed. The customer's service will be completed successfully if

$$\int_{\tau}^{\tau+U} \mu_{l(s)} ds \geq V.$$

Otherwise, the customer abandons the system during service at time $\tau + U$.

[2] We assume I is exogenous for now. We discuss how to choose it optimally below.

We assume customers' service times, patience times for waiting, and patience times during service are mutually independent and follow the distributions G, F_q, and F, respectively, for analytical tractability. To avoid subtle technical issues, we assume that all of the distribution functions are absolutely continuous and F_q is strictly increasing. Without loss of generality, we rescale the time by normalizing the mean service time to 1, i.e.,

$$\int_0^\infty [1-G(x)]dx = 1. \tag{2}$$

Naturally, the amount of service each customer receives per unit time from an agent decreases as the agent chats with more customers. Therefore we assume that

$$\mu_1 > \mu_2 > \cdots > \mu_I. \tag{3}$$

Next we introduce the notation used throughout the paper. Define

$$T_i = \frac{V}{\mu_i} \wedge U, \quad i = 1, \ldots, I. \tag{4}$$

The random variable T_i is the time a customer spends in service if that customer *always* receives service from an agent at level i. Also we set

$$\alpha_i = \frac{1}{\mathbb{E}[T_i]} \quad \text{and} \quad \hat{d}_i = i\alpha_i. \tag{5}$$

Here $\hat{d}_i$ is the total rate (per unit time) that customers depart (by service completion or abandonment during service) from an agent always serving i customers. We define

$$P_i^{Ab} = \mathbb{P}\Big(\frac{V}{\mu_i} > U\Big) \quad \text{for each } i = 1, 2, \ldots, I. \tag{6}$$

The term P_i^{Ab} can be understood as the probability that a customer abandons during service if the customer is *always* served by an agent at level i. It follows directly from (3) and (6) that

$$P_1^{Ab} < P_2^{Ab} < \cdots < P_I^{Ab}. \tag{7}$$

2.2. Intuitive Explanation for Our Approximation

Our definition of CSC systems is not yet complete as we still need to describe a routing policy that determines which agent an arriving customer should be routed to (if there is more than one available agent), and whether a customer should be routed to the queue or to one of the agents upon arrival. As we explained above, we need to identify efficient and inefficient levels to be able to use the policy proposed in Tezcan and Zhang (2014). However, in order to be able to identify these levels, we need accurate approximations for system performance and the system performance depends on the routing policy!

We resolve this dilemma by using a simple (but impractical) routing policy, whose performance can be approximated by applying existing results, to determine efficient and inefficient levels. Consider a routing policy that only allows each agent to work *only* at a predetermined level. To demonstrate the details, consider a system where agents can serve at most two levels with the following parameters: the arrival rate is 20 customers per unit time; service rates for levels 1 and 2 are $\hat{d}_1 = 1$ and $\hat{d}_2 = 3$; and the number of servers is $N = 10$. To be able to meet the demand while keeping the agents at the lowest possible level we assign five agents to level 1 and five agents to level 2. If an agent is assigned to level 2 and there is only one customer assigned to that agent we assume that the service rate is still fixed at μ_2. Because agents are assumed to be serving customers at a fixed level, the probability of abandonment during service in steady state can be determined from the number of servers assigned to each level. (We ignore the abandonment from queue for now for simplicity.) Specifically, in this case the probability of abandonment can be approximated by $P_1^{Ab} * 5/20 + P_2^{Ab} * 15/20$ because in steady state five customers per unit time are routed to level 1 and the rest to level 2[3]. This is the approximation we use for abandonment probability in steady state. In addition we use the performance of agents at each level to determine whether a level is efficient or not, as will be explained in the next section.

Before we proceed we highlight the need to use a more sophisticated policy than the simple policy we just explained. First, the exact arrival rate is not known in advance (and it might be time dependent). Therefore it is not clear how the allocation of servers to different levels can be done in practice. Second, not allowing servers to change levels freely may degrade the system performance since the server pool is effectively divided into two smaller independent server pools. However the steady state of this simple routing scheme can be used for identifying efficient and inefficient levels because we will show that this steady state is identical in the limit to that of a more sophisticated routing policy that does not require knowledge of the arrivals rate and allocates agents to levels efficiently. We next introduce the concept of efficient and inefficient levels and then describe the routing policy we use.

2.3. Efficient and Inefficient Levels

One of the fundamental results in Tezcan and Zhang (2014) is the fact that if agents work at certain inefficient levels, the system performance may deteriorate. They also developed routing policies that "avoid" having agents in these levels (asymptotically). The definition of inefficient levels was motivated by an asymptotic analysis but that analysis does not extend to general distributions,

[3] The validity of this approximation can be proved using the result for $G/GI/N+GI$ queues in Whitt (2006) because the two queueing systems with separate pool of agents operate as many-server queueing systems where agents work at a fixed rate.

except in certain trivial cases. However, using our approximations (introduced in the previous section), we can define the efficiency concept (see §2.3.1) and motivate the definition of "efficiency" using a static planning program (see §2.3.2).

2.3.1. Definition: A level i is said to be *inefficient* if

$$\hat{d}_i < \hat{d}_{i'} \quad \text{for some } 1 \le i' < i, \tag{8}$$

or if there exist k_1 and k_2, such that $1 \le k_1 < i < k_2 \le I$ and

$$(P_{k_2}^{Ab}\hat{d}_{k_2} - P_{k_1}^{Ab}\hat{d}_{k_1})\hat{d}_i \le (P_{k_2}^{Ab}\hat{d}_{k_2} - P_i^{Ab}\hat{d}_i)\hat{d}_{k_1} + (P_i^{Ab}\hat{d}_i - P_{k_1}^{Ab}\hat{d}_{k_1})\hat{d}_{k_2}. \tag{9}$$

All other levels are referred to as *efficient* levels. The precise form of these definitions is based on a static planning problem which we will describe below but for which we give an intuitive explanation here. By (7), customers served at level i have a higher abandonment probability during service than those at level $i' < i$. If condition (8) holds, the throughput of an agent working at level i is also lower than that of an agent at level i'. Thus, it is not desirable to have any agents work at level i. The intuition behind (9) is more intricate and is based on the fact that if (9) holds, allocating an agent to levels k_1 and k_2 for a certain amount of time will result in a higher throughput and a lower probability of abandonment than having that agent serve customers at level i that is in between these two levels.

The definition of efficiency for levels 1 and I is slightly different and obviously condition (9) cannot be checked for levels 1 and I and condition (8) cannot be checked for level 1. Level 1 is said to be efficient if $\mu_1 > \mu_i$ for all $i = 1, \ldots, N$, which holds by assumption (3). If level 1 is inefficient then it is more efficient to have agents at level 2 or above because then the abandonment rate does not increase but the service rate does. On the other hand, level I is efficient if

$$\left(1 - P_I^{Ab}\right)\hat{d}_I \ge \left(1 - P_i^{Ab}\right)\hat{d}_i, \quad \text{for all } i = 1, \ldots, I. \tag{10}$$

(Note that (10) implies that (8) cannot hold for $i = I$ by (7).) Intuitively, (10) implies that the departure rate due to completion of service by an agent working at the maximum level I should be higher than that for any other level. In fact, we show in Lemma EC.2 that if (10) does not hold, it is not optimal to use level I, under the assumption that our approximations are exact and we use a non-idling policy in the sense that agents continue accepting customers up to level I (see Legrosa and Jouinib (2018) for the case when this decision is made dynamically). Hence if (10) does not hold, it is optimal to have customers wait in queue instead of having them served by an agent at level I and we can restrict the maximum level to $I - 1$. If (10) is still invalid for level $I - 1$, we will keep reducing the maximum level by 1 until (10) is valid for a level, which will then be set as

the maximum level of the CSC system. Thus we assume that level I is efficient for the rest of the paper. We highlight the fact that (10) gives a simple condition that can be used to determine the *maximum number of customers* an agent should simultaneously serve. For notational simplicity we denote the set of efficient levels by $\mathscr{F} = \{i_1, i_2, \ldots, i_J\}$ where J is the total number of efficient levels and $i_1 < i_2 < \ldots < i_J$. Note that we have $i_1 = 1$ and $i_J = I$.

2.3.2. Static Planning Problem We motivate the definition of efficient levels based on the solution of a static planning problem for CSC systems, which we describe next. We will later also demonstrate numerically that having agents at inefficient levels decreases system performance. In addition, we use the solution of the static planning problem to identify the invariant state of the fluid models.

Static planning problems are used as a standard initial step to analyze complex queueing networks, see for example Williams (1998) and Harrison (2000). The goal of a static planning program is to gain insight into how best to allocate resources to different tasks in the long run. We next present the static planning program discussed in Tezcan and Zhang (2014). The only difference between our formulation and that in Tezcan and Zhang (2014) is the fact that we use approximations for the P_i^{Ab} we discussed above. For fixed λ and N, consider

$$\min_{\{\lambda_i \geq 0,\ i=1,\ldots,I+1\}} \quad \sum_{i=1}^{I} \lambda_i P_i^{Ab} + \lambda_{I+1} \tag{11}$$

$$s.t. \quad \sum_{i=1}^{I} \frac{\lambda_i}{\hat{d}_i} \leq N, \tag{12}$$

$$\sum_{i=1}^{I+1} \lambda_i \geq \lambda. \tag{13}$$

Intuitively λ_i represents the rate at which customers are served by agents at level i in the long run for $1 \leq i \leq I$ and λ_{I+1} can be viewed as the rate at which customers abandon from queue. Thus, the objective in (11) is to minimize the abandonment rate by choosing appropriate λ_i's. Constraint (12) states that λ_i's must be chosen so that the number of required agents (based on Little's law) does not exceed the capacity N. Constraint (13) implies that all arriving customers must depart from the system.

The following result, which extends Lemma 1 in Tezcan and Zhang (2014) to general service and patience times, establishes the reason we defined efficient levels as in (8) and (9).

Lemma 1. (i) *If $\lambda \leq \hat{d}_1 N$, an optimal solution of the routing LP is given by*

$$\lambda_1^* = \lambda, \text{ and } \lambda_i^* = 0 \text{ for } i > 1.$$

(ii) *If* $\lambda \geq \hat{d}_I N$, *an optimal solution of the routing LP is given by*

$$\lambda_I^* = \hat{d}_I N,\ \lambda_{I+1}^* = \lambda - \hat{d}_I N,\ \textit{and}\ \lambda_i^* = 0\ \textit{for}\ i < I.$$

(iii) *If* $\hat{d}_1 N < \lambda < \hat{d}_I N$, *an optimal solution of the routing LP is given by*

$$\lambda_{i_j^*}^* = \frac{\hat{d}_{i_j^*}}{\hat{d}_{i_{j+1}^*} - \hat{d}_{i_j^*}} \left(\hat{d}_{i_{j+1}^*} N - \lambda\right), \quad \lambda_{i_{j+1}^*}^* = \lambda - \lambda_{i_j^*}^* = \frac{\hat{d}_{i_{j+1}^*}}{\hat{d}_{i_{j+1}^*} - \hat{d}_{i_j^*}} \left(\lambda - \hat{d}_{i_j^*} N\right), \tag{14}$$

where

$$i_{j+1}^* := \min\left\{i : \hat{d}_i \geq \lambda/N,\ i \in \mathscr{F}\right\}. \tag{15}$$

and $\lambda_i^* = 0$ *for* $i \neq i_j^*, i_{j+1}^*$.[4]

The proof is similar to that in Tezcan and Zhang (2014) and we just need to verify that certain properties of efficient levels still hold under general distributions, see §EC.1 for details. For the rest of the paper, we refer to those levels that have positive arrival rates in the optimal solution as *basic levels*. Lemma 1 shows that if a level is inefficient, then it is *suboptimal* to have agents working at this level in the long run and agents should serve customers only at efficient levels. This result motivates our definition of efficient and inefficient levels. Also, by Lemma 1(iii), if there is more than one basic level, these basic levels must be two consecutive efficient levels. This follows from the fact that as agents serve fewer customers at efficient levels, the abandonment probability decreases. Hence it is optimal to keep agents at the lowest indexed efficient levels while making sure that the system has enough capacity to serve all customers, when possible. In the next section, we will describe a routing policy that allocates agents among different levels in an optimal way.

Remark 1 (Exponential service and patience times). When service and patience times are exponentially distributed, the definition of inefficiency can be stated in a much simpler form. To demonstrate, assume that V and U are independent and follow exponential distributions with rates 1 and ν, respectively. By (5) and (6), $\hat{d}_i = i(\mu_i + \nu)$ and $P_i^{Ab} = \nu/(\mu_i + \nu)$. Hence, $P_i^{Ab}\hat{d}_i = i\nu$. Thus (9) simplifies to

$$\hat{d}_i \leq \frac{k_2 - i}{k_2 - k_1}\hat{d}_{k_1} + \frac{i - k_1}{k_2 - k_1}\hat{d}_{k_2}. \tag{16}$$

This is equivalent to condition (5) in Tezcan and Zhang (2014). With a little algebra, it can be checked that the abandonment rate ν does not play a role in (16) in determining the efficiency of a level in the exponential case unlike in the case with general distributions where P_i^{Ab} depends on the entire distribution of patience times during service.

[4] The optimal solution in (14) is well defined because we have $\hat{d}_{i_{j+1}^*} > \hat{d}_{i_j^*}$ by Lemma EC.1.

2.4. Routing Policy

The static planning problem provides insights into how agents should be allocated to different levels in the long run. However, it is not clear how this can be accomplished dynamically by routing arriving customers to available agents. Assuming that service and patience times are exponentially distributed, Tezcan and Zhang (2014) studied this problem and devised novel routing policies that were shown to be asymptotically optimal in terms of minimizing the steady-state probability of abandonment. We focus on one of the policies proposed in Tezcan and Zhang (2014) that does not require knowledge of the exact arrival rate. Other policies studied in Tezcan and Zhang (2014) can be similarly analyzed once the efficient and inefficient levels are identified.

Consider the following policy. Let i denote the index of the lowest indexed non-empty (i.e. there are agents working at that level) level and i_j denote the index of the efficient level with the highest index below i or set $i_j = i$ if i is efficient. Denote by

$$\mathscr{U}_{i_j} = \{i_j + 1, \ldots, i_{j+1} - 1\} \tag{17}$$

all of the inefficient levels strictly between the two efficient levels i_j and i_{j+1}. The proposed policy routes a new arrival as follows:

- If $i = 0$, route the customer to an agent at level 0.
- If $1 \leq i < I$, route the customer to an agent at the *highest* non-empty level in $\{i_j\} \cup \mathscr{U}_{i_j} = \{i_j, \ldots, i_{j+1} - 1\}$.
- If $i = I$, the customer has to join the queue.

We denote this policy by π. The lack of dependence on the arrival rate makes this policy fairly robust and easy to implement. The intuition behind this policy is to force agents away from levels in $\mathscr{U}_{i_j}$ to efficient levels, see Tezcan and Zhang (2014) for more details. Also if $\mathscr{U}_{i_j} = \emptyset$ (equivalently $i_j + 1 = i_{j+1}$), for all $i_j \in \mathscr{F}$, i.e., all levels are efficient, then this policy reduces to the lightest-load-first policy (i.e., customers are routed to one of the least busy agents) in Luo and Zhang (2013).

We will later show that this policy achieves the optimal allocations of arrivals identified by the static planning problem in the fluid limits. However we are not able to extend the asymptotic optimality of π established in Tezcan and Zhang (2014) for exponential service and patience time distributions. This is mainly due to the difficulty in analyzing the asymptotic behavior of the underlying fluid model as t goes to infinity.

3. Model Formulation

In this section, we present an asymptotic analysis of CSC systems in the many-server regime. We use a measure-valued state descriptor to model CSC systems as a Markovian process because

using the standard head count processes (as commonly done in traditional queueing theory under Markovian assumptions) is not sufficient with general distributions. We use a measure that keeps track of the *remaining* service and patience times of each customer in the system following the previous work on many-server systems, similar to Zhang (2013), Gromoll et al. (2002) and Gromoll (2004). However, even such a state space descriptor is not rich enough because customers who are served by the same agent also move between levels together when the agent finishes serving one of the customers or is assigned a new customer. Yet modeling this in detail does not yield insightful results and removing the connection among customers does not result in a huge information loss, hence we make the following simplifying assumption.

Assumption 1 (A Modified System). *Assume that each agent moving from level i to level $i-1$ causes $i-1$* ***randomly selected*** *customers to leave level i and join level $i-1$; and each agent moving from level $i-1$ to level i causes $i-1$* ***randomly selected*** *customers to leave level $i-1$ and join level i.*

Clearly the queueing model under Assumption 1 is not identical to the underlying chat service system. We believe, however, that the queueing model under this assumption is very similar to the original CSC system for two reasons: i) When service and patience times are exponential, the modified and the original systems are equivalent in distribution due to the memoryless property; and ii) for general distributions these two systems perform almost identically in terms of various performance metrics in various numerical experiments (for example, the difference in the steady-state abandonment probability is less than 10^{-4} on average across in a variety of scenarios, see §EC.2 for details). For the rest of this paper, our analysis will focus on the modified system without any further mention.

3.1. Measure-valued Process

Let $\mathcal{L}_i(t)$ denote a measure describing the status of all the customers who are currently served by level i (for $i \geq 1$) agents at time t. More precisely, set $C_x \times C_y = (x,\infty)\times(y,\infty)$. Then $\mathcal{L}_i(t)(C_x \times C_y)$, $x, y \geq 0$, denotes the number of customers with remaining service time larger than x and remaining patience time during service larger than y. (In general, we can use a Borel set $B \subset \mathbb{R}^2_+$ instead of $C_x \times C_y$.) From the definition, we have

$$\mathcal{L}_i(t)(\mathbb{R}^2_+) = iZ_i(t), \tag{18}$$

where $Z_i(t)$ is the number of agents at level i at time t, for $i = 1, \ldots, I$. The number of idle (level 0) servers is given simply by

$$Z_0(t) = N - \sum_{i=1}^{I} Z_i(t) \tag{19}$$

as there are N agents in total.

If all agents are completely occupied, i.e., they are at level I, arrivals must wait in the queue and will be served later according to the FCFS discipline. When there are customers in the queue, the system dynamics will become exactly the same as call center models if we view the pool as $I \cdot N$ agents. To capture the dynamics of customers waiting in queue we use a *virtual buffer* as in Zhang (2013). The idea behind the virtual buffer is to keep customers in the queue until it is their turn for service even when their patience is exhausted. Specifically, when an agent becomes available, the customer who has been waiting in the queue for the longest is admitted to service if his or her remaining patience time is non-negative; otherwise that customer abandons the system. This process is repeated until a customer with positive remaining patience time is identified or until all customers have abandoned the queue. Working with the virtual buffer simplifies the analysis and the actual queue can easily be recovered from this process as we explain next.

Set $C_x = (x, \infty)$ and let $\mathcal{R}(t)(C_x)$, $x \in \mathbb{R}$, denote the number of customers in the queue with remaining patience time for waiting larger than x at time t. Then the number of customers in the actual queue can be expressed as $Q(t) = \mathcal{R}(t)(C_0)$. Clearly the following *non-idling constraint* must be satisfied at any time $t \geq 0$,

$$Q(t)(N - Z_I(t)) = 0. \tag{20}$$

3.2. Dynamics of the Modified System

To present the dynamic equations that govern the evolution of the system under Assumption 1, we introduce an operator, $\mathcal{X}$, on measures. Let $\mathcal{X} = \sum_{j=1}^{J} \delta_{(v_j, u_j)}$ where $\delta_{(v_j,u_j)}$ denote the Dirac point measure at $(v_j, u_j) \in \mathbb{R}^2_+$. We use Φ^k for $k \leq J$ to denote a random selection operator defined by

$$\Phi^k(\mathcal{X}) = \sum_{i=1}^{k} \delta_{(v_{j_i}, u_{j_i})}, \tag{21}$$

where the set of indices $\{j_1, \ldots, j_k\}$ are chosen randomly from $\{1, \ldots, J\}$.

Server Pool: Assume that customers arrive according to the renewal process $\Lambda(\cdot)$ with rate λ. For each $i = 1, \ldots, I$, let $A_i(t)$ denote the number of "arrivals" to level i, that is, those customers whose service commences at level i by time t. For levels 1 to $I-1$, these are those customers who, at the time of their arrival, were routed to an agent at level $i-1$ and so matched with $i-1$ customers at level $i-1$. For level I, $A_I(t)$ captures not only the customers who were routed to an agent at level $I-1$ but also those customers who commenced service at level I after waiting in the queue upon a departure from level I. Also, let $S_i(t)$ denote the number of customers who have departed the system (due to either service completion or abandonment during service) from level i by time t.

Let $\tau_{i,j}$ denote the time of the jth arrival at level i, and $v_{i,j}$ and $u_{i,j}$ denote the service and patience times of this arrival, respectively. We use $M_{i,i-1}$ to denote the number of times agents go to level $i-1$ from level i, and similarly $M_{i-1,i}$ to denote the number of times agents go to level i from level $i-1$. For notational simplicity, we set $M_{0,-1} = M_{-1,0} = M_{I,I+1} = M_{I+1,I} = 0$. The measure-valued process $\mathcal{L}_i(\cdot)$ satisfies the following *stochastic dynamic equation*:

$$\begin{aligned}
\mathcal{L}_i(t)(C_x \times C_y) = \mathcal{L}_i(0)(C_{x+\mu_i t} \times C_{y+t}) &- \int_0^t \Phi^{i-1}\left(\mathcal{L}_i(s)\right)(C_{x+\mu_i(t-s)} \times C_{y+t-s}) dM_{i,i-1}(s) \\
&+ \int_0^t \Phi^{i-1}\left(\mathcal{L}_{i-1}(s)\right)(C_{x+\mu_i(t-s)} \times C_{y+t-s}) dM_{i-1,i}(s) \\
&+ \sum_{j=1}^{A_i(t)} \mathbb{1}_{\{v_{i,j} > x+\mu_i(t-\tau_{i,j}), u_{i,j} > y+t-\tau_{i,j}\}} \\
&+ \int_0^t \Phi^{i}\left(\mathcal{L}_{i+1}(s)\right)(C_{x+\mu_i(t-s)} \times C_{y+t-s}) dM_{i+1,i}(s) \\
&- \int_0^t \Phi^{i}\left(\mathcal{L}_{i}(s)\right)(C_{x+\mu_i(t-s)} \times C_{y+t-s}) dM_{i,i+1}(s)
\end{aligned} \tag{22}$$

for any $x, y \geq 0$ and $i = 1, \ldots, I$, where we use the convention that $\Phi^0(\mathcal{X}) \equiv \mathbf{0}$, with $\mathbf{0}$ denoting the zero measure. The first term on the right-hand side of (22) captures the influence from the initial state. In order for a customer at level i at time 0 to be still at level i at time t and with their remaining service and patience times to be larger than x and y, respectively, his or her remaining service and patience times at time 0 must be larger than $x+\mu_i t$ and $y+t$, respectively. The second term captures the fact that an agent moving from level i to level $i-1$ causes $i-1$ randomly selected customers to leave level i, so we have to remove $i-1$ customers from this level using the random selection operator Φ^{i-1}. The third and fourth terms captures the impact of customers routed to level $i-1$. Each customer who is routed to level $i-1$ brings $i-1$ customers from level $i-1$ to level i in addition to himself. These customers are accounted for in the third term. Also if there are customers in the queue when a service is completed, the customer at the head of the queue will be immediately served at level I, without changing the state of the agent to which the customer is assigned. Similar to the second and the third terms, each agent moving from level $i+1$ to level i will cause i customers to move from level $i+1$ to level i, which is described in the fifth term; and each agent moving from level i to level $i+1$ will cause i customers to leave level i, which is described in the last term in (22).

The processes $M_{i,i-1}$ and $M_{i-1,i}$ satisfy

$$M_{i,i-1}(t) = \begin{cases} \int_0^t \mathbb{1}_{\{Q(s-)=0\}} dS_I(s), & i = I, \\ S_i(t), & i = 1, \ldots, I-1, \end{cases} \tag{23}$$

$$M_{i-1,i}(t) = \begin{cases} \int_0^t \mathbb{1}_{\{Z_I(s-)<N\}} dA_I(s), & i = I, \\ A_i(t), & i = 1, \ldots, I-1. \end{cases} \tag{24}$$

The two equations above follow from the fact that once a customer ends a chat with an agent at level i for $i \in \{1, \ldots, I-1\}$ and leaves the system, that agent goes to level $i-1$. Similarly, once a customer is assigned to an agent at level i for $i \in \{0, 1, \ldots, I-2\}$, the agent goes to level $i+1$. However, this is not true for a customer leaving from level I: after a customer departs from level I, an agent at level I goes to level $I-1$ only if the queue is empty; otherwise that agent remains at level I. Similarly a customer can be assigned to an agent at level $I-1$ only if not all agents are at level I. It can be seen from (23) and (24) that

$$M_{i,i-1}(t) - M_{i-1,i}(t) = S_i(t) - A_i(t) \quad \text{for all } i = 1, \ldots, I. \tag{25}$$

From the above discussion, we also have the following balance equation for the number of agents at each level:

$$Z_i(t) = Z_i(0) - S_i(t) + A_i(t) + S_{i+1}(t) - A_{i+1}(t) \quad \text{for } i = 0, 1, \ldots, I, \tag{26}$$

where we assume $S_0 = A_0 = S_{I+1} = A_{I+1} \equiv 0$ to omit a separate discussion for levels 0 and I. For notational simplicity, we assume that those customers who are initially present in the system have been there for a certain bounded amount of time. We also assume that the actual service and patience times (in queue and in service) of customers who are in the system at time zero have the same distributions as other customers.

Buffer: Let $R(t) = \mathcal{R}(t)(\mathbb{R})$ denote the total number of customers in the virtual buffer. Initially, there are $R(0)$ customers in the virtual buffer. Index them by $j = -R(0)+1, \ldots, 0$ according to their arrival time a_j, which is a negative number indicating how long the jth customer had been there by time 0. Similarly, index the newly arrivals on the time interval $(0, t]$ by $j = 1, 2, \ldots, \Lambda(t)$ in the order of arrival with a_j being the jth arrival time. For both customers initially in the virtual buffer and those who are newly arrivals, let u_j^q be the patience time for waiting of the jth customer. Define

$$B(t) = \Lambda(t) - R(t). \tag{27}$$

It is clear that at time t the index of the head-of-the-line customer in the virtual buffer is $B(t)+1$. Moreover, $B(t) - B(s)$ can be viewed as the number of customers who leave the virtual buffer and is about to be admitted into service during time interval $(s, t]$.

Denote by γ_j the time when the jth customer starts service for all $j \geq -R(0)+1$. Note that the jth customer enters service only if $\gamma_j - a_j$ is less than the patience time u_j^q. Then

$$\mathcal{R}(t)(C_x) = \sum_{j=B(t)+1}^{\Lambda(t)} \mathbb{1}_{\{u_j^q > x+t-a_j\}} \tag{28}$$

for any $x \in \mathbb{R}$. And the cumulative number of customers who have entered service can be written as

$$E(t) = \sum_{j=-R(0)+1}^{B(t)} \mathbb{1}_{\{u_j^q > \gamma_j - a_j\}} = \sum_{i=1}^{I} A_i(t), \tag{29}$$

where the second equality follows from the fact that customers who enter service will commence their service at a certain level. The abandonment process, $D(t)$, can be recovered from the following balance equation of the physical queue:

$$Q(t) = Q(0) + \Lambda(t) - D(t) - E(t). \tag{30}$$

Departure process: Similar to how (22) captures the system dynamics, the following equation determines how the departure process S_i from level i, $i = 1, \ldots, I$, evolves. Define the set

$$\mathcal{A}(x, y) = \{(x', y') \in \mathbb{R}_+^2 : x' \le x \text{ or } y' \le y\} = (C_x \times C_y)^c. \tag{31}$$

Then

$$\begin{aligned} S_i(t) = {} & \mathcal{L}_i(0)(\mathcal{A}_i(\mu_i t, t)) - \int_0^t \Phi^{i-1}\left(\mathcal{L}_i(s)\right)\left(\mathcal{A}(\mu_i(t-s), t-s)\right) dM_{i,i-1}(s) \\ & + \int_0^t \Phi^{i-1}\left(\mathcal{L}_{i-1}(s)\right)\left(\mathcal{A}(\mu_i(t-s), t-s)\right) dM_{i-1,i}(s) \\ & + \sum_{j=1}^{A_i(t)} \mathbb{1}_{\{v_{i,j} \le \mu_i(t-\tau_{i,j}) \text{ or } u_{i,j} \le t-\tau_{i,j}\}} \\ & + \int_0^t \Phi^{i}\left(\mathcal{L}_{i+1}(s)\right)\left(\mathcal{A}(\mu_i(t-s), t-s)\right) dM_{i+1,i}(s) \\ & - \int_0^t \Phi^{i}\left(\mathcal{L}_{i}(s)\right)\left(\mathcal{A}(\mu_i(t-s), t-s)\right) dM_{i,i+1}(s). \end{aligned} \tag{32}$$

Allocation of arrivals: The final process we define captures the allocation of customers to available servers; that is, process A_i, $i = 1, \ldots, I$. Any static priority policy basically specifies a one-to-one mapping $p : \{0, \ldots, I-1\} \to \{0, \ldots, I-1\}$ such that for any $i, j \in \{0, \ldots, I-1\}$, level j has priority over level i if and only if $p(j) < p(i)$. This means that a new arrival cannot be routed to a level i agent whenever there are agents at any level j with $p(j) < p(i)$. Therefore any *static priority* policy has to satisfy

$$\int_0^t \sum_{\{j=0,\ldots,I-1 : p(j) < p(i)\}} Z_j(s) dA_{i+1}(s) = 0, \quad i = 1, \ldots, I-1. \tag{33}$$

Note that under our policy proposed in §2.4 level 0 has the highest priority hence we set $p(0) = 0$. For other levels the priorities under this rule can be set as follows:

$$\begin{cases} p(i_j) = i_{j+1} - 1,\ p(i_j + 1) = i_{j+1} - 2, \ldots,\ p(i_{j+1} - 1) = i_j, & \mathscr{U}_{i_j} \neq \emptyset, \\ p(i_j) = i_j, & \mathscr{U}_{i_j} = \emptyset. \end{cases} \tag{34}$$

As mentioned above, if $\mathscr{U}_{i_j} = \emptyset$ for all $i_j \in \mathscr{F}$, then the policy π simply becomes the lightest-load-first policy with $p(i) = i$, $i = 0, \ldots, I-1$.

4. Asymptotic Analysis

In this section we first introduce a deterministic measure-valued fluid model, and then show that it serves as the fluid limit of the CSC system in the many-server asymptotic regime.

A fluid model. The underlying idea behind constructing fluid models is to replace the stochastic components in the system dynamics with their corresponding distributional but deterministic information. We use the bar sign to indicate fluid model processes associated with the queuing processes we defined above. Specifically, $\bar{A}_i$ is the fluid amount of "arrivals" to level i and $\bar{S}_i$ is the fluid amount of departures from level i, for $i=1,\ldots,I$. Moreover, $\bar{M}_{i,i-1}$ is the fluid amount of agents moving to level $i-1$ from level i, and similarly $\bar{M}_{i-1,i}$ is the fluid amount of agents moving to level i from level $i-1$. For the CSC model, we first construct the *fluid dynamic equation* for the **server pool**, corresponding to (22) as follows:

$$\begin{aligned}\bar{\mathcal{L}}_i(t)(C_x\times C_y) &= \bar{\mathcal{L}}_i(0)(C_{x+\mu_i t}\times C_{y+t})\\ &\quad -\int_0^t \frac{i-1}{i\bar{Z}_i(s)}\bar{\mathcal{L}}_i(s)(C_{x+\mu_i(t-s)}\times C_{y+t-s})d\bar{M}_{i,i-1}(s)\\ &\quad +\int_0^t \frac{1}{\bar{Z}_{i-1}(s)}\bar{\mathcal{L}}_{i-1}(s)(C_{x+\mu_i(t-s)}\times C_{y+t-s})d\bar{M}_{i-1,i}(s)\\ &\quad +\int_0^t G^c(x+\mu_i(t-s))F^c(y+t-s)d\bar{A}_i(s)\\ &\quad +\int_0^t \frac{i}{(i+1)\bar{Z}_{i+1}(s)}\bar{\mathcal{L}}_{i+1}(s)(C_{x+\mu_i(t-s)}\times C_{y+t-s})d\bar{M}_{i+1,i}(s)\\ &\quad -\int_0^t \frac{1}{\bar{Z}_i(s)}\bar{\mathcal{L}}_i(s)(C_{x+\mu_i(t-s)}\times C_{y+t-s})d\bar{M}_{i,i+1}(s),\end{aligned} \tag{35}$$

$t\geq 0,\ x,y\geq 0$ for all $i=1,\ldots,I$, where we again take $\bar{M}_{-1,0}=\bar{M}_{0,-1}=\bar{M}_{I,I+1}=\bar{M}_{I+1,I}=0$. However, if $\bar{Z}_i(t)=0$ the fluid model equation is defined as follows

$$\frac{1}{\bar{Z}_i(t)}\bar{\mathcal{L}}_i(t)(C_x\times C_y)d\bar{M}_{i,i-1}(t)=0 \tag{36}$$

and

$$\begin{aligned}\frac{1}{\bar{Z}_i(t)}\bar{\mathcal{L}}_i(t)(C_x\times C_y)d\bar{M}_{i,i+1}(t) &= \frac{1}{\bar{Z}_{i-1}(t)}\bar{\mathcal{L}}_{i-1}(t)(C_x\times C_y)d\bar{M}_{i-1,i}(t)+G^c(x)F^c(y)d\bar{A}_i(t)\\ &\quad +\frac{i}{(i+1)\bar{Z}_{i+1}(t)}\bar{\mathcal{L}}_{i+1}(t)(C_x\times C_y)d\bar{M}_{i+1,i}(t).\end{aligned} \tag{37}$$

When $\bar{Z}_i(t)=0$ we need a separate equation because $\bar{\mathcal{L}}_i(t)/\bar{Z}_i(t)$ is not well defined. Intuitively, (36) indicates that the rate at which the fluid content moves from level i to level $i-1$ should be 0 when there are no agents at level i. Meanwhile, customers from the adjacent levels will be immediately pushed to level $i+1$ by new arrivals who are routed to level i. Thus, (37) means that the customers moving between levels i and $i+1$ consist of a mix of customers from levels $i-1$ and $i+1$.

Corresponding to (23) and (24),

$$\bar{M}_{i,i-1}(t) = \bar{S}_i(t) \quad \text{and} \quad \bar{M}_{i-1,i}(t) = \bar{A}_i(t) \quad \text{for } i = 1, \ldots, I-1. \tag{38}$$

The processes $\bar{M}_{I,I-1}$ and $\bar{M}_{I-1,I}$ satisfy

$$\int_0^t \bar{Q}(s) d\bar{M}_{I,I-1}(s) = \int_0^t \bar{Q}(s) d\bar{M}_{I-1,I}(s) = 0, \tag{39}$$

and

$$d\bar{M}_{I,I-1}(t) = d\bar{S}_I(t) \quad \text{and} \quad d\bar{M}_{I-1,I}(t) = d\bar{A}_I(t) \quad \text{if } \bar{Z}_I(t) < N. \tag{40}$$

Moreover,

$$\bar{M}_{i,i-1}(t) - \bar{M}_{i-1,i}(t) = \bar{S}_i(t) - \bar{A}_i(t) \quad \text{for all } i = 1, \ldots, I. \tag{41}$$

The fluid amount of customers at level i satisfies

$$\bar{\mathcal{L}}_i(t)(\mathbb{R}_+^2) = i\bar{Z}_i(t) \quad \text{for } i = 1, \ldots, I, \tag{42}$$

and

$$\bar{Z}_0(t) = N - \sum_{i=1}^{I} \bar{Z}_i(t). \tag{43}$$

Also

$$\bar{Z}_i(t) = \bar{Z}_i(0) - \bar{S}_i(t) + \bar{A}_i(t) + \bar{S}_{i+1}(t) - \bar{A}_{i+1}(t) \quad \text{for } i = 0, 1, \ldots, I. \tag{44}$$

Similar to (26), we also set $\bar{S}_0 = \bar{A}_0 = \bar{S}_{I+1} = \bar{A}_{I+1} \equiv 0$ to make (44) compatible with $i = 0$ and I.

Corresponding to (32), the (fluid) departure process $\bar{S}_i$ from level i satisfies

$$\begin{aligned}
\bar{S}_i(t) = {} & \bar{\mathcal{L}}_i(0)(\mathcal{A}_i(\mu_i t, t)) \\
& - \int_0^t \frac{i-1}{i\bar{Z}_i(s)} \bar{\mathcal{L}}_i(s)(\mathcal{A}(\mu_i(t-s), t-s)) d\bar{M}_{i,i-1}(s) \\
& + \int_0^t \frac{1}{\bar{Z}_{i-1}(s)} \bar{\mathcal{L}}_{i-1}(s)(\mathcal{A}(\mu_i(t-s), t-s)) d\bar{M}_{i-1,i}(s) \\
& + \int_0^t \left[1 - G^c(\mu_i(t-s)) F^c(t-s)\right] d\bar{A}_i(s) \\
& + \int_0^t \frac{i}{(i+1)\bar{Z}_{i+1}(s)} \bar{\mathcal{L}}_{i+1}(s)(\mathcal{A}(\mu_i(t-s), t-s)) d\bar{M}_{i+1,i}(s) \\
& - \int_0^t \frac{1}{\bar{Z}_i(s)} \bar{\mathcal{L}}_i(s)(\mathcal{A}(\mu_i(t-s), t-s)) d\bar{M}_{i,i+1}(s).
\end{aligned} \tag{45}$$

The fluid dynamics for the **buffer** is given by

$$\bar{\mathcal{R}}(t)(C_x) = \lambda \int_{t-\frac{\bar{R}(t)}{\lambda}}^{t} F_q^c(x+t-s)ds, \quad t \geq 0,\ x \in \mathbb{R}, \tag{46}$$

where $\bar{R}(t) = \bar{\mathcal{R}}(t)(\mathbb{R})$ is the fluid content in the virtual buffer. (Recall that F_q is the patience time distribution for waiting in queue and we set $F_q^c(\cdot) = 1 - F_q(\cdot)$, the complementary cumulative distribution of F_q.) Also

$$\bar{B}(t) = \bar{\Lambda}(t) - \bar{R}(t). \tag{47}$$

The (fluid) queue content can be represented as $\bar{Q}(t) = \bar{\mathcal{R}}(t)(C_0)$, which satisfies the balance equation

$$\bar{Q}(t) = \bar{Q}(0) + \bar{\Lambda}(t) - \bar{D}(t) - \bar{E}(t). \tag{48}$$

Here $\bar{\Lambda}(t) = \lambda t$ is the external arrival process, $\bar{D}(t)$ is the abandonment process from the buffer, and the total amount that enters service is

$$\bar{E}(t) = \int_0^t F_q^c\Big(\frac{\bar{R}(s)}{\lambda}\Big) d\bar{B}(s) = \sum_{i=1}^{I} \bar{A}_i(t). \tag{49}$$

The static priority policy (33) corresponds to

$$\int_0^t \sum_{\{j=0,\ldots,I-1: p(j)<p(i)\}} \bar{Z}_j(s) d\bar{A}_{i+1}(s) = 0, \quad i = 1, \ldots, I-1. \tag{50}$$

For our policy π, $p(\cdot)$ is defined as in (34).

The following non-idling constraint always holds for all $t \geq 0$:

$$\bar{Q}(t)(N - \bar{Z}_I(t)) = 0. \tag{51}$$

We refer to (35)–(51) as the *fluid model* and any tuple $(\bar{\mathcal{R}}, \bar{\mathcal{L}}, \bar{R}, \bar{Q}, \bar{Z}, \bar{\Lambda}, \bar{B}, \bar{D}, \bar{E}, \bar{A}, \bar{S}, \bar{M})$ that satisfies (35)–(51) as a *fluid model* solution.

Fluid limits. We next show that the limit of the fluid scaled queueing processes in the many-server regime satisfies the fluid model equations. Consider a sequence of CSC systems indexed by $n = 1, 2, \ldots$ (thus we append a superscript n to the notation for the corresponding stochastic processes). Assume that both the arrival rate and the number of agents increase to infinity. More precisely

$$\frac{\Lambda^n(\cdot)}{n} \Rightarrow \lambda \cdot \ \text{ and } \ \frac{N^n}{n} \to N, \ \text{as } n \to \infty, \tag{52}$$

where $\Rightarrow$ denotes weak convergence in Skorohod (J_1) topology. Define the fluid scaled processes as

$$\bar{X}^n(t) = \frac{X^n(t)}{n}, \tag{53}$$

where X^n is a symbolic notation for $\mathcal{R}^n$, $\mathcal{L}_i^n$, R^n, Q^n, Z_i^n, Λ^n, B^n, D^n, E^n, A_i^n, S_i^n, $M_{i,i-1}^n$ and $M_{i-1,i}^n$. We assume that the initial states satisfy

$$\bar{\mathcal{R}}^n(0) \Rightarrow \bar{\mathcal{R}}(0), \quad \bar{\mathcal{L}}_i^n(0) \Rightarrow \bar{\mathcal{L}}_i(0), \quad i=1,\ldots,I, \tag{54}$$

for measures $\bar{\mathcal{R}}(0)$ and $\bar{\mathcal{L}}(0)=(\bar{\mathcal{L}}_1(0),\ldots,\bar{\mathcal{L}}_I(0))$ satisfying

$$\bar{\mathcal{R}}(0)(\{x\})=0 \text{ for any } x\in\mathbb{R}, \tag{55}$$

$$\bar{\mathcal{L}}_i(0)(\{x\}\times\mathbb{R}_+)=\bar{\mathcal{L}}_i(0)(\mathbb{R}_+\times\{y\})=0 \text{ for any } x,y\geq 0. \tag{56}$$

Theorem 1 (Fluid Limits). *In the many-server regime specified by* (52)*, if the initial state satisfies* (54)–(56)*, then the sequence of fluid scaled stochastic processes* $\{(\bar{\mathcal{R}}^n,\bar{\mathcal{L}}^n,\bar{R}^n,\bar{Q}^n,\bar{Z}^n,\bar{\Lambda}^n,\bar{B}^n,\bar{D}^n,\bar{E}^n,\bar{A}^n,\bar{S}^n,\bar{M}^n):n\in\mathbb{N}\}$ *under any static priority policy* (33) *is tight in the Skorohod* (J_1) *topology. Denote by* $\bar{Z}_i(\cdot)$, $i=1,\ldots,I$, *the weak limit of* $\bar{Z}_i^n(\cdot)$. *Assume* $\bar{Z}_i(\cdot)$, $i=1,\ldots,I$, *switches between 0 and positive values only finitely many times in any bounded time interval, then every weak limit of the fluid scaled stochastic processes satisfies the fluid model equations* (35)–(51).

The proof, presented in §EC.5, consists of two major steps. The first step is to show that the sequence is tight (which implies that every subsequence has a converging subsequence). The second step is to verify that the limit of any convergent subsequence satisfies the fluid model equations.

Remark 2 (Connection to Exponential Service and Patience Times). To facilitate the understanding of the fluid model equations we consider exponential service time and exponential patience time during service, i.e., $G^c(x)=e^{-x}$ and $F^c(x)=e^{-\nu x}$. By (18), at time t there are $i\bar{Z}(t)$ customers being served at level i. Index them by $k=1,\ldots,iZ_i^n(t)$. Note that the order could be arbitrary. We also use $v_{i,k}$ and $u_{i,k}$ to denote the remaining service time and remaining patience time during service of the kth customer at time t. Then by definition

$$\bar{\mathcal{L}}_i^n(t)(C_x\times C_y)=\frac{1}{n}\sum_{k=1}^{iZ_i^n(t)}\mathbb{1}_{\{v_{i,k}>x,u_{i,k}>y\}}.$$

By the memoryless property, $v_{i,k}$'s follow the same distribution as G and $u_{i,k}$'s follow distribution F. It then follows from the tightness proved in Theorem 1 and the Glivenko-Cantelli estimate (EC.12) that

$$\bar{\mathcal{L}}_i(t)(C_x\times C_y)=i\bar{Z}_i(t)e^{-x}e^{-\nu y} \tag{57}$$

for all levels with $\bar{Z}_i(t)>0$. Obviously, (57) also holds for the case with $\bar{Z}_i(t)=0$. Plugging the above equation to (35) and (45) yields

$$d\bar{S}_i(t)=i(\mu_i+\nu)\bar{Z}_i(t)dt. \tag{58}$$

Then the fluid dynamic equation (35) becomes

$$\begin{aligned}\bar{Z}_i(t) = \bar{Z}_i(0)e^{-(\mu_i+\nu)t} - (i-1)\int_0^t (\mu_i+\nu)\bar{Z}_i(s)e^{-(\mu_i+\nu)(t-s)}ds + \int_0^t e^{-(\mu_i+\nu)(t-s)}d\bar{A}_i(s) \\ + (i+1)\int_0^t (\mu_{i+1}+\nu)\bar{Z}_{i+1}(s)e^{-(\mu_i+\nu)(t-s)}ds - \int_0^t e^{-(\mu_i+\nu)(t-s)}d\bar{A}_{i+1}(s).\end{aligned} \tag{59}$$

The proof of the above two equations is placed in Lemma EC.3. Taking derivatives of both sides yields the following ordinary differential equation:

$$\dot{\bar{Z}}_i(t) = -i(\mu_i+\nu)\bar{Z}_i(t) + \dot{\bar{A}}_i(t) + (i+1)(\mu_{i+1}+\nu)\bar{Z}_{i+1}(t) - \dot{\bar{A}}_{i+1}(t), \tag{60}$$

which is precisely the same fluid dynamic equation for exponential service and patience times in Tezcan and Zhang (2014).

5. Invariant State

In this section we identify invariant states of the fluid model of CSC systems. First we show in Proposition 1 that there will be at most two levels the agents will provide service in the invariant state and those levels must be efficient and consecutive (when there are two). This proves that the proposed policy avoids having agents at inefficient levels. Then we prove in Theorem 2 that there exists an invariant state, which can be stated in a relatively simple closed form, when two basic levels are non-adjacent or when there is only one basic level. Unfortunately, we are not able to obtain a similar result if the basic levels are adjacent, so instead we study two special cases in Theorem 3 and show that the form of invariant state we obtain in Theorem 2 is still valid. Finally we propose an approximation for the systems that are not covered by Theorems 2 and 3. We will later use the invariant states to derive approximations for various performance metrics in §6 and we will verify the accuracy of these approximations numerically in §7.

Definition: A state $(\bar{\mathcal{L}}(\infty), \bar{\mathcal{R}}(\infty))$ is said to be an *invariant state* of the fluid model if $(\bar{\mathcal{L}}(0), \bar{\mathcal{R}}(0)) = (\bar{\mathcal{L}}(\infty), \bar{\mathcal{R}}(\infty))$. Then

$$(\bar{\mathcal{L}}(t), \bar{\mathcal{R}}(t)) = (\bar{\mathcal{L}}(\infty), \bar{\mathcal{R}}(\infty)) \tag{61}$$

is a solution to the fluid model (35)–(51) for all $t > 0$.

For an invariant state $(\bar{\mathcal{L}}(\infty), \bar{\mathcal{R}}(\infty))$, let

$$\lim_{\delta\to 0} \frac{\bar{\mathcal{L}}_i(\infty)(\mathcal{A}(\mu_i\delta, \delta))}{\delta} =: \lambda_i \tag{62}$$

(the limit exists a.e. by Lemma EC.7). Then by (61) and (EC.52), $\bar{S}_i(t) = \lambda_i t$. From (44), we have

$$\bar{A}_i(t) = \bar{S}_i(t) = \lambda_i t \quad \text{for } i = 1, \ldots, I. \tag{63}$$

Similar to the static planning problem (11)-(13), the rates $(\lambda_1, \ldots, \lambda_I)$ can be interpreted as a long-run allocation of the external arrivals to each service level. We now show that there can be at most two non-negative arrival rates allocated to efficient levels in the invariant state. This also verifies that the routing policy π avoids having agents at inefficient levels in the long run. The proof of the following proposition is placed in §EC.4.

Proposition 1. *If $(\bar{\mathcal{L}}(\infty), \bar{\mathcal{R}}(\infty))$ is an invariant state of the fluid model (35)–(51), then there can be at most two efficient levels $i_j, i_{j+1} \in \mathscr{F}$ satisfying $\lambda_{i_j} > 0$ and $\lambda_{i_{j+1}} > 0$.*

We next identify invariant states in several special cases.

Theorem 2. *Let $(\lambda_1^*, \ldots, \lambda_I^*)$ be defined as follows under the following cases:*

(i) *If $\lambda \le \hat{d}_1 N$, then $\lambda_1^* = \lambda$ and $\lambda_i^* = 0$ for $i > 1$.*

(ii) *If $\lambda \ge \hat{d}_I N$, then $\lambda_I^* = \hat{d}_I N$ and $\lambda_i^* = 0$ for $i < I$.*

(iii) *If $\hat{d}_1 N < \lambda < \hat{d}_I N$ and $\hat{d}_{i_{j+1}^*} = \lambda/N$, then $\lambda_{i_{j+1}^*}^* = \lambda$ and $\lambda_i^* = 0$ for $i \ne i_{j+1}^*$.*

(iv) *If $\hat{d}_1 N < \lambda < \hat{d}_I N$, $\hat{d}_{i_{j+1}^*} > \lambda/N$ and $i_{j+1}^* \ne i_j^* + 1$, then $\lambda_{i_j^*}^*$ and $\lambda_{i_{j+1}^*}^*$ are given as in (14), and $\lambda_i^* = 0$ for $i \ne i_j^*, i_{j+1}^*$.*

The CSC fluid model (35)–(51) has an invariant state $(\bar{\mathcal{L}}(\infty), \bar{\mathcal{R}}(\infty))$ defined as follows:

If $\lambda_i^* > 0$, then

$$\bar{\mathcal{L}}_i(\infty)(C_x \times C_y) = \lambda_i^* \int_0^\infty G^c(x + \mu_i s) F^c(y + s) ds, \quad x, y \ge 0, \tag{64}$$

$$\bar{Z}_i(\infty) = \frac{\lambda_i^*}{\hat{d}_i}. \tag{65}$$

If $\lambda_i^* = 0$ but $\lambda_{i+1}^* > 0$ then

$$\bar{\mathcal{L}}_i(\infty) = \mathbf{0}, \ \bar{Z}_i(\infty) = 0, \ \text{and} \tag{66}$$

$$\frac{1}{i\bar{Z}_i(\infty)} \bar{\mathcal{L}}_i(\infty) = \frac{1}{(i+1)\bar{Z}_{i+1}(\infty)} \bar{\mathcal{L}}_{i+1}(\infty). \tag{67}$$

If $\lambda_i^* = 0$ and $\lambda_{i+1}^* = 0$ then

$$\bar{\mathcal{L}}_i(\infty) = \mathbf{0}, \ \bar{Z}_i(\infty) = 0, \ \text{and } \bar{\mathcal{L}}_i(\infty)/\bar{Z}_i(\infty) = \mathbf{0}. \tag{68}$$

And $\bar{\mathcal{R}}(\infty)$ is given by

$$\bar{\mathcal{R}}(\infty)(C_x) = \lambda \int_0^w F_q^c(x + s) ds, \quad x \in \mathbb{R}, \tag{69}$$

where w is a unique solution to $F_q(w) = \max\left(\frac{\lambda - \hat{d}_I N}{\lambda}, 0\right)$.

The proof is presented in §EC.4. In cases (i), (ii) and (iii) all customers are served only at a single level and in case (iv) customers are served at two basic levels that are non-adjacent. It is easy to verify that the arrival rates, λ_i^*'s, for the invariant states agree with the optimal solution of the static planning problem (11)-(13) in these cases. Also the invariant state of the buffer $\bar{\mathcal{R}}(\infty)$ is identical to that of the $G/GI/N+GI$ queue described by (3.13) in Zhang (2013), which is expected since overloaded CSC systems are similar to multi-server queues where all servers serve I customers.

We will show below that the system performance mainly depends on the invariant states of basic levels (those with $\lambda_i^*>0$ or those with $\bar{Z}_i(\infty)>0$). However, we will need the invariant states of non-basic levels in proving Theorem 2. Also because $\bar{\mathcal{L}}_i/\bar{Z}_i$ is the limit of $\bar{\mathcal{L}}_i^n/\bar{Z}_i^n$ we express its limit separately. The limit is well defined, even when $\bar{Z}_i(\infty)=0$, as we proved in §4.

Unfortunately, if the two basic levels i_j and i_{j+1} are adjacent, i.e., $i_{j+1}=i_j+1$, then we cannot obtain a closed-form expression for the invariant state. In the following theorem, we present two special cases for which the closed-form invariant state can still be obtained and has the same form as part (iv) of Theorem 2. The proof is provided in §EC.4.

Theorem 3. *If $\hat{d}_1 N<\lambda<\hat{d}_I N$, $\hat{d}_{i^*_{j+1}}>\lambda/N$ and $i^*_{j+1}=i^*_j+1$, and one of the following conditions holds*

Condition I: *service times and patience times during service follow exponential distributions, i.e., $G^c(x)=e^{-x}$ and $F^c(y)=e^{-\nu y}$,*

Condition II: *customers have unlimited patience during service, i.e., $F(y)\equiv 0$ for any $y\geq 0$,*

then the fluid models of the CSC systems (35)–(51) *have the invariant state given in part (iv) of Theorem 2.*

For general service and patience time distributions it can be easily verified that the simple form (64) is no longer an invariant state of the fluid model if there are exactly two basic levels and these two levels are adjacent (i.e., $i^*_{j+1}=i^*_j+1$). The main issue here is that because customers move between these two levels, the distributions of customers' service and patience times in the invariant state interact in a complicated manner and we are not able to capture this interaction in a closed form. This is not an issue, for example, when the two basic levels are non-adjacent because they do not interact or when both service and patience times have exponential distributions, in which case the remaining service and patience time distributions are identical because of the memoryless property of the exponential distribution.

For the rest of this paper we use (64) as an approximation for the invariant state of the fluid model. To obtain further insight as to when this approximation is accurate, it is easy to see that the invariant state is given as in part (iv) of Theorem 2 when $\mu_{i^*_j}=\mu_{i^*_{j+1}}$. Clearly this contradicts

our assumption (3) but it does imply that if the service rates between two adjacent levels are not significantly different, (64) might provide a sensible approximation for the system's invariant state. Next, we will build approximations for various performance measures based on this approximation.

6. Approximations

In this section we build approximations based on the results in §5, in a manner similar to Whitt (2006) and Bassamboo and Randhawa (2016) who focus on call center models. We mainly focus on the probability of abandonment, and the mean and variance of time in the system in steady state. Other performance metrics can also be estimated based on the invariant state of the fluid model.

6.1. Approximations for Probability of Abandonment

We now use the invariant state of the CSC model to develop our approximation for the probability of abandonment from each level, which serves as the building block for other approximation formulae to follow in §6.2 and §6.3. The results in this section also support our idea of approximation presented in §2.2, which we have been using throughout the paper.

It follows from (64) that

$$\bar{\mathcal{L}}_i(\infty)(C_{\mu_i\delta}\times C_\delta)=\lambda_i^*\int_\delta^\infty G^c(\mu_i s)F^c(s)ds$$

for any $\delta>0$. Therefore the total departure rate due to service completion and abandonment from level i is given by

$$\psi_i=\lim_{\delta\to 0}\frac{\bar{\mathcal{L}}_i(\infty)(C_0\times C_0)-\bar{\mathcal{L}}_i(\infty)(C_{\mu_i\delta}\times C_\delta)}{\delta}=-\frac{d}{d\delta}\bar{\mathcal{L}}_i(\infty)(C_{\mu_i\delta}\times C_\delta)\Big|_{\delta=0}=\lambda_i^*.$$

The rate at which fluid content at level i departs that level by abandonment, ψ_i^a, is

$$\begin{aligned}\psi_i^a&=\lim_{\delta\to 0}\frac{\bar{\mathcal{L}}_i(\infty)(C_0\times C_0)-\bar{\mathcal{L}}_i(\infty)(C_0\times C_\delta)}{\delta}\\&=-\frac{d}{d\delta}\bar{\mathcal{L}}_i(\infty)(C_0\times C_\delta)\Big|_{\delta=0}\\&=\lambda_i^*\int_0^\infty G^c(\mu_i s)f(s)ds,\end{aligned}$$

where f is the pdf associated with the distribution function F and the last equality follows from (64). Hence the proportion of customers who abandon the system among those who depart the system from level i is given by

$$\frac{\psi_i^a}{\psi_i}=\int_0^\infty G^c(\mu_i s)f(s)ds=P_i^{Ab},\tag{70}$$

where the last equality follows from the definition of P_i^{Ab} in (6).

Remark 3. Based on this approximation the abandonment probability during service is given by $\sum_{i=1}^{I} \lambda_i^* P_i^{Ab}$. This approximation is identical to that we described in §2.2, providing further evidence that our definitions of efficient and inefficient levels are valid.

Remark 4. By (70), our approximation for the probability of abandonment depends on the information regarding both the service and patience time distributions. We also demonstrate this via numerical experiments below where we show that the abandonment probability can change as much as by 31.6% when we switch from log-normal service and patience time distributions to exponential ones without altering their mean and variance. This is significantly different from the approximations for the traditional many-server systems based on fluid models. In those models the abandonment probability (in the fluid limit) only depends on the mean service time and not on the distribution of service or abandonment times. However, the other performance measures (such as expected time in queue) may also depend on the entire patience time distribution for those systems, see Whitt (2006), Bassamboo and Randhawa (2016), Long and Zhang (2014) for more details.

We next use (70) to build approximations for the other performance metrics for underloaded and overloaded systems.

6.2. Approximations for Underloaded Systems

First assume that the system is underloaded, i.e.,

$$\lambda < \hat{d}_I N. \tag{71}$$

Under this condition, the system nominally has sufficient capacity to serve all customers. We next provide approximations in steady state for the probability of abandonment P^{Ab}, the expected time in system $\mathbb{E}[W]$, the standard deviation of time in system $\text{stdev}(W)$, and the conditional expected time in system given that the customer will eventually complete service successfully, $\mathbb{E}[W|S]$, and abandon the system, $\mathbb{E}[W|A]$.

If $\hat{d}_1 N < \lambda < \hat{d}_I N$, then by the invariant state of the fluid limit and (14), a fraction

$$q_{i_j^*} = \frac{\hat{d}_{i_{j+1}^*} N - \lambda}{\hat{d}_{i_{j+1}^*} - \hat{d}_{i_j^*}} \cdot \frac{\hat{d}_{i_j^*}}{\lambda}$$

of arriving customers is served by an agent at level i_j^*, and the remaining $1 - q_{i_j^*}$ is served by an agent at level i_{j+1}^* in the fluid invariant state. According to (70) the probability of abandonment for those customers served by level i_j^* agents can be approximated by $P_{i_j^*}^{Ab}$. Hence, the probability that an arriving customer abandons the system in steady state can be approximated by

$$P^{Ab} = q_{i_j^*} P_{i_j^*}^{Ab} + (1 - q_{i_j^*}) P_{i_{j+1}^*}^{Ab}.$$

By (65) the expected number of agents at level i, N_i, is given by $N_i = \lambda_i^*/\hat{d}_i$. Using Little's Law, we have

$$\mathbb{E}[W] = \frac{i_j^* N_{i_j^*} + i_{j+1}^* N_{i_{j+1}^*}}{\lambda}. \tag{72}$$

Next we focus on the conditional expected time in system and the variance of the time in system. Let

$$S_i^{(c)} = \mathbb{E}\left[\frac{V}{\mu_i} | \frac{V}{\mu_i} \le U\right] \quad \text{and} \quad S_i^{(a)} = \mathbb{E}\left[U | \frac{V}{\mu_i} > U\right], \quad i = 1, \ldots, I, \tag{73}$$

where U and V are defined in §2.1. In words, $S_i^{(c)}$ is the conditional expected service time of a customer given that the customer's service is completed and $S_i^{(a)}$ is the conditional expected patience time of a customer given that the customer abandons service, if the customer is served by an agent at level i in steady state. Then, conditional on the level of the agent that a customer is served by, we have

$$\mathbb{E}[W|S] = \frac{q_{i_j^*}\left(1 - P_{i_j^*}^{Ab}\right)}{1 - P^{Ab}} S_{i_j^*}^{(c)} + \frac{\left(1 - q_{i_j^*}\right)\left(1 - P_{i_{j+1}^*}^{Ab}\right)}{1 - P^{Ab}} S_{i_{j+1}^*}^{(c)},$$

and

$$\mathbb{E}[W|A] = \frac{q_{i_j^*} P_{i_j^*}^{Ab}}{P^{Ab}} S_{i_j^*}^{(a)} + \frac{\left(1 - q_{i_j^*}\right) P_{i_{j+1}^*}^{Ab}}{P^{Ab}} S_{i_{j+1}^*}^{(a)}.$$

Next we consider the standard deviation of the time spent in system in steady state. Conditional on the level of the agent that a customer is served by, we have

$$\mathbb{E}\left[W^2\right] = q_{i_j^*} \mathbb{E}[T_{i_j^*}^2] + \left(1 - q_{i_j^*}\right) \mathbb{E}[T_{i_{j+1}^*}^2], \tag{74}$$

where T_i is defined as in (4). Therefore, the standard deviation of the time spent in system in steady state, stdev(W), can be approximated by

$$\text{stdev}(W) = \left(\mathbb{E}\left[W^2\right] - \left(\mathbb{E}[W]\right)^2\right)^{1/2}, \tag{75}$$

where $\mathbb{E}[W^2]$ is defined as in (74) and $\mathbb{E}[W]$ is defined as in (72).

If $\lambda \le \hat{d}_1 N$, then by Theorem 2 all of the arrivals will be served at level 1 in the fluid invariant state, thus $P^{Ab} = P_1^{Ab}$. By (65), the expected number of agents at level 1, N_1, is given by $N_1 = \lambda/\hat{d}_1$. Applying Little's law yields $\mathbb{E}[W] = 1/\hat{d}_1$. Since all customers are served by agents at level 1,

$$\mathbb{E}[W|S] = S_1^{(c)} \quad \text{and} \quad \mathbb{E}[W|A] = S_1^{(a)}.$$

Moreover, $\mathbb{E}[W^2] = \mathbb{E}[T_1^2]$ and so the standard deviation can be found as in (75).

6.3. Approximations for Overloaded Systems

We now turn our attention to overloaded systems, i.e., when the inequality in (71) is reversed. The approximations we build in this case follows closely those for the multi-server queue (e.g., Whitt (2006)). However there are still certain differences due to the fact that customers can also abandon during service in CSC systems.

First note that a customer can exit the system in three different ways: i) abandonment from the queue, ii) abandonment during service, and iii) service completion. From (69), a customer have to wait in queue for w time units before entering service. Then, the probability that a customer abandons from queue is $F_q(w)$ and reaches service is $F_q^c(w) = 1 - F_q(w)$. Because all customers are served at level I, the probability of abandonment during service provided that the customer reaches service is P_I^{Ab}. Therefore, the probability that a customer abandons the system in steady state is given by

$$P^{Ab} = F_q(w) + F_q^c(w)P_I^{Ab}.$$

Next we find an approximation for the expected time in system. By Theorem 2, the expected number of customers in queue is given by $\lambda \int_0^w F_q^c(s)ds$ and the expected number of agents at level I is N. Therefore we have

$$\mathbb{E}[W] = \frac{\lambda \int_0^w F_q^c(s)ds + IN}{\lambda} \tag{76}$$

by Little's law. We now consider the expected time in system conditional on the exit point of a customer. Let A_q denote the event that a customer abandons the queue. Then the conditional expected time in system in steady state given that a customer abandons from queue is given by

$$\mathbb{E}[W|A_q] = \mathbb{E}[Y_q|Y_q \leq w], \tag{77}$$

where Y_q is a random variable with distribution F_q. For those who reach service, the expected total time in system in steady state is given by

$$\mathbb{E}[W|A_q^c] = w + \frac{1}{\alpha_I},$$

where we use A_q^c to denote the event that a customer reaches service and α_I is defined in (5). To approximate the standard deviation of time in system, conditional on whether a customer enters service or not, we can obtain

$$\begin{aligned}\mathbb{E}[W^2] &= \mathbb{E}[W^2|A_q]\mathbb{P}(A_q) + \mathbb{E}[W^2|A_q^c]\mathbb{P}(A_q^c) \\ &= \mathbb{E}[Y_q^2|Y_q \leq w]F_q(w) + \mathbb{E}[(w+T_I)^2]F_q^c(w),\end{aligned}$$

where Y_q is the same as the one in (77) and T_I is defined in (4). The standard deviation can easily be obtained from this equation and (76).

7. Numerical Experiments

In this section, we present the results of extensive numerical experiments in systems with the number of agents ranging from 25 to 100 and in two different experiment sets. We have three goals: i) to demonstrate the accuracy of our approximations based on the asymptotic analysis; ii) to show that the distribution of service and patience times have a significant impact on the performance of CSC systems; and iii) to demonstrate that carefully crafted routing policies can significantly improve system performance.

In §7.1 we explain the parameters used in our experiments and in §7.2 we present the results when all the service levels are efficient. We illustrate in §7.3 the effect of inefficient levels on system performance. Due to space constraints we mainly focus on the probability of abandonment in underloaded systems. Results on other performance measures can be found in Appendix EC.6, where we also present the results of additional experiments for overloaded systems.

7.1. Simulation Parameters

We consider two different experimental settings with the main difference being that customers are less patient in the first than in the second setting. In both settings we set $I = 6$ and assume that arrivals follow a Poisson process and that customers' patience for waiting in queue has an exponential distribution with mean 1.

In the first experimental setting we let $\mu = \{4, 3.8, 3.3, 3, 2.75, 2.5\}$ and consider three different pairs of values of λ and N, the arrival rate and the number of agents. The details are presented in Table 1(a). For each (λ, N) pair, we simulate the system under three different combinations of service and patience time distributions (see Table 2(a) for details). From here on we use "expo(x)" to denote an exponential random variable with mean x and "ln(x, y)" to denote a log-normal distribution with mean x and variance y.

System	λ	N
1_1	281.25	25
2_1	562.5	50
3_1	1125	100

(a) Experiment set 1

System	λ	N
1_2	375	25
2_2	750	50
3_2	1500	100

(b) Experiment set 2

Table 1 Arrival rates and number of agents in each set of experiments

The setup of the second set of experiments is similar. We set $\mu = \{10, 7, 5.1, 4, 3.3, 2.8\}$ and use three different pairs of values of λ and N presented in Table 1(b). For each (λ, N) pair we simulate the system under three different combinations of service and patience time distributions presented in Table 2(b). Therefore, we consider nine systems in total in each set of experiments.

Combination	Service Time	Patience Time
I_1	expo(1)	expo(1)
II_1	ln(1,0.2)	expo(1)
III_1	ln(1,1)	ln(1,1)

(a) Experiment set 1

Combination	Service Time	Patience Time
I_2	expo(1)	expo(2)
II_2	ln(1,0.2)	expo(2)
III_2	ln(1,1)	ln(2,4)

(b) Experiment set 2

Table 2 Combinations of service and patience time distributions

We choose arrival rates such that the agents are distributed between two basic levels (i_j^* and i_{j+1}^*) at different ratios in different experiments to explore its affect on our approximations. For example, in systems 1_1 through 3_1, the arrival rates are chosen so that when the service and patience times are exponential we have $Z_{i_j^*} = Z_{i_{j+1}^*}$ and in systems 1_2 through 3_2 there is only one basic level. (The optimal values of $Z_{i_j^*}$ and $Z_{i_{j+1}^*}$ for each setting are presented in Appendix EC.6.) In addition the parameters are chosen to observe the effect of the coefficient of variation of service times on the accuracy of our approximations – they are lower in experiments II_1 and II_2.

We run each simulation long enough to observe 2 million arrivals. The first 10% of the simulation time is regarded as the warm-up period, and thus is discarded when computing steady-state performance metrics. The last 10% of the simulation time is also discarded to avoid the potential impact of customers who are still in service at the end of the simulation.

7.2. Experiments with All Efficient Levels

In both set of experiments all the levels are efficient under the service rates specified in §7.1 and the experimental parameters in Tables 1 and 2. Hence in these cases the proposed policy π reduces to the lightest-load-first policy that gives priority to the least busy agents and chooses one randomly when necessary. In this section we explore how various parameters affect the accuracy of our approximations and demonstrate the impact of service and abandonment time distributions on system performance.

Combination	System	$P^{Ab}{}_{\text{sim}}$	$P^{Ab}{}_{\text{approx}}$	Rel. Error (%)
I_1	1_1	0.2234($\pm$0.0004)	0.2222	0.54
	2_1	0.2227($\pm$0.0004)	0.2222	0.22
	3_1	0.2223($\pm$0.0003)	0.2222	0.04
II_1	1_1	0.2520($\pm$0.0005)	0.2511	0.36
	2_1	0.2513($\pm$0.0004)	0.2511	0.08
	3_1	0.2510($\pm$0.0004)	0.2511	0.04
III_1	1_1	0.1530($\pm$0.0003)	0.1519	0.72
	2_1	0.1524($\pm$0.0003)	0.1519	0.33
	3_1	0.1521($\pm$0.0003)	0.1519	0.13

Table 3 Comparison of simulation results and approximations for P^{Ab} of experiment set 1

The results of the first and second sets of simulation experiments are presented in Tables 3 and 4,

respectively, where we show the results for the abandonment probability along with the relative error of our approximations for each combination. For example, in system 1_1 when both distributions are exponential, our approximations underestimate (compared with the simulation results) the abandonment probability by 0.54%. More detailed results of the experiments along with 95% confidence intervals are presented in Appendix EC.6.

We first point out that the service and patience time distributions have a significant effect on the performance of the systems in both sets of experiments. To illustrate this, we note that the average abandonment probability when both distributions are exponential is around 22% in systems $1_1 - 3_1$ and it is only 15% when both distributions are log-normal with the mean and standard deviation of service and patience distributions kept fixed. Similarly, in the second set of experiments, the results are similar with an average abandonment probability of 7.3% vs. 1.9% for combinations I_2 and III_2, respectively. This should come as no surprise in the light of our approximations for P_i^{Ab} in (6), where both distributions play a role.

In the first set of experiments, our approximations are highly accurate. In almost all the experiments, errors are less than 1%, with an average of just 0.27%. The relative errors of the approximations for expected time in system, conditional expected time in system for abandoned and served customers, and standard deviation of time in system are about the same (see Appendix EC.6). The quality of our approximations improves with system size, as expected.

Combination	System	${P^{Ab}}_{\text{sim}}$	${P^{Ab}}_{\text{approx}}$	Rel. Error (%)
I_2	1_2	0.0752($\pm$0.0003)	0.0667	11.30
	2_2	0.0728($\pm$0.0003)	0.0667	8.38
	3_2	0.0702($\pm$0.0002)	0.0667	4.99
II_2	1_2	0.0806($\pm$0.0003)	0.0748	7.20
	2_2	0.0781($\pm$0.0003)	0.0748	4.23
	3_2	0.0766($\pm$0.0002)	0.0748	2.35
III_2	1_2	0.0212($\pm$0.0001)	0.0187	11.79
	2_2	0.0197($\pm$0.0001)	0.0187	5.08
	3_2	0.0189($\pm$0.0001)	0.0187	1.06

Table 4 Comparison of simulation results and approximations for P^{Ab} of experiment set 2

The quality of the approximations in the second set of experiments is relatively worse, especially when the number of agents is equal to 25. For larger systems the relative error decreases: on average, for systems with 50 and 100 agents, the relative errors are around 5.5% and 2.8%, respectively.

When the agents are estimated to be more evenly distributed between two basic levels, our approximations are much more accurate. For example, in experiment I_1, our estimates for the expected number of agents at levels 2 and 3 are equal (see Table EC.2 in Appendix EC.6) and in experiment I_2, all of the agents are estimated to be working at level 2 (see Table EC.5 in

Appendix EC.6) when service times and patience times during service are exponential. This is mainly due to the fact that in the actual system, especially when N is small, "second-order" fluctuations have a bigger impact on experiments when agents are unevenly distributed. This is in a way similar to the analysis of traditional queueing systems because fluid limits do not provide very accurate estimates in heavy traffic, but approximations based on diffusion limits, which capture the second-order fluctuations, are reasonably accurate. The diffusion limits of CSC systems have been studied in Cui and Tezcan (2016) under exponential assumptions. For general distributions, we leave the diffusion analysis to future research.

7.3. Experiments with Inefficient Levels

In order to illustrate the effect of inefficient levels we run simulations using the first experimental setting except we set $\mu_3 = 2.9$ and $\mu_4 = 2.8$ (instead of their original values $\mu_3 = 3.3$ and $\mu_4 = 3$) making levels 3 and 4 inefficient under each distribution pair in Table 2(a). We carry out simulation experiments with this change for the same arrival rates and number of agents given in Table 1(a). Now the order of priority for policy π becomes $p(1) = 1,\ p(2) = 4,\ p(3) = 3,\ p(4) = 2,\ p(5) = 5$, which is obviously different from the lightest-load-first policy, and we have $i_j^* = 2$ and $i_{j+1}^* = 5$ in all experiments.

The results of the relative errors and the improvements from using our proposed policies are presented in Table 5. Specifically, in the last column titled "Improvement", we display the improvements in abandonment probability if the proposed policy π is used as opposed to the lightest-load-first policy. We observe that the abandonment probability can be reduced significantly by as much as 12.34% with an average of 8.5%.

Combination	System	$P^{Ab}{}_{\text{sim-lightest}}$	$P^{Ab}{}_{\text{sim-}\pi}$	$P^{Ab}{}_{\text{approx-}\pi}$	Rel. Error (%)	Improvement (%)
I_1	1_1	0.2441($\pm$0.0005)	0.2303($\pm$0.0003)	0.2259	1.91	5.65
	2_1	0.2451($\pm$0.0005)	0.2286($\pm$0.0004)	0.2259	1.18	6.73
	3_1	0.2461($\pm$0.0006)	0.2275($\pm$0.0005)	0.2259	0.70	7.56
II_1	1_1	0.2809($\pm$0.0005)	0.2619($\pm$0.0003)	0.2578	1.57	6.76
	2_1	0.2828($\pm$0.0004)	0.2603($\pm$0.0004)	0.2578	0.96	7.96
	3_1	0.2842($\pm$0.0005)	0.2593($\pm$0.0004)	0.2578	0.58	8.76
III_1	1_1	0.1801($\pm$0.0003)	0.1629($\pm$0.0002)	0.1590	2.39	9.55
	2_1	0.1819($\pm$0.0003)	0.1614($\pm$0.0003)	0.1590	1.49	11.27
	3_1	0.1832($\pm$0.0002)	0.1605($\pm$0.0003)	0.1590	0.93	12.34

Table 5 Comparison of simulation results and approximations for P^{Ab} of experiment set 1 with inefficient levels

Under the column "Rel. Error", we present the error of the approximation for the abandonment probability relative to the simulation result. The errors of our approximations are slightly higher in this case than those observed in the previous section, especially when the number of agents is

equal to 25. However, even in systems with 25 agents the average error is less than 2.39% across all combinations and performance metrics. Besides, for larger systems the relative error is much lower: on average, for systems with 50 and 100 agents, the relative error is around 1.15% and 0.72%, respectively. The main reason behind increased errors is the fact that the inefficient levels between two basic levels are asymptotically empty (i.e. there are no agents working at those levels) at all times but they are not in finite size systems because of the randomness in arrivals and service completions.

8. Conclusions

In this paper, we analyze CSC systems with generally distributed service and patience times. Typically these systems have multiple agents and each agent can serve multiple customers simultaneously. These unique features make the analysis challenging, especially when combined with general service and patience time distributions. We present a tractable alternative system to serve as a proxy for the original CSC system. We then use measure-valued processes and construct equations that capture the dynamics of the alternative system. We then establish the fluid limits of these processes and show that they satisfy a set of fluid model equations. We then analyze the invariant state of the fluid model and obtain approximations for various performance metrics of the system in the steady state based on these invariant states.

Our numerical experiments demonstrate that our approximations are accurate in general and easy to calculate once the service and patience time distributions are determined. Due to their simplicity, our approximations would be especially effective i) in making staffing decisions even when the arrival rate itself is random (see Bassamboo et al. (2010)) and ii) in performing various kinds of what-if analyses, for example, when the system manager can influence the service rates, for example, via additional agent training. In employing our approximations, however, caution must be taken when the service rates between two adjacent basic levels are significantly different. Nevertheless, we did not observe a significant degradation in the performance of our approximations even in these cases in our numerical experiments.

Our results rely on several assumptions that can be verified in future research. First, we did not prove the convergence of the fluid model solutions to the invariant state and it is not clear if the invariant state is unique. Second, we used a modified system for tractability without establishing analytically whether or not it is a good approximation for the original system. Third, we did not try to optimize the routing decisions and instead used the ones that have been established to be asymptotically optimal when service and patience times are exponential. Finally, we assumed that the arrival rate is constant. Having said that, our approximations can still be used to manage systems by dividing the day into non-overlapping intervals if the arrival rate does not change too

rapidly, see Gans et al. (2003). However, if the arrival rate change quickly (compared to service times), analysis similar to Liu and Whitt (2014) might be more practical.

Acknowledgments

Tolga Tezcan's research is supported in part by NSF Grants CMMI-0954126 and CMMI-1130346. Jiheng Zhang's research is supported in part by GRF Grants No. 16200617 and 16501015 from Hong Kong Research Grants Council.

References

Akşin, Z., M. Armony, and V. Mehrotra (2007). The modern call center: A multi-disciplinary perspective on operations management research. *Production and Operations Management 16*(6), 665–688.

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

Bassamboo, A. and R. S. Randhawa (2016). Scheduling homogeneous impatient customers. *Management Science 62*(7), 2129–2147.

Bassamboo, A., R. S. Randhawa, and A. Zeevi (2010). Capacity sizing under parameter uncertainty: Safety staffing principles revisited. *Mgt. Sci. 56*(10), 1668–1686.

Billingsley, P. (1968). *Convergence of probability measures*. Wiley Series in Probability and Statistics: Probability and Statistics. New York: John Wiley & Sons Inc.

Billingsley, P. (1999). *Convergence of probability measures* (Second ed.). Wiley Series in Probability and Statistics. New York: John Wiley & Sons Inc.

Cui, L. and T. Tezcan (2016). Approximations for chat service systems using many-server diffusion limits. *Mathematics of Operations Research 41*(3), 775–807.

Dai, J. G. and R. J. Williams (1996). Existence and uniqueness of semimartingale reflecting brownian motions in convex polyhedrons. *Theory of Probability & Its Applications 40*(1), 1–40.

Ethier, S. N. and T. G. Kurtz (1986). *Markov processes*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. New York: John Wiley & Sons Inc.

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

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

Gromoll, H. C. (2004). Diffusion approximation for a processor sharing queue in heavy traffic. *Ann. Appl. Probab. 14*(2), 555–611.

Gromoll, H. C., A. L. Puha, and R. J. Williams (2002). The fluid limit of a heavily loaded processor sharing queue. *Ann. Appl. Probab. 12*(3), 797–859.

Gromoll, H. C., P. Robert, and B. Zwart (2008). Fluid limits for processor sharing queues with impatience. *Math. Oper. Res. 33*(2), 375–402.

Harrison, J. M. (2000). Brownian models of open processing networks: canonical representation of workload. *Ann. Appl. Probab. 10*(1), 75–103.

ICMI (2013). Extreme engagement in the multichannel contact center: Leveraging the emerging channels research report and best practices guide. ICMI Research Report.

International, T. (2011). Best practices: Online chat sales. Benchmarking study — White paper.

Kallenberg, O. (1986). *Random measures* (Fourth ed.). Berlin: Akademie-Verlag.

Legrosa, B. and O. Jouinib (2018). On the scheduling of operations in a chat contact center. Technical report, Paris School of Business.

Liu, Y. and W. Whitt (2014). Algorithms for time-varying networks of many-server fluid queues. *INFORMS Journal on Computing 26*(1), 59–73.

Long, Z. and J. Zhang (2014). Convergence to equilibrium states for fluid models of many-server queues with abandonment. *Oper. Res. Lett. 42*(6–7), 388–393.

Luo, J. and J. Zhang (2013). Staffing and control of instant messaging contact centers. *Oper. Res. 61*(2), 328–343.

Puha, A. L. and R. J. Williams (2004). Invariant states and rates of convergence for a critical fluid model of a processor sharing queue. *Ann. Appl. Probab. 14*(2), 517–554.

Reed, J. E. and A. R. Ward (2008). Approximating the $GI/GI/1+GI$ queue with a nonlinear drift diffusion: hazard rate scaling in heavy traffic. *Math. Oper. Res. 33*(3), 606–644.

Royden, H. L. (1988). *Real analysis* (Third ed.). New York: Macmillan Publishing Company.

Rudin, W. (1987). *Real and complex analysis* (Third ed.). New York: McGraw-Hill Book Co.

Sen, P. and J. Singer (1994). *Large Sample Methods in Statistics: An Introduction with Applications*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis.

Tezcan, T. and B. Behzad (2012). Robust design and control of call centers with flexible interactive voice response systems. *Manufacturing & Service Operations Management 14*(3), 386–401.

Tezcan, T. and J. Zhang (2014). Routing and staffing in customer service chat systems with impatient customers. *Oper. Res. 62*(4), 943–956.

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

Williams, R. J. (1998). Diffusion approximations for open multiclass queueing networks: sufficient conditions involving state space collapse. *Queueing Syst. 30*(1-2), 27–88.

Zhang, J. (2013). Fluid models of many-server queues with abandonment. *Queueing Syst. 73*(2), 147–193.

Zhang, J., J. G. Dai, and B. Zwart (2009). Law of Large Number Limits of Limited Processor-Sharing Queues. *Math. Oper. Res. 34*(4), 937–970.

Zhang, J., J. G. Dai, and B. Zwart (2011). Diffusion Limits of Limited Processor-Sharing Queues. *Ann. Appl. Probab. 21*(2), 745–799.

Zhang, J. and B. Zwart (2008). Steady state approximations of limited processor sharing queues in heavy traffic. *Queueing Syst. 60*(3-4), 227–246.

Appendix: Customer Service Chat Systems with General Service and Patience Times

We prove Lemma 1 in §EC.1. In §EC.2 we present the numerical results to compare the original CSC system with the modified system. Then we present the related results about the fluid model in §EC.3 and prove the results of the invariant state in §EC.4. The proofs of the convergence of the stochastic model to its fluid limits are presented in §EC.5. In the end, the details of the results of the simulation experiments in §7 appear in §EC.6.

EC.1. Proofs of Efficient and Inefficient Levels

Proof of Lemma 1. The proof of the result is similar to that of Lemma 1 in Tezcan and Zhang (2014) once we establish the properties of efficient levels, namely (EC.1), (EC.2) and Lemma EC1 in Tezcan and Zhang (2014). First it can be easily checked that condition (9) is equivalent to the following:

$$\left(\hat{d}_{k_2}-\hat{d}_{k_1}\right)\left(P_{k_2}^{Ab}\hat{d}_{k_2}-P_i^{Ab}\hat{d}_i\right)\leq\left(\hat{d}_{k_2}-\hat{d}_i\right)\left(P_{k_2}^{Ab}\hat{d}_{k_2}-P_{k_1}^{Ab}\hat{d}_{k_1}\right), \tag{EC.1}$$

$$\left(\hat{d}_{k_2}-\hat{d}_{k_1}\right)P_i^{Ab}\hat{d}_i\geq\left(\hat{d}_i-\hat{d}_{k_1}\right)P_{k_2}^{Ab}\hat{d}_{k_2}+\left(\hat{d}_{k_2}-\hat{d}_i\right)P_{k_1}^{Ab}\hat{d}_{k_1}. \tag{EC.2}$$

Note that the above two equivalent conditions are identical to those in Remark EC1 of Tezcan and Zhang (2014), though in the context of general service and patience time distributions. Also, by (3), $P_{i'}>P_i$ if $i'>i$. We show in Lemma EC.1 that results of Lemma EC1 in Tezcan and Zhang (2014) holds in the current case as well. The proof is then identical to that of Lemma 1 in Tezcan and Zhang (2014). □

Lemma EC.1. *Assume that* (3) *and* (10) *hold.*

(i) *If for a level* j, $1<j<I$, $\hat{d}_j=\hat{d}_{j'}$ *for some* $j'<j$, *then level* j *cannot be efficient.*

(ii) *For any efficient level* i_j

$$\left(1-P_{i_j}^{Ab}\right)\hat{d}_{i_j}\geq\left(1-P_i^{Ab}\right)\hat{d}_i,\quad \textit{for all } i\leq i_j. \tag{EC.3}$$

Proof. Assume that for $j>1$, $\hat{d}_j=\hat{d}_{j'}$ for some $j'<j$. Set $k_1=j', i=j$ and set $k_2=I$. Then we have $\hat{d}_{k_1}=\hat{d}_i$. And by (7) and (10),

$$\begin{aligned}(\hat{d}_{k_2}-\hat{d}_{k_1})(P_{k_2}^{Ab}\hat{d}_{k_2}-P_i^{Ab}\hat{d}_i) &= (\hat{d}_{k_2}-\hat{d}_i)(P_{k_2}^{Ab}\hat{d}_{k_2}-P_i^{Ab}\hat{d}_{k_1})\\ &\le (\hat{d}_{k_2}-\hat{d}_i)(P_{k_2}^{Ab}\hat{d}_{k_2}-P_{k_1}^{Ab}\hat{d}_{k_1}).\end{aligned}$$

Hence j is inefficient by (EC.1).

Now assume that i_j is efficient. By (10) and by part (i) $\hat{d}_I>\hat{d}_{i_j}>\hat{d}_i$ for any $i<i_j$. Then, by (EC.2) and (10), for any $i<i_j$

$$\begin{aligned}P_{i_j}^{Ab}\hat{d}_{i_j} &\le \frac{\left(\hat{d}_{i_j}-\hat{d}_i\right)}{\left(\hat{d}_I-\hat{d}_i\right)}P_I^{Ab}\hat{d}_I+\frac{\left(\hat{d}_I-\hat{d}_{i_j}\right)}{\left(\hat{d}_I-\hat{d}_i\right)}P_i^{Ab}\hat{d}_i\\ &\le \frac{\left(\hat{d}_{i_j}-\hat{d}_i\right)}{\left(\hat{d}_I-\hat{d}_i\right)}\left(\hat{d}_I-\hat{d}_i+P_i^{Ab}d_i\right)+\frac{\left(\hat{d}_I-\hat{d}_{i_j}\right)}{\left(\hat{d}_I-\hat{d}_i\right)}P_i^{Ab}\hat{d}_i\\ &= \left(\hat{d}_{i_j}-\hat{d}_i\right)+P_i^{Ab}\hat{d}_i\end{aligned}$$

giving the desired result. □

Lemma EC.2. *If there exists $j=1,2,\ldots,I-1$ such that $\hat{d}_I<\hat{d}_j$ or $(1-P_I^{Ab})\hat{d}_I<(1-P_j^{Ab})\hat{d}_j$, then any optimal solution of the routing linear program (11)–(13) must satisfy $\lambda_I^*=0$.*

Proof. If $\hat{d}_I<\hat{d}_j$ for some $j=1,2,\ldots,I-1$ the results follows from the proof of Lemma 1. So assume that $\hat{d}_I\ge\hat{d}_j$ for all $j=1,2,\ldots,I-1$.

Assume that $(1-P_I^{Ab})\hat{d}_I<(1-P_j^{Ab})\hat{d}_j$ for some $j=1,2,\ldots,I-1$. We prove the result by contradiction. Assume that given λ, for an optimal solution we have $\lambda_I^*>0$. Choose level j such that $(1-P_I^{Ab})\hat{d}_I<(1-P_j^{Ab})\hat{d}_j$. Consider the following solution of the routing linear program (11)–(13),

$$\lambda_i=\begin{cases}\lambda_i^* & \text{if } i\in\{1,\ldots,I\}\setminus\{j,I\},\\ \lambda_j^*+\lambda_I^*\frac{\hat{d}_j}{\hat{d}_I} & \text{if } i=j,\\ 0 & \text{if } i=I,\\ \lambda_{I+1}^*+\lambda_I^*(1-\frac{\hat{d}_j}{\hat{d}_I}) & \text{if } i=I+1.\end{cases}$$

Note that $\hat{d}_j\le\hat{d}_I$. It can also be easily seen that the above is a feasible solution of the static planning problem (11)–(13). Moreover, the objective function value with this solution satisfies

$$\begin{aligned}&\sum_{i=1}^{I}\lambda_iP_i^{Ab}+\lambda_{I+1}-(\sum_{i=1}^{I}\lambda_i^*P_i^{Ab}+\lambda_{I+1}^*)\\ &=\lambda_I^*(1-P_I^{Ab})-\lambda_I^*(1-P_j^{Ab})\frac{\hat{d}_j}{\hat{d}_I}<0,\end{aligned}$$

where the last inequality follows from the assumption that $(1-P_I^{Ab})\hat{d}_I<(1-P_j^{Ab})\hat{d}_j$. Hence any optimal solution of (11)–(13) must satisfy $\lambda_I^*=0$. □

EC.2. Difference between the Original and the Modified Systems

In this section we present the results of the numerical experiments to compare the original CSC system with the modified system under Assumption 1. We consider three systems $1_1 - 3_1$ with different arrival rate and number of agents; see Table 1(a). for details. For patience and service time distributions, we use the combinations from Table 2(a). The service rate $\mu = \{4, 3.8, 3.3, 3, 2.75, 2.5\}$ is also kept the same as that of the first experiment set in §7.1. We apply our proposed policy π to the original and the modified systems.

Combination	System	$P^{Ab}{}_{\text{sim-original}}$	$P^{Ab}{}_{\text{sim-modified}}$	Difference
I_1	1_1	$0.2234(\pm 0.0004)$	$0.2231(\pm 0.0005)$	3×10^{-4}
	2_1	$0.2227(\pm 0.0004)$	$0.2225(\pm 0.0005)$	2×10^{-4}
	3_1	$0.2223(\pm 0.0003)$	$0.2220(\pm 0.0005)$	3×10^{-4}
II_1	1_1	$0.2520(\pm 0.0005)$	$0.2521(\pm 0.0009)$	1×10^{-4}
	2_1	$0.2513(\pm 0.0004)$	$0.2514(\pm 0.0003)$	1×10^{-4}
	3_1	$0.2510(\pm 0.0004)$	$0.2511(\pm 0.0003)$	1×10^{-4}
III_1	1_1	$0.1530(\pm 0.0003)$	$0.1529(\pm 0.0002)$	1×10^{-4}
	2_1	$0.1524(\pm 0.0003)$	$0.1522(\pm 0.0002)$	2×10^{-4}
	3_1	$0.1521(\pm 0.0003)$	$0.1519(\pm 0.0001)$	2×10^{-4}

Table EC.1 Comparison of simulation results P^{Ab} of the original and modified systems

Table EC.1 summarizes the difference in the steady-state abandonment probability between the original and the modified systems under a range of different parameter setting described in the above. Though the original CSC system and the modified one are equivalent under exponential service and patience time distributions, we also simulate the result as a benchmark. Obviously these two systems are also nearly identical for general distributions (other performance metrics besides abandonment probability are also very close).

EC.3. Analysis of the Fluid Model

Lemma EC.3. *Consider the CSC fluid model* (35)–(51). *If customers' service times and patience times during service follow exponential distributions, saying that* $G^c(x) = e^{-x}$ *and* $F^c(y) = e^{-\nu y}$, *then* (58) *and* (59) *hold.*

Proof. Setting x, y in (35) to be zero and plugging (57) yield

$$\begin{aligned}
i\bar{Z}_i(t) &= i\bar{Z}_i(0)e^{-(\mu_i+\nu)t} - (i-1)\int_0^t e^{-(\mu_i+\nu)(t-s)}d\bar{M}_{i,i-1}(s) + (i-1)\int_0^t e^{-(\mu_i+\nu)(t-s)}d\bar{M}_{i-1,i}(s) \\
&\quad + \int_0^t e^{-(\mu_i+\nu)(t-s)}d\bar{A}_i(s) + i\int_0^t e^{-(\mu_i+\nu)(t-s)}d\bar{M}_{i+1,i}(s) - i\int_0^t e^{-(\mu_i+\nu)(t-s)}d\bar{M}_{i,i+1}(s) \\
&= i\bar{Z}_i(0)e^{-(\mu_i+\nu)t} - (i-1)\int_0^t e^{-(\mu_i+\nu)(t-s)}d\bar{S}_i(s) + i\int_0^t e^{-(\mu_i+\nu)(t-s)}d\bar{A}_i(s) \\
&\quad + i\int_0^t e^{-(\mu_i+\nu)(t-s)}d\bar{S}_{i+1}(s) - i\int_0^t e^{-(\mu_i+\nu)(t-s)}d\bar{A}_{i+1}(s), \qquad \text{(EC.4)}
\end{aligned}$$

where the first equality follows from (35), and the second one is due to (41). One can check that the definition at $\bar{Z}(\cdot)=0$ in (36) and (37) is also satisfied.

On the other hand, by (31) and (57) we have $\bar{\mathcal{L}}_i(t)(\mathcal{A}(x,y))=i\bar{Z}_i(t)(1-e^{-(x+\nu y)})$. Plugging this to (45) yields

$$\begin{aligned}\bar{S}_i(t)&=i\bar{Z}_i(0)(1-e^{-(\mu_i+\nu)t})-(i-1)\int_0^t(1-e^{-(\mu_i+\nu)(t-s)})d\bar{M}_{i,i-1}(s)+(i-1)\int_0^t(1-e^{-(\mu_i+\nu)(t-s)})d\bar{M}_{i-1,i}(s)\\&\quad+\int_0^t(1-e^{-(\mu_i+\nu)(t-s)})d\bar{A}_i(s)+i\int_0^t(1-e^{-(\mu_i+\nu)(t-s)})d\bar{M}_{i+1,i}(s)-i\int_0^t(1-e^{-(\mu_i+\nu)(t-s)})d\bar{M}_{i,i+1}(s)\\&=i\bar{Z}_i(0)(1-e^{-(\mu_i+\nu)t})-(i-1)\int_0^t(1-e^{-(\mu_i+\nu)(t-s)})d\bar{S}_i(s)+i\int_0^t(1-e^{-(\mu_i+\nu)(t-s)})d\bar{A}_i(s)\\&\quad+i\int_0^t(1-e^{-(\mu_i+\nu)(t-s)})d\bar{S}_{i+1}(s)-i\int_0^t(1-e^{-(\mu_i+\nu)(t-s)})d\bar{A}_{i+1}(s),\end{aligned}$$

where the last equality follows from (41). Applying the chain rule yields

$$\begin{aligned}d\bar{S}_i(t)&=(\mu_i+\nu)\Big[i\bar{Z}_i(0)e^{-(\mu_i+\nu)t}-(i-1)\int_0^t e^{-(\mu_i+\nu)(t-s)}d\bar{S}_i(s)+i\int_0^t e^{-(\mu_i+\nu)(t-s)}d\bar{A}_i(s)\\&\qquad+i\int_0^t e^{-(\mu_i+\nu)(t-s)}d\bar{S}_{i+1}(s)-i\int_0^t e^{-(\mu_i+\nu)(t-s)}d\bar{A}_{i+1}(s)\Big]dt\\&=(\mu_i+\nu)i\bar{Z}_i(t)dt,\end{aligned}$$

where the last equality follows from (EC.4). This also proves (58). Thus, the above together with (EC.4) immediately implies (59). □

EC.4. Analysis of the Invariant State

Proof of Proposition 1. Let $(\bar{\mathcal{L}}(\infty),\bar{\mathcal{R}}(\infty))$ denote an invariant state. First we show that $\lambda_i=0$ for any inefficient level $i\notin\mathscr{F}$. Suppose that there exists $i\notin\mathscr{F}$ such that $\lambda_i>0$ and let k be one of these levels.. By (63), $\bar{A}_k(t)=\lambda_k t$, where $\lambda_k>0$. Based on our policy π described in (34), $p(k)<p(k-1)$ since $k\notin\mathscr{F}$. This together with (50) yields, for all $t\geq 0$,

$$0=\bar{Z}_k(t)d\bar{A}_k(t)=\lambda_k\bar{Z}_k(t)dt.$$

The above implies $\lambda_k\bar{Z}_k(t)=0$. Thus, $\bar{Z}_k(t)=0$ since $\lambda_k>0$. However, by (42) and (62), $\lambda_k=0$ whenever $\bar{Z}_k(t)=0$. This clearly is a contradiction. Therefore no such k exists. This proves that $\lambda_i=0$ for all $i\notin\mathscr{F}$.

Now we prove that there are at most two efficient levels with $\lambda_i>0$, $i\in\mathscr{F}$ and that there cannot be any efficient levels between these two levels. Let $i_{j+1}\in\mathscr{F}$ be the efficient level with the highest index among those efficient levels with $\lambda_i>0$. By (63), $\bar{A}_{i_{j+1}}(t)=\lambda_{i_{j+1}}t$ and from our assumption $\lambda_{i_{j+1}}>0$. Recall that $\mathscr{F}=\{i_1,i_2,\ldots,i_J\}$, where $i_1<i_2<\ldots<i_J$ (see §2.3.1). Again by our policy

π described in (34), we have $p(i) < p(i_{j+1}-1) = i_j$ for all $i = i_1, i_2, \ldots, i_{j-1}$. This together with (50) yields, for all $t \geq 0$,

$$0 = \bar{Z}_i(t) d\bar{A}_{i_{j+1}}(t) = \lambda_{i_{j+1}} \bar{Z}_i(t) dt, \quad i = i_1, i_2, \ldots, i_{j-1}.$$

Since $\lambda_{i_{j+1}} > 0$, the above implies $\bar{Z}_i(t) = 0$ for all $i = i_1, i_2, \ldots, i_{j-1}$. This with (42) and (62) yields $\lambda_i = 0$ for all $i = i_1, i_2, \ldots, i_{j-1}$, giving the desired result. □

Proof of Theorem 2. Let $(\bar{\mathcal{L}}(\infty), \bar{\mathcal{R}}(\infty))$ denote the state defined in the theorem. First assume that one of the conditions in (i), (ii), (iii) or (iv) of the theorem holds. Let

$$\bar{A}_i(t) = \lambda_i^* t. \tag{EC.5}$$

We prove that $(\bar{\mathcal{L}}(\infty), \bar{\mathcal{R}}(\infty))$ with $\bar{A}_i$ is an invariant state. By (5), $\hat{d}_i = i / \int_0^\infty G^c(\mu_i s) F^c(s) ds$.

From the conditions of the theorem, two basic levels are non-adjacent or there is only one basic level. Let i_j denote one of these levels. Then, $\bar{Z}_{i_j-1}(t) = \bar{Z}_{i_j+1}(t) = 0$ for all $t \geq 0$. By the definition of the invariant state (64)–(68) and (EC.5), (36) and (37) are satisfied if $\bar{Z}_i(\infty) = 0$. Then, (64)–(68) satisfy (35) for all $i = 1, \ldots, I$. Other fluid model equations (38)–(51) are verified similarly. One can verify that $\bar{\mathcal{R}}(\infty)$ defined as in (69) is a solution to (46) in a similar way. □

Proof of Theorem 3. Assume that $\hat{d}_1 N < \lambda < \hat{d}_I N$, $\hat{d}_{i^*_{j+1}} > \lambda/N$ and $i^*_{j+1} = i^*_j + 1$. Let $\bar{\mathcal{L}}(\infty)$ defined as in (64)–(68) and $\bar{A}_i$ be defined as in (EC.5) for all $i = 1, \ldots, I$. By the definition of the invariant state (64)–(68) and (EC.5), (36) and (37) are satisfied if $\bar{Z}_i(\infty) = 0$. Therefore, (35) becomes

$$\begin{aligned}
\bar{\mathcal{L}}_{i^*_j}(\infty)(C_x \times C_y) = {} & \lambda^*_{i^*_j} \int_0^\infty G^c(x + \mu_{i^*_j} s) F^c(y+s) ds \\
& + i^*_j \lambda^*_{i^*_{j+1}} \int_0^\infty \frac{1}{i^*_{j+1} \bar{Z}_{i^*_{j+1}}(\infty)} \bar{\mathcal{L}}_{i^*_{j+1}}(\infty)(C_{x+\mu_{i^*_j} s} \times C_{y+s}) ds \\
& - i^*_j \lambda^*_{i^*_{j+1}} \int_0^\infty \frac{1}{i^*_j \bar{Z}_{i^*_j}(\infty)} \bar{\mathcal{L}}_{i^*_j}(\infty)(C_{x+\mu_{i^*_j} s} \times C_{y+s}) ds,
\end{aligned} \tag{EC.6}$$

$$\begin{aligned}
\bar{\mathcal{L}}_{i^*_{j+1}}(\infty)(C_x \times C_y) = {} & \lambda^*_{i^*_{j+1}} \int_0^\infty G^c(x + \mu_{i^*_{j+1}} s) F^c(y+s) ds \\
& - i^*_j \lambda^*_{i^*_{j+1}} \int_0^\infty \frac{1}{i^*_{j+1} \bar{Z}_{i^*_{j+1}}(\infty)} \bar{\mathcal{L}}_{i^*_{j+1}}(\infty)(C_{x+\mu_{i^*_{j+1}} s} \times C_{y+s}) ds \\
& + i^*_j \lambda^*_{i^*_{j+1}} \int_0^\infty \frac{1}{i^*_j \bar{Z}_{i^*_j}(\infty)} \bar{\mathcal{L}}_{i^*_j}(\infty)(C_{x+\mu_{i^*_{j+1}} s} \times C_{y+s}) ds,
\end{aligned} \tag{EC.7}$$

and $\bar{\mathcal{L}}_i(\infty) = \mathbf{0}$ for all $i \neq i^*_j, i^*_{j+1}$. Note that the last two terms on the right-hand side of (EC.6) and (EC.7) are the interactions between the two adjacent basic levels.

Assume that Condition I holds. By (64)

$$\frac{1}{i \bar{Z}_i(\infty)} \bar{\mathcal{L}}_i(\infty)(C_x \times C_y) = e^{-x} e^{-vy} \quad \text{for } i = i^*_j, i^*_{j+1}.$$

This satisfies (EC.6) and (EC.7).

Now assume that Condition II holds. By (64)

$$\bar{\mathcal{L}}_i(\infty)(C_x \times C_y) = \lambda_i^* \int_0^\infty G^c(x+\mu_i s)ds \quad \text{for } i = i_j^*, i_{j+1}^*.$$

This with (42) gives

$$\frac{1}{i\bar{Z}_i(\infty)}\bar{\mathcal{L}}_i(\infty)(C_x \times C_y) = \frac{\int_0^\infty G^c(x+\mu_i s)ds}{\int_0^\infty G^c(\mu_i s)ds} = \int_0^\infty G^c(x+s)ds \quad \text{for } i = i_j^*, i_{j+1}^*.$$

This clearly satisfies (EC.6) and (EC.7). The other fluid model equations can be checked similarly.

□

EC.5. Analysis of the Stochastic Model

By (22)

$$\begin{aligned}
\bar{\mathcal{L}}_i^n(t)(C_x \times C_y) = {} & \bar{\mathcal{L}}_i^n(t_0)(C_{x+\mu_i(t-t_0)} \times C_{y+(t-t_0)}) \\
& - \int_{t_0}^t \frac{1}{n}\Phi^{i-1}\left(\mathcal{L}_i^n(s)\right)(C_{x+\mu_i(t-s)} \times C_{y+t-s})dM_{i,i-1}^n(s) \\
& + \int_{t_0}^t \frac{1}{n}\Phi^{i-1}\left(\mathcal{L}_{i-1}^n(s)\right)(C_{x+\mu_i(t-s)} \times C_{y+t-s})dM_{i-1,i}^n(s) \\
& + \frac{1}{n}\sum_{j=A_i^n(t_0)+1}^{A_i^n(t)} \mathbb{1}_{\{v_{i,j}^n > x+\mu_i(t-\tau_{i,j}^n), u_{i,j}^n > y+t-\tau_{i,j}^n\}} \\
& + \int_{t_0}^t \frac{1}{n}\Phi^{i}\left(\mathcal{L}_{i+1}^n(s)\right)(C_{x+\mu_i(t-s)} \times C_{y+t-s})dM_{i+1,i}^n(s) \\
& - \int_{t_0}^t \frac{1}{n}\Phi^{i}\left(\mathcal{L}_{i}^n(s)\right)(C_{x+\mu_i(t-s)} \times C_{y+t-s})dM_{i,i+1}^n(s).
\end{aligned} \tag{EC.8}$$

Actually, the above equation is a shifted fluid scaled dynamic equation treating t_0 as a start point. When setting $t_0 = 0$, (EC.8) becomes the fluid scaled version of (22). Let $\bar{S}_i^n(t_0,t) = \bar{S}^n(t) - \bar{S}^n(t_0)$. Then, similar to (EC.8), we can also shift (32) to t_0 and it becomes

$$\begin{aligned}
\bar{S}_i^n(t_0,t) = {} & \bar{\mathcal{L}}_i^n(t_0)(\mathcal{A}_i(\mu_i(t-t_0), t-t_0)) \\
& - \int_{t_0}^t \frac{1}{n}\Phi^{i-1}\left(\mathcal{L}_i^n(s)\right)(\mathcal{A}(\mu_i(t-s),(t-s)))dM_{i,i-1}^n(s) \\
& + \int_{t_0}^t \frac{1}{n}\Phi^{i-1}\left(\mathcal{L}_{i-1}^n(s)\right)(\mathcal{A}(\mu_i(t-s),(t-s)))dM_{i-1,i}^n(s) \\
& + \frac{1}{n}\sum_{j=A_i^n(t_0)+1}^{A_i^n(t)} \mathbb{1}_{\{v_{i,j}^n \le \mu_i(t-\tau_{i,j}^n) \text{ or } u_{i,j}^n \le t-\tau_{i,j}^n\}} \\
& + \int_{0}^t \Phi^{i}\left(\mathcal{L}_{i+1}^n(s)\right)(\mathcal{A}(\mu_i(t-s),(t-s)))dM_{i+1,i}^n(s) \\
& - \int_{0}^t \Phi^{i}\left(\mathcal{L}_{i}^n(s)\right)(\mathcal{A}(\mu_i(t-s),(t-s)))dM_{i,i+1}^n(s).
\end{aligned} \tag{EC.9}$$

We assume all random variables and processes associated with the nth system are defined on the probability space $(\Omega^n, \mathcal{F}^n, \mathbb{P}^n)$.

EC.5.1. Tightness

To prove the tightness we use an approach similar to that in Zhang (2013), which studied measure-valued process underlying a many-server queue with a single pool and a single customer class. By Theorem 3.7.2 in Ethier and Kurtz (1986), it suffices to verify (a) the compact containment condition and (b) the oscillation bound in the following two subsections.

EC.5.1.1. Compact Containment The objective of this subsection is to prove the compact containment, Lemma EC.4 below. To state the result, we need to introduce the concept of compactness in the space of measures. Let $\mathbf{M}$ denote the space of all non-negative Borel measures on $\mathbb{R}_+^2$ equipped with Prohorov metric (see §6 in Billingsley (1999) for details). A set $\mathbf{K} \subset \mathbf{M}$ is relatively compact if $\sup_{\xi \in \mathbf{K}} \xi(\mathbb{R}_+^2) < \infty$, and there exists a sequence of nested compact sets $\mathcal{B}_j \subset \mathbb{R}_+^2$ such that $\cup \mathcal{B}_j = \mathbb{R}_+^2$ and

$$\lim_{j \to \infty} \sup_{\xi \in \mathbf{K}} \xi(\mathcal{B}_j^c) = 0, \tag{EC.10}$$

where $\mathcal{B}_j^c$ denotes the complement of $\mathcal{B}_j$; see Kallenberg (1986), Theorem A7.5.

Lemma EC.4. *Fix $T > 0$. For any $\eta > 0$ there exists a compact set $\mathbf{K} \subset \mathbf{M}$ such that*

$$\liminf_{n \to \infty} \mathbb{P}^n \Big(\bar{\mathcal{R}}^n(t) \in \mathbf{K} \text{ and } \bar{\mathcal{L}}_i^n(t) \in \mathbf{K} \text{ for all } i = 1, \ldots, I \text{ and } t \in [0, T] \Big) \geq 1 - \eta.$$

To make the presentation self-contained, we briefly cite the Glivenko-Cantelli estimates (e.g., Appendix B in Zhang (2013)). Define

$$\bar{\mathcal{E}}_i^n(l) = \frac{1}{n} \sum_{j=1}^{\lfloor nl \rfloor} \delta_{(v_{i,j}^n, u_{i,j}^n)}, \tag{EC.11}$$

where $\delta_{(x,y)}$ denotes the Dirac measure of point (x, y) on $\mathbb{R} \times \mathbb{R}$. Recall that $\{v_{i,j}^n\}_{j=1}^\infty$ is i.i.d. sequence of random variables following distribution G, and $\{u_{i,j}^n\}_{j=1}^\infty$ is i.i.d. sequence of random variables following distribution F. Denote ν_G and ν_F the probability measures corresponding to the service time distribution G and the patience time distribution during service F, respectively. Introduce the family of testing functions

$$\mathcal{V} = \left\{ 1_{C_x \times C_y}(\cdot, \cdot) : x, y \in \mathbb{R} \right\}.$$

There exists a function $\bar{f} : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}$ satisfying

$$\bar{f} \text{ is increasing and unbounded,}$$
$$f \leq \bar{f} \text{ for all } f \in \mathcal{V},$$
$$\langle \bar{f}^2, \nu \rangle < \infty,$$

where $\langle \bar{f}^2, \nu \rangle$ denote the integration of function $\bar{f}^2$ with respect to measure ν, see Appendix B in Zhang (2013). Denote $\bar{\mathcal{V}} = \{\bar{f}\} \cup \mathcal{V}$. The function $\bar{f}$ is referred to as the envelop function for $\mathcal{V}$. It follows from Lemma 5.1 in Zhang et al. (2009) and Lemma B.1 in Zhang (2013) that

$$\lim_{n\to\infty} \mathbb{P}^n\Big(\Omega^n_{\mathrm{GC}}(L)\Big) = 1, \tag{EC.12}$$

for any fixed $L > 0$, where the event $\Omega^n_{\mathrm{GC}}(L)$ is defined as

$$\Omega^n_{\mathrm{GC}}(L) = \Big\{ \max_{i\in\{1,\ldots,I\}} \sup_{l\in[0,L]} \sup_{f\in\bar{\mathcal{V}}} \Big| \langle f, \bar{\mathcal{E}}^n_i(l)\rangle - l\langle f, (\nu_F, \nu_G)\rangle \Big| \le \epsilon_{\mathrm{GC}}(n) \Big\}, \tag{EC.13}$$

for some function $\epsilon_{\mathrm{GC}}(\cdot)$ which vanishes at infinity. Intuitively, on the event $\Omega^n_{\mathrm{GC}}(L)$ (whose probability goes to 1 as $n \to \infty$), the measures $\bar{\mathcal{E}}^n_i(l)$ is "close" to ν.

We also need to introduce another "good" events to work with later in our analysis. It follows from condition (52) and Lemma 5.2 in Zhang et al. (2009) that

$$\lim_{n\to\infty} \mathbb{P}^n\Big(\Omega^n_{\Lambda}(T)\Big) = 1, \tag{EC.14}$$

for any fixed $T > 0$, where the event Ω^n_{Λ} is defined as

$$\Omega^n_{\Lambda}(T) = \Big\{ \sup_{t\in[0,T]} |\bar{\Lambda}^n(t) - \lambda t| < \epsilon_E(n) \Big\}, \tag{EC.15}$$

for some function $\epsilon_E(\cdot)$ which vanishes at infinity.

Proof of Lemma EC.4. The buffer part of the customer service chat systems is identical that of the call center model studied in Zhang (2013), therefore the compact containment property of $\bar{\mathcal{R}}^n$ follows from the same argument in Lemma 5.1 of Zhang (2013). Hence we mainly focus on the compact containment property of $\bar{\mathcal{L}}^n_i$.

Fix $\eta > 0$. First, for any $i \in \{1, \ldots, I\}$ and $t \ge 0$, $\bar{\mathcal{L}}^n_i(t)(\mathbb{R}^2_+) \le I \cdot N < \infty$ due to the fluid scaling. It remains to verify (EC.10). By the convergence of the initial condition (54), for any $\epsilon > 0$, there exists a relatively compact set $\mathbf{K}_0 \subset \mathbf{M}$ such that

$$\liminf_{n\to\infty} \mathbb{P}^n\Big(\bar{\mathcal{R}}^n(0) \in \mathbf{K}_0 \text{ and } \bar{\mathcal{L}}^n_i(0) \in \mathbf{K}_0 \text{ for all } i \in \{1, \ldots, I\}\Big) \ge 1 - \eta/2.$$

Denote the event in the above probability by $\Omega^n_{0,c}$. On this event, by the definition of relatively compact set in the space $\mathbf{M}$, there exists a function $\kappa_0(\cdot)$ with $\lim_{x\to\infty} \kappa_0(x) = 0$ such that

$$\sum_{i=1}^{I} \bar{\mathcal{L}}^n_i(0)(C_x \times C_x) \le \kappa_0(x). \tag{EC.16}$$

Define

$$\mu_{\min} = \min_{i\in\{1,\ldots,I\}} \mu_i. \tag{EC.17}$$

On the event Ω_Λ^n,

$$0 \le \bar{\Lambda}^n(t) \le 2\lambda T, \quad t \in [0,T] \tag{EC.18}$$

for all large enough n. By (EC.8), on the event Ω_Λ we have

$$\begin{aligned}\sum_{i=1}^I \bar{\mathcal{L}}_i^n(t)(C_x \times C_x) &\le \sum_{i=1}^I \bar{\mathcal{L}}_i^n(0)(C_{x+\mu_{\min}t} \times C_{x+t}) + \sum_{i=1}^I \frac{1}{n} \sum_{j=1}^{A_i^n(t)} \mathbb{1}_{\{v_{i,j}^n > x+\mu_{\min}(t-\tau_{i,j}^n), u_{i,j}^n > x+t-\tau_{i,j}^n\}} \\ &\le \kappa_0(x) + \sum_{i=1}^I \frac{1}{n} \sum_{j=1}^{A_i^n(t)} \mathbb{1}_{\{v_{i,j}^n > x, u_{i,j}^n > x\}}. \end{aligned} \tag{EC.19}$$

We can think of (EC.19) in terms of the total mass: at time t, those with remaining service time large than x must be either one of those initially in the system and with remaining service time larger than $x+\mu_{\min}t$ or arrive after with remaining service time larger than $x+\mu_{\min}(t-\tau)$ if he arrives at time τ. Since we are on the event Ω_Λ^n (see (EC.18)) and $\Omega_{\rm GC}^n(2\lambda T)$,

$$\langle \bar{f}, \frac{1}{n} \sum_{i=1}^I \sum_{j=1}^{A_i^n(t)} \delta_{(v_{i,j}^n, u_{i,j}^n)} \rangle \le 2\lambda T \langle \bar{f}, \nu \rangle + 1,$$

for all large n. Applying Markov's inequality to (EC.19),

$$\bar{\mathcal{L}}_i^n(t)(C_x \times C_x) \le \kappa_0(x) + \frac{2\lambda T \langle \bar{f}, \nu \rangle + 1}{\bar{f}(x)}, \tag{EC.20}$$

which converges to 0 as $x \to \infty$. So we can define the set $\mathbf{K} = \{\xi \in \mathbf{M} : \xi(\mathbb{R}\times\mathbb{R}) \le 1, \xi(C_x \times C_x) \le \kappa_0(x) + \frac{2\lambda T\langle \bar{f}, \nu\rangle + 1}{\bar{f}(x)}\}$, which is compact in $\mathbf{M}$ according to the definition. On the event $\Omega_{0,c}^n \cap \Omega_\Lambda^n \cap \Omega_{\rm GC}^n(2\lambda T)$ (which has probability larger than $1-\eta$ for all large n), $\bar{\mathcal{L}}_i^n(t) \in \mathbf{K}$ for all $i \in \{1,\dots,I\}$ and $t \in [0,T]$. Thus the desired result follows from (EC.12) and (EC.14). $\square$

EC.5.1.2. Oscillation Bound The oscillation of a function $\zeta(\cdot)$ taking values in the metric space $\mathbf{M}$ with metric $\mathbf{d}$ on a fixed interval $[0,T]$ is defined as

$$\mathbf{w}_T(\zeta(\cdot), \delta) = \sup_{s,t \in [0,T], |s-t|<\delta} \mathbf{d}[\zeta(s), \zeta(t)].$$

If the metric space is $\mathbb{R}$, we just use the Euclidean metric; if the space is all finite measures, we use the Prohorov metric defined in §6 of Billingsley (1999). For $\nu_1, \nu_2 \in \mathbf{M}$, the Prohorov metric is defined as

$$\begin{aligned}\mathbf{d}[\nu_1,\nu_2] = \inf \big\{ &\epsilon > 0 : \nu_1(A) \le \nu_2(A^\epsilon) + \epsilon \text{ and} \\ &\nu_2(A) \le \nu_1(A^\epsilon) + \epsilon \text{ for all Borel set } A \subset \mathbb{R} \big\}, \end{aligned} \tag{EC.21}$$

where $A^\epsilon = \{b \in \mathbb{R} : \inf_{a\in A} |a-b| < \epsilon\}$.

The second major step to prove tightness is to show that the oscillation is small with large probability, we show this next.

Lemma EC.5. *Fix $T>0$. For each $\epsilon,\eta>0$ there exists a $\delta>0$ (depending on ϵ and η) such that*

$$\liminf_{n\to\infty}\mathbb{P}^n\Big(\mathbf{w}_T(\bar{\mathcal{R}}^n(\cdot),\delta)\le\epsilon\Big)\ge 1-\eta, \tag{EC.22}$$

$$\liminf_{n\to\infty}\mathbb{P}^n\Big(\max_{i\in\{1,\dots,I\}}\mathbf{w}_T(\bar{\mathcal{L}}_i^n(\cdot),\delta)\le\epsilon\Big)\ge 1-\eta, \tag{EC.23}$$

$$\liminf_{n\to\infty}\mathbb{P}^n\Big(\max_{i\in\{1,\dots,I\}}\mathbf{w}_T(\bar{A}_i^n(\cdot),\delta)\le\epsilon\Big)\ge 1-\eta, \tag{EC.24}$$

$$\liminf_{n\to\infty}\mathbb{P}^n\Big(\max_{i\in\{1,\dots,I\}}\mathbf{w}_T(\bar{S}_i^n(\cdot),\delta)\le\epsilon\Big)\ge 1-\eta. \tag{EC.25}$$

The rest of this section is devoted to the proof of this result. We begin with the following auxiliary result. Given $\kappa>0$ we define

$$\Delta_\kappa(x,y):=C_x\times C_y\setminus C_{x+\kappa}\times C_{y+\kappa}. \tag{EC.26}$$

Lemma EC.6. *Fix $T>0$. For each $\epsilon,\eta>0$ there exists $\kappa>0$ (depending on ϵ and η) such that*

$$\liminf_{n\to\infty}\mathbb{P}^n\Big(\max_{i\in\{1,\dots,I\}}\sup_{t\in[0,T]}\sup_{x,y\in\mathbb{R}_+}\bar{\mathcal{L}}_i^n(t)(\Delta_\kappa(x,y))\le\epsilon\Big)\ge 1-\eta. \tag{EC.27}$$

Proof. Fix $\epsilon>0$ and $\eta>0$. Similar to the proof of Lemma EC.4, we only consider the event $\Omega_{0,c}^n\cap\Omega_\Lambda^n\cap\Omega_{\mathrm{GC}}^n$ for the rest of this proof. The customers who receive service must be either those initially in the server pool or those who arrive after time 0. We index the customers initially at level i by $l=1,\dots,iZ_i^n(0)$ according to the time spent during service $w_{i,l}^n$ by time 0. Recall that $w_{i,l}^n$ is assumed to be bounded. Also let $s_{i,l}^n$ denote the amount of service of lth such received by time 0. And $v_{i,l}^{n,o}$ and $u_{i,l}^{n,o}$ denote the remaining service time and remaining patience during service of the lth such customer. In view of (1), we use $\mu_{i,l}^o(s)$, $s\in[0,t]$, to denote the service rate of this customer at time s. Similarly, we index those customers who arrived after time 0 and whose service commences at level i by $j=1,\dots,A_i^n(t)$ based on their service start time $\tau_{i,j}^n$ for $j=1,\dots,A_i^n(t)$. We also use $\mu_{i,j}(s)$, $s\in[0,t]$, to denote the service rate of the jth customer at time s. We use $v_{i,j}^n$ and $u_{i,j}^n$ to denote the service time and patience time during service of the jth such customer, respectively. Then by the definition of $\bar{\mathcal{L}}_i^n(t)$, we have

$$\begin{aligned}\sum_{i=1}^I\bar{\mathcal{L}}_i^n(t)(\Delta_\kappa(x,y))=&\frac{1}{n}\sum_{i=1}^I\sum_{l=1}^{iZ_i^n(0)}\delta_{(v_{i,l}^{n,o},u_{i,l}^{n,o})}\big(\Delta_\kappa(x+\int_0^t\mu_{i,l}^o(s)ds,y+t)\big)\\&+\frac{1}{n}\sum_{i=1}^I\sum_{j=1}^{A_i^n(t)}\delta_{(v_{i,j}^n,u_{i,j}^n)}\big(\Delta_\kappa(x+\int_{t-\tau_{i,j}^n}^t\mu_{i,j}(s)ds,y+t-\tau_{i,j}^n)\big).\end{aligned}$$

Note that $\delta_{(v_{i,l}^{n,o},u_{i,l}^{n,o})}\big(\Delta_\kappa(x+\int_0^t\mu_{i,l}^o(s)ds,y+t)\big)$, $l=1,\dots,iZ_i^n(0)$, and $\delta_{(v_{i,j}^n,u_{i,j}^n)}\big(\Delta_\kappa(x+$

$\int_{t-\tau_{i,j}^n}^t \mu_{i,j}(s)ds, y+t-\tau_{i,j}^n)\big)$, $j=1,\ldots,A_i^n(t)$, are Bernoulli random variables, which are independent and their variances are all bounded by 1. Then by Kolmogorov's strong law of large numbers (Theorem 2.3.10 in Sen and Singer (1994)), we have a.s.

$$\begin{aligned}\sum_{i=1}^I \bar{\mathcal{L}}_i^n(t)(\Delta_\kappa(x,y)) \le \frac{1}{n}\sum_{i=1}^I \sum_{l=1}^{iZ_i^n(0)} \mathbb{E}\left[\delta_{(v_{i,l}^{n,o},u_{i,l}^{n,o})}\big(\Delta_\kappa(x+\int_0^t \mu_{i,l}^o(s)ds, y+t)\big)\right] \\ + \frac{1}{n}\sum_{i=1}^I \sum_{j=1}^{A_i^n(t)} \mathbb{E}\left[\delta_{(v_{i,j}^n,u_{i,j}^n)}(\Delta_\kappa(x+\int_{t-\tau_{i,j}^n}^t \mu_{i,j}(s)ds, y+t-\tau_{i,j}^n))\right] + \frac{\epsilon}{2}\end{aligned} \tag{EC.28}$$

for all large n.

Now we consider the first term on the right-hand side of (EC.28). We have

$$\begin{aligned}&\mathbb{E}\left[\delta_{(v_{i,l}^{n,o},u_{i,l}^{n,o})}\big(\Delta_\kappa(x+\int_0^t \mu_{i,l}^o(s)ds, y+t)\big)\right] \\ &\le \mathbb{E}\left[\mathbb{1}_{\{v_{i,l}^{n,o}\in(x+\int_0^t \mu_{i,l}^o(s)ds, x+\int_0^t \mu_{i,l}^o(s)ds+\kappa]\}}\right] + \mathbb{E}\left[\mathbb{1}_{\{u_{i,l}^{n,o}\in(y+t,y+t+\kappa]\}}\right] \\ &= \frac{1}{G^c(s_{i,l}^n)}\left[G\Big(x+s_{i,l}^n+\int_0^t \mu_{i,l}^o(s)ds+\kappa\Big) - G\Big(x+s_{i,l}^n+\int_0^t \mu_{i,l}^0(s)ds\Big)\right] \\ &\quad + \frac{1}{F^c(w_{i,l}^n)}\left[F(y+w_{i,l}^n+t+\kappa) - F(y+w_{i,l}^n+t)\right].\end{aligned}$$

Note that because $v_{i,l}^{n,o}$ is the remaining service time of the lth customer initially at level i, it follows distribution function $1-\frac{G^c(s_{i,l}^n+x)}{G^c(s_{i,l}^n)}$. Similarly, for this customer, $u_{i,l}^{n,o}$. Because $w_{i,l}^n$ is bounded, $s_{i,l}^n$ is also bounded because $s_{i,l}^n \le \mu_{\max} w_{i,l}^n$, where $\mu_{\max} = \max_{i\in\{1,\ldots,I\}} \mu_i$. Therefore,

$$\begin{aligned}&\frac{1}{G^c(s_{i,l}^n)}\left[G\Big(x+s_{i,l}^n+\int_0^t \mu_{i,l}^o(s)ds+\kappa\Big) - G\Big(x+s_{i,l}^n+\int_0^t \mu_{i,l}^o(s)ds\Big)\right] \\ &+ \frac{1}{F^c(w_{i,l}^n)}\left[F(y+w_{i,l}^n+t+\kappa) - F(y+w_{i,l}^n+t)\right] \le \frac{\epsilon}{4NI^2}\end{aligned}$$

for κ small enough, where N is defined in (52). It then follows from the above two inequalities that

$$\frac{1}{n}\sum_{i=1}^I \sum_{l=1}^{iZ_i^n(0)} \mathbb{E}\left[\delta_{(v_{i,l}^{n,o},u_{i,l}^{n,o})}\big(\Delta_\kappa(x+\int_0^t \mu_{i,l}^o(s)ds, y+t)\big)\right] \le \sum_{i=1}^I i\bar{Z}_i^n(0)\frac{\epsilon}{4NI^2} \le \frac{\epsilon}{4}.$$

Now we consider the second term on the right-hand side of (EC.28). By (29) and (30), we have

$$\bar{A}_i^n(t) \le \bar{E}^n(t) \le \bar{\Lambda}^n(t) + \bar{Q}^n(0) \le M_0 + 2\lambda T. \tag{EC.29}$$

In fact, on the event $\Omega_{0,c}^n \cap \Omega_\Lambda^n$, $\bar{Q}^n(0) < M_0$ for some constant M_0 by Lemma EC.4 and $\bar{\Lambda}^n(t) \le 2\lambda T$. On the other hand,

$$\mathbb{E}\left[\delta_{(v_{i,j}^n,u_{i,j}^n)}(\Delta_\kappa(x+\int_{t-\tau_{i,j}^n}^t \mu_{i,j}(s)ds, y+t-\tau_{i,j}^n))\right]$$

$$\leq \mathbb{E}\left[\mathbb{1}_{\{v^n_{i,j}\in(x+\int_{t-\tau^n_{i,j}}^t \mu_{i,j}(s)ds, x+\int_{t-\tau^n_{i,j}}^t \mu_{i,j}(s)ds+\kappa]\}}\right] + \mathbb{E}\left[\mathbb{1}_{\{u^n_{i,j}\in(y+t-\tau^n_{i,j}, y+t-\tau^n_{i,j}+\kappa]\}}\right]$$
$$= G\Big(x+\int_{t-\tau^n_{i,j}}^t \mu_{i,l}(s)ds+\kappa\Big) - G\Big(x+\int_{t-\tau^n_{i,j}}^t \mu_{i,l}(s)ds\Big) + F(y+t+\kappa) - F(y+t).$$

Also

$$G\Big(x+\int_{t-\tau^n_{i,j}}^t \mu_{i,l}(s)ds+\kappa\Big) - G\Big(x+\int_{t-\tau^n_{i,j}}^t \mu_{i,l}(s)ds\Big) + F(y+t+\kappa) - F(y+t) \leq \frac{\epsilon}{4I(M_0+2\lambda T)}$$

for κ small enough. Then we can conclude from the above two inequalities and (EC.29) that

$$\frac{1}{n}\sum_{i=1}^I \sum_{j=1}^{A^n_i(t)} \mathbb{E}\left[\delta_{(v^n_{i,j},u^n_{i,j})}(\Delta_\kappa(x+\int_{t-\tau^n_{i,j}}^t \mu_{i,j}(s)ds, y+t-\tau^n_{i,j}))\right] \leq \sum_{i=1}^I \bar{A}^n_i(t)\frac{\epsilon}{4I(M_0+2\lambda T)} \leq \frac{\epsilon}{4}.$$

It then follows from (EC.28) that $\sum_{i=1}^I \bar{\mathcal{L}}^n_i(t)(\Delta_\kappa(x,y)) \leq \epsilon$. This completes the proof. □

Proof of Lemma EC.5. Fix $\epsilon>0$ and $\eta>0$. Similar to the compact containment property of $\bar{\mathcal{R}}^n$ in Lemma EC.4, the proof of the oscillation of $\bar{\mathcal{R}}^n$ in (EC.22) also follows from the same argument in Lemma 5.4 of of Zhang (2013). So we will focus on (EC.23)–(EC.25). To this end, we just need to restrict the stochastic processes on the event $\Omega^n_{0,c}\cap\Omega^n_{t,s}\cap\Omega^n_\Lambda\cap\Omega^n_{\rm GC}$, which has probability larger than $1-\eta$ for large enough n. Note that $\Omega^n_{t,s}$ is denoted to be the event in (EC.27).

Fix $\delta>0$ and choose $t_0<t$ such that $t-t_0<\delta$. We use Lemma EC.6 to study the oscillations in the departure process during this interval. To simplify the notation, let $\bar{f}^n(t_0,t)=f(t)-f(t_0)$, for any function f. Recall that $\mathcal{A}(t)(x,y)=\{(x',y')\in\mathbb{R}^2_+ : x'\leq x \text{ or } y'\leq y\}$ is the compliment of $C_x\times C_y$ defined in (31). Define $\mu_{\max}=\max_{i\in\{1,\dots,I\}}\mu_i$. By (EC.9) for any level i

$$\sum_{i=1}^I \bar{S}^n_i(t_0,t) \leq \sum_{i=1}^I \bar{\mathcal{L}}^n_i(t_0)(\mathcal{A}(\mu_{\max}(t-t_0),t-t_0)) + \sum_{i=1}^I \frac{1}{n}\sum_{j=A^n_i(t_0)+1}^{A^n_i(t)} \mathbb{1}_{\{v^n_{i,j}\leq\mu_{\max}(t-\tau^n_{i,j}) \text{ or } u^n_{i,j}\leq t-\tau^n_{i,j}\}}. \tag{EC.30}$$

This follows from the argument we use to arrive (EC.19); the departures during $(t_0,t]$ must be either those customers initially in system at t_0 and with remaining service time less than $\mu_{\max}(t-t_0)$ or remaining patience time less than $t-t_0$, or those newly arrivals with remaining service time less than $\mu_{\max}(t-\tau)$ or remaining patience time less than $t-\tau$ if the customer arrives at time τ.

By (EC.27), we can choose δ sufficiently small such that the first term on the right-hand side of (EC.30) is less than $\epsilon/2$. On the other hand, we have shown in (EC.29) that $\bar{A}^n_i(t)\leq M_0+2\lambda T$, where M_0 is chosen as in (EC.29). Since $\tau^n_{i,j}\in[t_0,t]$

$$\frac{1}{n}\sum_{j=A^n_i(t_0)+1}^{A^n_i(t)} \mathbb{1}_{\{v^n_{i,j}\leq\mu_{\max}(t-\tau^n_{i,j}) \text{ or } u^n_{i,j}\leq t-\tau^n_{i,j}\}}$$

$$\leq \frac{1}{n}\sum_{j=A_i^n(t_0)+1}^{A_i^n(t)} \mathbb{1}_{\{v_{i,j}^n \leq \mu_{\max}(t-t_0) \text{ or } u_{i,j}^n \leq t-t_0\}}$$
$$\leq (\bar{A}_i^n(t) - \bar{A}_i^n(t_0))(G(\mu_{\max}(t-t_0)) + F(t-t_0)) + \frac{\epsilon}{4I},$$

where the last inequality in the above follows from Glivenko-Cantelli estimate (EC.13). For distribution functions F and G, we can choose δ small enough such that,

$$G(\mu_{\max}(t-t_0)) + F(t-t_0) \leq \frac{\epsilon}{4I(M_0 + 2\lambda T)}.$$

Hence, from the above two inequalities, we can conclude that the second term on the right-hand side of (EC.30) is bounded by $\epsilon/2$. Thus,

$$\sum_{i=1}^{I} \bar{S}_i^n(t_0, t) \leq \epsilon, \tag{EC.31}$$

for t and t_0 close enough. This proves (EC.25).

By the definition of Ω_Λ^n in (EC.15), we have $\bar{\Lambda}^n(t_0, t) \leq \epsilon$, for $t - t_0$ small enough. Because each customer enters service either upon a service completion or if upon arrival we have

$$\sum_{i=1}^{I} \bar{A}_i^n(t_0, t) \leq \bar{\Lambda}^n(t_0, t) + \sum_{i=1}^{I} \bar{S}_i^n(t_0, t) \leq 2\epsilon. \tag{EC.32}$$

Thus (EC.24) holds.

Next we prove the oscillation bound for $\bar{\mathcal{L}}_i^n$. Let $C \subset \mathbb{R}_+^2$ be a Borel subset and define the "shift" of set C by (a, b) as

$$C + (a, b) = \{(x + a, y + b) | (x, y) \in C\}.$$

Note that the fluid scaled dynamic equation (EC.8) still holds for any such Borel set if we replace $C_x \times C_y$ and $C_{x+\mu_i(t-s)} \times C_{y+t-s}$ with any Borel set C and its shift $C + (\mu_i(t-s), t-s)$, respectively. Thus (EC.8) becomes

$$\begin{aligned}
\bar{\mathcal{L}}_i^n(t)(C) = \bar{\mathcal{L}}_i^n(t_0)(C + (\mu_i(t-t_0), t-t_0)) \\
- \int_{t_0}^{t} \frac{1}{n}\Phi^{i-1}\left(\mathcal{L}_i^n(s)\right)(C + (\mu_i(t-s), t-s))dM_{i,i-1}^n(s) \\
+ \int_{t_0}^{t} \frac{1}{n}\Phi^{i-1}\left(\mathcal{L}_{i-1}^n(s)\right)(C + (\mu_i(t-s), t-s))dM_{i-1,i}^n(s) \\
+ \frac{1}{n}\sum_{j=A_i^n(t_0)+1}^{A_i^n(t)} \delta_{(v_{i,j}^n, u_{i,j}^n)}(C + (\mu_i(t-\tau_{i,j}^n), t-\tau_{i,j}^n)) \\
+ \int_{t_0}^{t} \frac{1}{n}\Phi^{i}\left(\mathcal{L}_{i+1}^n(s)\right)(C + (\mu_i(t-s), t-s))dM_{i+1,i}^n(s) \\
- \int_{t_0}^{t} \frac{1}{n}\Phi^{i}\left(\mathcal{L}_i^n(s)\right)(C + (\mu_i(t-s), t-s))dM_{i,i+1}^n(s).
\end{aligned} \tag{EC.33}$$

Let C^ϵ denote the ϵ-enlargement of the set C, i.e., $C^\epsilon = \{(x',y') \in \mathbb{R}^2_+ | \max(|x'-x|, |y'-y|) < \epsilon,\ x,y \in C\}$. Choose $\delta < \epsilon/(1+\mu_{\max})$, then $C + (\mu_i(t-s), t-s) \subset C^\epsilon$ for all $t_0 \le s \le t \le t_0 + \delta$. So the above equation together with (23) and (24) implies that

$$\begin{aligned}\bar{\mathcal{L}}_i^n(t)(C) &\le \bar{\mathcal{L}}_i^n(t_0)(C^\epsilon) + i\bar{A}_i^n(t_0,t) + i\bar{S}_{i+1}^n(t_0,t)\\ &\le \bar{\mathcal{L}}_i^n(t_0)(C^\epsilon) + 3I\epsilon,\end{aligned}$$

where the last inequality follows from (EC.31) and (EC.32). Since C is arbitrary, let $C_0 = C + (\mu_i(t-t_0), t-t_0)$. We also have that $C \subset {C_0}^\epsilon$ when $t_0 \le s \le t \le t_0 + \delta$. Therefore (EC.33) also yields the other direction of bound estimate

$$\bar{\mathcal{L}}_i^n(t)({C_0}^\epsilon) \ge \bar{\mathcal{L}}_i^n(t_0)(C_0) - (i-1)\bar{S}_i^n(t_0,t) - i\bar{A}_{i+1}^n(t_0,t).$$

These with (EC.31) and (EC.32) give

$$\bar{\mathcal{L}}_i^n(t_0)(C_0) \le \bar{\mathcal{L}}_i^n(t)({C_0}^\epsilon) + 3I\epsilon.$$

By the definition of Prohorov metric between two finite measures, we have $\mathbf{d}[\bar{\mathcal{L}}_i^n(t), \bar{\mathcal{L}}_i^n(t_0)] \le 3I\epsilon$. This gives (EC.23) since ϵ is arbitrary. □

EC.5.2. Convergence to Fluid Model

In this section we prove Theorem 1. It follows from Lemmas EC.4 and EC.5 that the sequence of fluid scaled processes $\{(\bar{\mathcal{R}}^n, \bar{\mathcal{L}}^n, \bar{S}^n, \bar{A}^n)\}_{n\in\mathbb{N}}$ is tight. Since $R^n(t) = \mathcal{R}^n(t)(\mathbb{R})$, $Q^n(t) = \mathcal{R}^n(t)(\mathbb{R}_+)$ and $E^n(t) = \sum_{i=1}^I A_i^n(t)$, the sequence of fluid scaled processes $\{(\bar{R}^n, \bar{Q}^n, \bar{E}^n)\}_{n\in\mathbb{N}}$ is also tight. The tightness of $\bar{M}_{i,i-1}^n$ and $\bar{M}_{i-1,i}^n$ follows from the fact that $\bar{M}_{i,i-1}^n(t) - \bar{M}_{i,i-1}^n(s) \le \bar{S}_i^n(t) - \bar{S}_i^n(s)$ and $\bar{M}_{i-1,i}^n(t) - \bar{M}_{i-1,i}^n(s) \le \bar{A}_i^n(t) - \bar{A}_i^n(s)$ for any $0 \le s \le t$ by (23) and (24) and that these processes are non-decreasing. The tightness of $\bar{Z}^n$, $\bar{B}^n$ and $\bar{D}^n$ can be seen from (26), (27) and (30), respectively. The tightness of the external arrival process $\bar{\Lambda}^n$ is given by (52). So every subsequence of the fluid scaled processes $\{(\bar{\mathcal{R}}^n, \bar{\mathcal{L}}^n, \bar{R}^n, \bar{Q}^n, \bar{Z}^n, \bar{\Lambda}^n, \bar{B}^n, \bar{D}^n, \bar{E}^n, \bar{A}^n, \bar{S}^n, \bar{M}^n) : n \in \mathbb{N}\}$ has a further subsequence which converges to some limit, denoted by $(\bar{\mathcal{R}}, \bar{\mathcal{L}}, \bar{R}, \bar{Q}, \bar{Z}, \bar{\Lambda}, \bar{B}, \bar{D}, \bar{E}, \bar{A}, \bar{S}, \bar{M})$. For notational simplicity, we still use index n for the convergent subsequence. By Skorohod representation theorem (Lemma C.1 in Zhang (2013)) we can map all the random objects to the same probability space so that all weak convergence becomes almost sure convergence (see the discussion of §5.2 in Zhang (2013) for technical details). Again for simplicity we use the same notation for these processes in the new space. Therefore,

$$(\bar{\mathcal{R}}^n, \bar{\mathcal{L}}^n, \bar{R}^n, \bar{Q}^n, \bar{Z}^n, \bar{\Lambda}^n, \bar{B}^n, \bar{D}^n, \bar{E}^n, \bar{A}^n, \bar{S}^n, \bar{M}^n) \to (\bar{\mathcal{R}}, \bar{\mathcal{L}}, \bar{R}, \bar{Q}, \bar{Z}, \bar{\Lambda}, \bar{B}, \bar{D}, \bar{E}, \bar{A}, \bar{S}, \bar{M}) \quad \text{(EC.34)}$$

almost surely as $n \to \infty$.

In order to complete the proof of Theorem 1, we need to verify that every such limit satisfies the fluid dynamic equations (35)–(51).

Proof of Theorem 1. Note that the fluid dynamic equations (38), (41)–(44), (47), (48), the second equation of (49), and as well the non-idling constraint (51) can be verified easily by the corresponding stochastic equations and convergence of these processes as stated in (EC.34). By Lemma 2.4 in Dai and Williams (1996), the fluid equation (50) also holds. Combining Lemma EC.5 with Theorem 15.5 of Billingsley (1968) yields that $\bar{Q}(t)=\bar{\mathcal{R}}(t)(C_0)$, $\bar{S}_i(t)$ and $\bar{A}_i(t)$ are continuous. Thus we also have the continuity of $\bar{Z}_i(t)$ by (44). Therefore, if $\bar{Z}_I(t)<N$, we have for all n large enough $Z_I^n(s)<N^n$ for $|s-t|<\delta$, $\delta>0$. It implies that $Q^n(s)=0$ for $|s-t|<\delta$ by (20). This together with the first entries of (23) and (24) yields $M^n_{I,I-1}(t)-M^n_{I,I-1}(s)=S^n_I(t)-S^n_I(s)$ and $M^n_{I-1,I}(t)-M^n_{I-1,I}(s)=A^n_I(t)-A^n_I(s)$ for $|s-t|<\delta$. Thus (40) holds. On the other hand, if $\bar{Q}(t)>0$, then we have for all n large enough $Q^n(s)>0$ for $|s-t|<\delta$, $\delta>0$. This yields $M^n_{I,I-1}(t)-M^n_{I,I-1}(s)=M^n_{I-1,I}(t)-M^n_{I-1,I}(s)=0$. Consequently, $d\bar{M}_{I,I-1}(t)=d\bar{M}_{I-1,I}(t)=0$. This proves (39). Moreover, since the buffer of the CSC systems is same as that of the call center model studied in Zhang (2013), the fluid equations (46) and the first equation of (49) that relate to the buffer follow the same argument in Lemma 5.5 of Zhang (2013).

It remains to verify (35) and (45). Comparing the stochastic dynamic equation (22) and the fluid one (35), we first show that as $n\to\infty$

$$\frac{1}{n}\sum_{j=1}^{A_i^n(t)}\mathbb{1}_{\{v^n_{i,j}>x+\mu_i(t-\tau^n_{i,j}),u^n_{i,j}>y+t-\tau^n_{i,j}\}}\to\int_0^t G^c(x+\mu_i(t-s))F^c(x+t-s)d\bar{A}_i(s). \tag{EC.35}$$

Let $0=t_0<t_1<\ldots<t_K=t$ be a partition of the interval $[0,t]$ such that $\max_{1\le k\le K}|t_k-t_{k-1}|<\delta$. Using the partition, we divide the integration into K parts. Since $\tau^n_{i,j}\in[t_k,t_{k+1}]$ for those $j\in[A_i^n(t_k)+1,A_i^n(t_{k+1})]$, we have

$$\begin{aligned}
&\frac{1}{n}\sum_{j=A_i^n(t_k)+1}^{A_i^n(t_{k+1})}\mathbb{1}_{\{v^n_{i,j}>x+\mu_i(t-\tau^n_{i,j}),u^n_{i,j}>y+t-\tau^n_{i,j}\}}\\
\le&\frac{1}{n}\sum_{j=A_i^n(t_k)+1}^{A_i^n(t_{k+1})}\mathbb{1}_{\{v^n_{i,j}>x+\mu_i(t-t_{k+1}),u^n_{i,j}>y+t-t_{k+1}\}}\\
\le&\bar{A}_i^n(t_k,t_{k+1})G^c(x+\mu_i(t-t_{k+1}))F^c(y+t-t_{k+1})+\epsilon,
\end{aligned}$$

where the last inequality follows from the Glivenko-Cantelli estimate (EC.13). Since ϵ can be arbitrary, we have

$$\limsup_{n\to\infty}\frac{1}{n}\sum_{j=1}^{A_i^n(t)}\mathbb{1}_{\{v^n_{i,j}>x+\mu_i(t-\tau^n_{i,j}),u^n_{i,j}>y+t-\tau^n_{i,j}\}}\le\sum_{k=0}^{K-1}G^c(x+\mu_i(t-t_{k+1}))F^c(y+t-t_{k+1})\bar{A}_i(t_k,t_{k+1}) \tag{EC.36}$$

as $n\to\infty$. By a similar argument, we also have

$$\liminf_{n\to\infty}\frac{1}{n}\sum_{j=1}^{A_i^n(t)}\mathbb{1}_{\{v_{i,j}^n>x+\mu_i(t-\tau_{i,j}^n),u_{i,j}^n>y+t-\tau_{i,j}^n\}}\geq\sum_{k=0}^{K-1}G^c(x+\mu_i(t-t_k))F^c(y+t-t_k)\bar{A}_i(t_k,t_{k+1}). \tag{EC.37}$$

The terms (EC.36) and (EC.37) are Riemann-Stieltjes upper and lower sum of the integral on the right-hand of (EC.35). Since the partition is arbitrary, (EC.36) and (EC.37) give (EC.35).

Next we prove that the four other terms on the right-hand side of (22) also converge to their corresponding terms in (35). We start from the second term on the right-hand side of (22) and show that as $n\to\infty$

$$\int_0^t\frac{1}{n}\Phi^{i-1}\left(\mathcal{L}_i^n(s)\right)\left(C_{x+\mu_i(t-s)}\times C_{y+t-s}\right)dM_{i,i-1}^n(s)\to\int_0^t\frac{i-1}{i\bar{Z}_i(s)}\bar{\mathcal{L}}_i(s)(C_{x+\mu_i(t-s)}\times C_{y+t-s})d\bar{M}_{i,i-1}(s). \tag{EC.38}$$

Recall that we assume $\bar{Z}_i(\cdot)$, $i=1,\ldots,I$, has only finitely many switches between 0 and positive value in any bounded time interval. Then on the interval $[0,t]$ we have a finite partition $0=a_0<a_1<\ldots<a_L=t$ for some $L<\infty$, where a_l's are the switch points of $\bar{Z}_i(\cdot)$. Then (EC.38) is equivalent to

$$\int_{a_l}^{a_{l+1}}\frac{1}{n}\Phi^{i-1}\left(\mathcal{L}_i^n(s)\right)\left(C_{x+\mu_i(t-s)}\times C_{y+t-s}\right)dM_{i,i-1}^n(s)\to\int_{a_l}^{a_{l+1}}\frac{i-1}{i\bar{Z}_i(s)}\bar{\mathcal{L}}_i(s)(C_{x+\mu_i(t-s)}\times C_{y+t-s})d\bar{M}_{i,i-1}(s) \tag{EC.39}$$

as $n\to\infty$ for all $l=0,\ldots,L-1$. Based on the definition of a_l's, either $\bar{Z}_i(s)=0$ for all $s\in(a_l,a_{l+1})$ or $\bar{Z}_i(s)>0$ for all $s\in(a_l,a_{l+1})$. Thus we consider the following two cases:

Case 1: Assume that $\bar{Z}_i(s)=0$ for all $s\in(a_l,a_{l+1})$. By (36), the right-hand side of (EC.39) equals to zero. So we just need to prove that as $n\to\infty$

$$\int_{a_l}^{a_{l+1}}\frac{1}{n}\Phi^{i-1}\left(\mathcal{L}_i^n(s)\right)\left(C_{x+\mu_i(t-s)}\times C_{y+t-s}\right)dM_{i,i-1}^n(s)\to 0. \tag{EC.40}$$

This follows since

$$\begin{aligned}&\limsup_{n\to\infty}\int_{a_l}^{a_{l+1}}\frac{1}{n}\Phi^{i-1}\left(\mathcal{L}_i^n(s)\right)\left(C_{x+\mu_i(t-s)}\times C_{y+t-s}\right)dM_{i,i-1}^n(s)\\ \leq&\lim_{n\to\infty}(i-1)\bar{M}_{i,i-1}^n(a_l,a_{l+1})=(i-1)\bar{M}_{i,i-1}(a_l,a_{l+1}).\end{aligned}$$

Obviously, the left-hand side of (EC.40) is non-negative. By (23), $\bar{M}_{i,i-1}(a_l,a_{l+1})\leq\bar{S}_i(a_l,a_{l+1})$. By Lemma EC.7 we have $\dot{\bar{S}}_i(t)=0$ for $t\in(a_l,a_{l+1})$ because $\bar{Z}_i(t)=0$ for all $t\in(a_l,a_{l+1})$. This with Lemma EC.8 implies $\bar{S}_i(a_l,a_{l+1})=0$. Thus, (EC.39) holds.

Case 2: Assume that $\bar{Z}_i(s)>0$ for all $s\in(a_l,a_{l+1})$. We also need to verify (EC.39) in this case. Note that number of customers who moved from level i to level $i-1$ during the time interval

(a_l, a_{l+1}) is given by there are $M^n_{i,i-1}(a_l, a_{l+1})$. Let $\tau_j \in (a_l, a_{l+1})$, $j = 1, \ldots, M^n_{i,i-1}(a_l, a_{l+1})$, denote the time points such that $M^n_{i,i-1}(\tau_j) - M^n_{i,i-1}(\tau_j -) > 0$. Then the left-hand side of (EC.39) can be written as

$$\begin{aligned}
&\int_{a_l}^{a_{l+1}} \frac{1}{n} \Phi^{i-1}\left(\mathcal{L}^n_i(s)\right)\left(C_{x+\mu_i(t-s)} \times C_{y+t-s}\right) dM^n_{i,i-1}(s) \\
=& \frac{1}{n} \sum_{j=1}^{M^n_{i,i-1}(a_l, a_{l+1})} \Phi^{i-1}\left(\mathcal{L}^n_i(\tau_j)\right)\left(C_{x+\mu_i(t-\tau_j)} \times C_{y+t-\tau_j}\right). \qquad \text{(EC.41)}
\end{aligned}$$

The term $\Phi^{i-1}\left(\mathcal{L}^n_i(\tau_j)\right)\left(C_{x+\mu_i(t-\tau_j)} \times C_{y+t-\tau_j}\right)$ is a hypergeometric random variable, representing the number of elements in $\mathcal{L}^n_i(\tau_j)(C_{x+\mu_i(t-\tau_j)} \times C_{y+t-\tau_j})$ out of $i-1$ draws from a total of $\mathcal{L}^n_i(\tau_j)(C_0 \times C_0) = iZ^n_i(\tau_j)$ elements. And it has variance

$$(i-1)\frac{\mathcal{L}^n_i(\tau_j)(C_{x+\mu_i(t-\tau_j)} \times C_{y+t-\tau_j})}{iZ^n_i(\tau_j)}\left(1 - \frac{\mathcal{L}^n_i(\tau_j)(C_{x+\mu_i(t-\tau_j)} \times C_{y+t-\tau_j})}{iZ^n_i(\tau_j)}\right)\frac{iZ^n_i(\tau_j) - (i-1)}{iZ^n_i(\tau_j) - 1} \le i - 1.$$

Since each hypergeometric random variable $\Phi^{i-1}\left(\mathcal{L}^n_i(\tau_j)\right)\left(C_{x+\mu_i(t-\tau_j)} \times C_{y+t-\tau_j}\right)$ is sampled independently and has a finite second moment, we can conclude from Kolmogorov's strong law of large numbers (Theorem 2.3.10 in Sen and Singer (1994)) that for any $\epsilon > 0$ we have

$$\begin{aligned}
&\frac{1}{n} \sum_{j=1}^{M^n_{i,i-1}(a_l, a_{l+1})} \Phi^{i-1}\left(\mathcal{L}^n_i(\tau_j)\right)\left(C_{x+\mu_i(t-\tau_j)} \times C_{y+t-\tau_j}\right) \\
\le& \frac{1}{n} \sum_{j=1}^{M^n_{i,i-1}(a_l, a_{l+1})} \mathbb{E}\left[\Phi^{i-1}\left(\mathcal{L}^n_i(\tau_j)\right)\left(C_{x+\mu_i(t-\tau_j)} \times C_{y+t-\tau_j}\right)\right] + \epsilon \\
=& \frac{1}{n} \sum_{j=1}^{M^n_{i,i-1}(a_l, a_{l+1})} \frac{i-1}{iZ^n_i(\tau_j)} \mathcal{L}^n_i(\tau_j)(C_{x+\mu_i(t-\tau_j)} \times C_{y+t-\tau_j}) + \epsilon
\end{aligned}$$

for all large n. Note that $\bar{\mathcal{L}}^n_i(\tau_j)/\bar{Z}^n_i(\tau_j) = \mathbf{0}$ whenever $\bar{Z}^n_i(\tau_j) = 0$. Here the last equality holds since the expectation of the hypergeometric random variable $\Phi^{i-1}\left(\mathcal{L}^n_i(\tau_j)\right)\left(C_{x+\mu_i(t-\tau_j)} \times C_{y+t-\tau_j}\right)$ is $\frac{i-1}{iZ^n_i(\tau_j)}\mathcal{L}^n_i(\tau_j)(C_{x+\mu_i(t-\tau_j)} \times C_{y+t-\tau_j})$. Plugging the above into (EC.41) and considering a further partition $a_l = t_0 < t_1 \ldots < t_K = a_{l+1}$ of the interval $[a_l, a_{l+1}]$ yields

$$\begin{aligned}
&\int_{a_l}^{a_{l+1}} \frac{1}{n} \Phi^{i-1}\left(\mathcal{L}^n_i(s)\right)\left(C_{x+\mu_i(t-s)} \times C_{y+t-s}\right) dM^n_{i,i-1}(s) \\
\le& \sum_{k=0}^{K-1} \sup_{s \in [t_k, t_{k+1}]} \frac{i-1}{i\bar{Z}^n_i(s)} \bar{\mathcal{L}}^n_i(s)(C_{x+\mu_i(t-s)} \times C_{y+t-s}) \bar{M}^n_{i,i-1}(t_k, t_{k+1}) + \epsilon
\end{aligned} \qquad \text{(EC.42)}$$

for all large n. Since $\bar{Z}_i(s) > 0$ for all $s \in (a_l, a_{l+1})$, by continuous mapping theorem $\bar{\mathcal{L}}^n_i(s)/\bar{Z}^n_i(s)$ converges u.o.c. to $\bar{\mathcal{L}}_i(s)/\bar{Z}_i(s)$ a.s. as $n \to \infty$ on the interval (a_l, a_{l+1}). Therefore

$$\limsup_{n \to \infty} \int_{a_l}^{a_{l+1}} \frac{1}{n} \Phi^{i-1}\left(\mathcal{L}^n_i(s)\right)\left(C_{x+\mu_i(t-s)} \times C_{y+t-s}\right) dM^n_{i,i-1}(s)$$

$$\leq \sum_{k=0}^{K-1} \sup_{s\in[t_k,t_{k+1}]} \frac{i-1}{i\bar{Z}_i(s)}\bar{\mathcal{L}}_i(s)(C_{x+\mu_i(t-s)}\times C_{y+t-s})\bar{M}_{i,i-1}(t_k,t_{k+1}). \tag{EC.43}$$

Using a similar argument, we can show the inequality in the other direction

$$\begin{aligned}&\liminf_{n\to\infty}\int_{a_l}^{a_{l+1}} \frac{1}{n}\Phi^{i-1}\left(\mathcal{L}_i^n(s)\right)(C_{x+\mu_i(t-s)}\times C_{y+t-s})dM_{i,i-1}^n(s)\\ &\geq \sum_{k=0}^{K-1} \inf_{s\in[t_k,t_{k+1}]} \frac{i-1}{i\bar{Z}_i(s)}\bar{\mathcal{L}}_i(s)(C_{x+\mu_i(t-s)}\times C_{y+t-s})\bar{M}_{i,i-1}(t_k,t_{k+1}).\end{aligned} \tag{EC.44}$$

The terms (EC.43) and (EC.44) are Riemann-Stieltjes upper and lower sum of the integral on the right-hand of (EC.39). Because the partition of (a_l, a_{l+1}) is arbitrary, we have (EC.39). This completes the proof of (EC.38) by combining the results of Cases 1 and 2.

Next we focus on the fifth term on the right-hand side of (22), which is very similar to the second term whose convergence we studied above. Specifically we show that as $n\to\infty$

$$\begin{aligned}&\int_0^t \frac{1}{n}\Phi^i\left(\mathcal{L}_{i+1}^n(s)\right)(C_{x+\mu_i(t-s)}\times C_{y+t-s})dM_{i+1,i}^n(s)\\ &\to \int_0^t \frac{i}{(i+1)\bar{Z}_{i+1}(s)}\bar{\mathcal{L}}_{i+1}(s)(C_{x+\mu_i(t-s)}\times C_{y+t-s})d\bar{M}_{i+1,i}(s).\end{aligned} \tag{EC.45}$$

Note that (EC.45) is similar to (EC.38) and can be obtained by replacing the index i in (EC.38) by $i+1$ (wherever applicable) and $i-1$ by $i+1$, except for the service rate term μ_i. However the actual value of μ_i does not play a role in the proof of (EC.38) hence proof of (EC.45) is identical.

We next consider the third term and the last term on the right-hand side of (22) together and show that as $n\to\infty$

$$\begin{aligned}&\int_0^t \frac{1}{n}\Phi^{i-1}\left(\mathcal{L}_{i-1}^n(s)\right)(C_{x+\mu_i(t-s)}\times C_{y+t-s})dM_{i-1,i}^n(s) - \int_0^t \frac{1}{n}\Phi^i\left(\mathcal{L}_i^n(s)\right)(C_{x+\mu_i(t-s)}\times C_{y+t-s})dM_{i,i+1}^n(s)\\ &\to \int_0^t \frac{1}{\bar{Z}_{i-1}(s)}\bar{\mathcal{L}}_{i-1}(s)(C_{x+\mu_i(t-s)}\times C_{y+t-s})d\bar{M}_{i-1,i}(s) - \int_0^t \frac{1}{\bar{Z}_i(s)}\bar{\mathcal{L}}_i(s)(C_{x+\mu_i(t-s)}\times C_{y+t-s})d\bar{M}_{i,i+1}(s).\end{aligned} \tag{EC.46}$$

Because $\bar{Z}_i(\cdot)$, $i=1,\ldots,I$, has only finitely many switches between 0 and non-zero values in any bounded time interval, there exists a finite partition $0=b_0<b_1<\ldots<b_J=t$, $J<\infty$, of the interval $[0,t]$ such that on each open interval (b_j,b_{j+1}) the values of $\bar{Z}_i$ and $\bar{Z}_{i-1}$ only have the following four situations: 1) $\bar{Z}_i(s)>0$ and $\bar{Z}_{i-1}(s)>0$ for all $s\in(b_j,b_{j+1})$; 2) $\bar{Z}_i(s)=0$ and $\bar{Z}_{i-1}(s)>0$ for all $s\in(b_j,b_{j+1})$; 3) $\bar{Z}_i(s)>0$ and $\bar{Z}_{i-1}(s)=0$ for all $s\in(b_j,b_{j+1})$; and 4) $\bar{Z}_i(s)=0$ and $\bar{Z}_{i-1}(s)=0$ for all $s\in(b_j,b_{j+1})$. Then it is enough to prove that (EC.46) holds for each interval, that is,

$$\int_{b_j}^{b_{j+1}} \frac{1}{n}\Phi^{i-1}\left(\mathcal{L}_{i-1}^n(s)\right)(C_{x+\mu_i(t-s)}\times C_{y+t-s})dM_{i-1,i}^n(s) \tag{EC.47}$$

$$- \int_{b_j}^{b_{j+1}} \frac{1}{n}\Phi^i\left(\mathcal{L}_i^n(s)\right)(C_{x+\mu_i(t-s)}\times C_{y+t-s})dM_{i,i+1}^n(s) \tag{EC.48}$$

$$\to \int_{b_j}^{b_{j+1}} \frac{1}{\bar{Z}_{i-1}(s)} \bar{\mathcal{L}}_{i-1}(s)(C_{x+\mu_i(t-s)} \times C_{y+t-s}) d\bar{M}_{i-1,i}(s)$$
$$- \int_{b_j}^{b_{j+1}} \frac{1}{\bar{Z}_i(s)} \bar{\mathcal{L}}_i(s)(C_{x+\mu_i(t-s)} \times C_{y+t-s}) d\bar{M}_{i,i+1}(s)$$

as $n \to \infty$ for all $j = 0, \ldots, J-1$. According to the values of $\bar{Z}_i$ and $\bar{Z}_{i-1}$ we study the four cases:

Case i: Assume that $\bar{Z}_i(s) > 0$ and $\bar{Z}_{i-1}(s) > 0$ for all $s \in (b_j, b_{j+1})$. In this situation, (EC.47) follows from the argument we used to prove (EC.43) and (EC.44).

Case ii: Assume that $\bar{Z}_i(s) = 0$ and $\bar{Z}_{i-1}(s) > 0$ for all $s \in (b_j, b_{j+1})$. As in Case 1, (EC.47) converges since $\bar{Z}_{i-1}(s) > 0$. So we just need to consider the limit of (EC.48) and show that

$$\int_{b_j}^{b_{j+1}} \frac{1}{n} \Phi^i\left(\mathcal{L}_i^n(s)\right)(C_{x+\mu_i(t-s)} \times C_{y+t-s}) dM_{i,i+1}^n(s)$$
$$\to \int_{b_j}^{b_{j+1}} \frac{1}{\bar{Z}_i(s)} \bar{\mathcal{L}}_i(s)(C_{x+\mu_i(t-s)} \times C_{y+t-s}) d\bar{M}_{i,i+1}(s)$$

as $n \to \infty$. The right-hand side of the above limit is defined through (37). Plugging (37) to the right-hand side of the above limit, it is equivalent to

$$\int_{b_j}^{b_{j+1}} \frac{1}{n} \Phi^i\left(\mathcal{L}_i^n(s)\right)(C_{x+\mu_i(t-s)} \times C_{y+t-s}) dM_{i,i+1}^n(s)$$
$$\to \int_{b_j}^{b_{j+1}} \frac{1}{\bar{Z}_{i-1}(s)} \bar{\mathcal{L}}_{i-1}(s)(C_{x+\mu_i(t-s)} \times C_{y+t-s}) d\bar{M}_{i-1,i}(s) + \int_{b_j}^{b_{j+1}} G^c(x+\mu_i(t-s)) F^c(y+t-s) d\bar{A}_i(s)$$
$$+ \int_{b_j}^{b_{j+1}} \frac{i}{(i+1)\bar{Z}_{i+1}(s)} \bar{\mathcal{L}}_{i+1}(s)(C_{x+\mu_i(t-s)} \times C_{y+t-s}) d\bar{M}_{i+1,i}(t) \tag{EC.49}$$

as $n \to \infty$. Replacing t_0 and t in (EC.8) by b_j and b_{j+1}, respectively, yields

$$\begin{aligned}
\bar{\mathcal{L}}_i^n(b_{j+1})(C_x \times C_y) = {} & \bar{\mathcal{L}}_i^n(b_j)(C_{x+\mu_i(b_{j+1}-t_0)} \times C_{y+(b_{j+1}-t_0)}) \\
& - \int_{b_j}^{b_{j+1}} \frac{1}{n} \Phi^{i-1}\left(\mathcal{L}_i^n(s)\right)(C_{x+\mu_i(b_{j+1}-s)} \times C_{y+b_{j+1}-s}) dM_{i,i-1}^n(s) \\
& + \int_{b_j}^{b_{j+1}} \frac{1}{n} \Phi^{i-1}\left(\mathcal{L}_{i-1}^n(s)\right)(C_{x+\mu_i(b_{j+1}-s)} \times C_{y+b_{j+1}-s}) dM_{i-1,i}^n(s) \\
& + \frac{1}{n} \sum_{j=A_i^n(b_j)+1}^{A_i^n(b_{j+1})} \mathbb{1}_{\{v_{i,j}^n > x+\mu_i(b_{j+1}-\tau_{i,j}^n), u_{i,j}^n > y+b_{j+1}-\tau_{i,j}^n\}} \\
& + \int_{b_j}^{b_{j+1}} \frac{1}{n} \Phi^i\left(\mathcal{L}_{i+1}^n(s)\right)(C_{x+\mu_i(b_{j+1}-s)} \times C_{y+a_{j+1}-s}) dM_{i+1,i}^n(s) \\
& - \int_{b_j}^{b_{j+1}} \frac{1}{n} \Phi^i\left(\mathcal{L}_i^n(s)\right)(C_{x+\mu_i(a_{j+1}-s)} \times C_{y+b_{j+1}-s}) dM_{i,i+1}^n(s),
\end{aligned}$$

where the first two terms converge to 0 since $\bar{Z}_i(b_j) = \bar{Z}_i(b_{j+1}) = 0$ due to the continuity of the fluid limit, the second term on the right-hand side converges to 0 similar to (EC.40), the third

term on the right-hand side converges to the first term on the right-hand side of (EC.49) (proof is similar to that of (EC.43) and (EC.44)), the fourth term on the right-hand side converges to the second term on the right-hand side of (EC.49) by (EC.35), and the fifth term on the right-hand side converges to the third term on the right-hand side of (EC.49) by (EC.45). It then follows that as $n \to \infty$

$$\begin{aligned}
&\int_{b_j}^{b_{j+1}} \frac{1}{n}\Phi^i\left(\mathcal{L}_i^n(s)\right)\left(C_{x+\mu_i(b_{j+1}-s)} \times C_{y+b_{j+1}-s}\right) dM_{i,i+1}^n(s) \\
\to &\int_{b_j}^{b_{j+1}} \frac{1}{\bar{Z}_{i-1}(s)}\bar{\mathcal{L}}_{i-1}(s)(C_{x+\mu_i(b_{j+1}-s)} \times C_{y+b_{j+1}-s}) d\bar{M}_{i-1,i}(s) \\
&\quad + \int_{b_j}^{b_{j+1}} G^c(x+\mu_i(b_{j+1}-s))F^c(y+b_{j+1}-s) d\bar{A}_i(s) \\
&\quad + \int_{b_j}^{b_{j+1}} \frac{i}{(i+1)\bar{Z}_{i+1}(s)}\bar{\mathcal{L}}_{i+1}(s)(C_{x+\mu_i(b_{j+1}-s)} \times C_{y+b_{j+1}-s}) d\bar{M}_{i+1,i}(s).
\end{aligned}$$

Because x and y are arbitrary non-negative number this implies (EC.49).

Case iii: Assume that $\bar{Z}_i(s) > 0$ and $\bar{Z}_{i-1}(s) = 0$ for all $s \in (b_j, b_{j+1})$. In this situation, the limit of (EC.48) follows from the same argument we used to prove (EC.43) and (EC.44) since $\bar{Z}_i(s) > 0$. Therefore we just need to show that as $n \to \infty$

$$\begin{aligned}
&\int_{b_j}^{b_{j+1}} \frac{1}{n}\Phi^{i-1}\left(\mathcal{L}_{i-1}^n(s)\right)\left(C_{x+\mu_i(t-s)} \times C_{y+t-s}\right) dM_{i-1,i}^n(s) \\
\to &\int_{b_j}^{b_{j+1}} \frac{1}{\bar{Z}_{i-1}(s)}\bar{\mathcal{L}}_{i-1}(s)(C_{x+\mu_i(t-s)} \times C_{y+t-s}) d\bar{M}_{i-1,i}(s).
\end{aligned} \tag{EC.50}$$

By (37) (just by changing indices)

$$\begin{aligned}
\frac{1}{\bar{Z}_{i-1}(t)}\bar{\mathcal{L}}_{i-1}(t)(C_x \times C_y) d\bar{M}_{i-1,i}(t) = &\frac{1}{\bar{Z}_{i-2}(t)}\bar{\mathcal{L}}_{i-2}(t)(C_x \times C_y) d\bar{M}_{i-2,i-1}(t) + G^c(x)F^c(y) d\bar{A}_{i-1}(t) \\
&+ \frac{i-1}{i\bar{Z}_i(t)}\bar{\mathcal{L}}_i(t)(C_x \times C_y) d\bar{M}_{i,i-1}(t).
\end{aligned}$$

Therefore, to prove (EC.50), it is enough to show that

$$\begin{aligned}
&\int_{b_j}^{b_{j+1}} \frac{1}{n}\Phi^{i-1}\left(\mathcal{L}_{i-1}^n(s)\right)\left(C_{x+\mu_i(t-s)} \times C_{y+t-s}\right) dM_{i-1,i}^n(s) \\
\to &\int_{b_j}^{b_{j+1}} \frac{1}{\bar{Z}_{i-2}(s)}\bar{\mathcal{L}}_{i-2}(s)(C_{x+\mu_i(t-s)} \times C_{y+t-s}) d\bar{M}_{i-2,i-1}(t) + \int_{b_j}^{b_{j+1}} G^c(x+\mu_i(t-s))F^c(y+t-s) d\bar{A}_{i-1}(s) \\
&\quad + \int_{b_j}^{b_{j+1}} \frac{i-1}{i\bar{Z}_i(t)}\bar{\mathcal{L}}_i(s)(C_{x+\mu_i(t-s)} \times C_{y+t-s}) d\bar{M}_{i,i-1}(s)
\end{aligned} \tag{EC.51}$$

as $n \to \infty$. As we argued for the relation between (EC.38) and (EC.45), the proof of (EC.51) follows from (EC.49).

Case iv: $\bar{Z}_i(s)=0$ and $\bar{Z}_{i-1}(s)=0$ for all $s\in(b_j,b_{j+1})$. In this situation, we need to consider the limit of (EC.47) and (EC.48) simultaneously. Their fluid limit is also defined through (37), which implies

$$\begin{aligned}&\frac{1}{\bar{Z}_{i-1}(t)}\bar{\mathcal{L}}_{i-1}(t)(C_x\times C_y)d\bar{M}_{i-1,i}(t)-\frac{1}{\bar{Z}_i(t)}\bar{\mathcal{L}}_i(t)(C_x\times C_y)d\bar{M}_{i,i+1}(t)\\=&-G^c(x)F^c(y)d\bar{A}_i(t)-\frac{i}{(i+1)\bar{Z}_{i+1}(t)}\bar{\mathcal{L}}_{i+1}(t)d\bar{M}_{i+1,i}(t).\end{aligned}$$

Plugging the above equation to the limit of (EC.47) and (EC.48), it suffices to prove that as $n\to\infty$

$$\begin{aligned}&\int_{b_j}^{b_{j+1}}\frac{1}{n}\Phi^{i-1}\left(\mathcal{L}^n_{i-1}(s)\right)(C_{x+\mu_i(t-s)}\times C_{y+t-s})dM^n_{i-1,i}(s)\\&\qquad-\int_{b_j}^{b_{j+1}}\frac{1}{n}\Phi^{i}\left(\mathcal{L}^n_{i}(s)\right)(C_{x+\mu_i(t-s)}\times C_{y+t-s})dM^n_{i,i+1}(s)\\\to&-\int_{b_j}^{b_{j+1}}G^c(x+\mu_i(t-s))F^c(y+t-s)d\bar{A}_i(s)\\&\qquad-\int_{b_j}^{b_{j+1}}\frac{i}{(i+1)\bar{Z}_{i+1}(s)}\bar{\mathcal{L}}_{i+1}(s)(C_{x+\mu_i(t-s)}\times C_{y+t-s})d\bar{M}_{i+1,i}(s).\end{aligned}$$

But we can still use the same argument we used to prove (EC.49), the only difference here is that we combine the limit of (EC.47) and (EC.48) together.

Combining these results with (EC.35), (EC.38) and (EC.45), we can conclude that (35) corresponds to a fluid limit of (22). The proof of (45) follows the same argument as that of (35). Thus we omit it for brevity. □

EC.5.2.1. Auxiliary Results

Lemma EC.7. *For any fluid limit, $\bar{S}_i$ is differentiable almost everywhere and the derivative $\dot{\bar{S}}_i(t):=(d/dt)\bar{S}_i(t)$ satisfies*

$$\dot{\bar{S}}_i(t)=\lim_{\delta\to0}\frac{\bar{\mathcal{L}}_i(t)(\mathcal{A}(\mu_i\delta,\delta))}{\delta},\tag{EC.52}$$

a.e. for $i=1,\dots,I$.

Proof. For any $i=1,\dots,I$, $\bar{S}_i$ is the cumulative amount of departure from level i, so it is nondecreasing. Thus, $\bar{S}_i$ is differentiable almost everywhere (see Royden (1988), Page 100). We prove (EC.52) using (EC.9). Following the same argument we used to prove (EC.35) we have the following limit for the fourth term on the right hand-side of (EC.9),

$$\lim_{n\to\infty}\frac{1}{n}\sum_{j=A^n_i(t_0)+1}^{A^n_i(t)}\mathbb{1}_{\{v^n_{i,j}\le\mu_i(t-\tau^n_{i,j})\text{ or }u^n_{i,j}\le t-\tau^n_{i,j}\}}=\int_{t_0}^{t}\left[1-G^c(\mu_i(t-s))F^c(t-s)\right]d\bar{A}_i(s).\tag{EC.53}$$

We use $\bar{Y}^n_{t_0}(t)$ to denote the other four terms on the right-hand side of (EC.9),

$$\begin{aligned}\bar{Y}^n_{t_0}(t) := &-\int_{t_0}^t \frac{1}{n}\Phi^{i-1}\left(\mathcal{L}^n_i(s)\right)\left(\mathcal{A}(\mu_i(t-s),(t-s))\right)dM^n_{i,i-1}(s)\\ &+\int_{t_0}^t \frac{1}{n}\Phi^{i-1}\left(\mathcal{L}^n_{i-1}(s)\right)\left(\mathcal{A}(\mu_i(t-s),(t-s))\right)dM^n_{i-1,i}(s)\\ &+\int_{t_0}^t \frac{1}{n}\Phi^{i}\left(\mathcal{L}^n_{i+1}(s)\right)\left(\mathcal{A}(\mu_i(t-s),(t-s))\right)dM^n_{i+1,i}(s)\\ &-\int_{t_0}^t \frac{1}{n}\Phi^{i}\left(\mathcal{L}^n_{i}(s)\right)\left(\mathcal{A}(\mu_i(t-s),(t-s))\right)dM^n_{i,i+1}(s),\end{aligned} \tag{EC.54}$$

where we append a subscript t_0 to $\bar{Y}^n_{t_0}(t)$ to emphasize the dependence on t_0. It can be seen from (EC.9), (EC.34) and (EC.53) that the limit of $\bar{Y}^n_{t_0}(t)$ exists. Let $\bar{Y}_{t_0}(t) = \lim_{n\to\infty} \bar{Y}^n_{t_0}(t)$. Then by (EC.9) and (EC.34) $\bar{S}_i$ satisfies

$$\bar{S}_i(t) = \bar{S}_i(t_0) + \bar{\mathcal{L}}_i(t_0)(\mathcal{A}_i(\mu_i(t-t_0),t-t_0)) + \int_{t_0}^t \left[1-G^c(\mu_i(t-s))F^c(t-s)\right]d\bar{A}_i(s) + \bar{Y}_{t_0}(t). \tag{EC.55}$$

Taking derivative of the above equation at t_0 yields

$$\dot{\bar{S}}_i(t_0) = \frac{d}{dt}\bar{\mathcal{L}}_i(t_0)(\mathcal{A}_i(\mu_i(t-t_0),t-t_0))\Big|_{t=t_0} + \frac{d}{dt}\bar{Y}_{t_0}(t)\Big|_{t=t_0}. \tag{EC.56}$$

Note that the first term on the right-hand side of (EC.56) is identical to the right-hand side of (EC.52). Thus it suffices to prove that $\frac{d}{dt}\bar{Y}_{t_0}(t)|_{t=t_0} = 0$.

By (EC.54)

$$\begin{aligned}\bar{Y}^n_{t_0}(t) \geq &-\int_{t_0}^t \frac{1}{n}\Phi^{i-1}\left(\mathcal{L}^n_i(s)\right)\left(\mathcal{A}(\mu_i(t-s),(t-s))\right)dM^n_{i,i-1}(s)\\ &-\int_{t_0}^t \frac{1}{n}\Phi^{i}\left(\mathcal{L}^n_{i}(s)\right)\left(\mathcal{A}(\mu_i(t-s),(t-s))\right)dM^n_{i,i+1}(s).\end{aligned} \tag{EC.57}$$

Now we consider the first term on the right-hand side of the above inequality. The number of customers that switches from level i to level $i-1$ during the time interval $(t_0,t]$ is given by $(i-1)M^n_{i,i-1}(t_0,t) := (i-1)[M^n_{i,i-1}(t) - M^n_{i,i-1}(t_0)]$. We index these customers by $l = 1,\ldots,(i-1)M^n_{i,i-1}(t_0,t)$ according to their switch time and we use τ^n_l to denote the switch time of lth customer. Let s^n_l be the amount of service that the lth customer has already received by time τ^n_l and w^n_l the time that the lth customer has already spent during service by time τ^n_l. Also let v^n_l and u^n_l denote the remaining service time and remaining patience during service of this customer, respectively. Then we have

$$\int_{t_0}^t \frac{1}{n}\Phi^{i-1}\left(\mathcal{L}^n_i(s)\right)\left(\mathcal{A}(\mu_i(t-s),(t-s))\right)dM^n_{i,i-1}(s) = \frac{1}{n}\sum_{l=1}^{(i-1)M^n_{i,i-1}(t_0,t)} \delta_{(v^n_l,u^n_l)}(\mathcal{A}(\mu_i(t-\tau^n_l),(t-\tau^n_l))). \tag{EC.58}$$

Note that $\delta_{(v_l^n,u_l^n)}(\mathcal{A}(\mu_i(t-s),(t-s)))$ has a binomial distribution with mean

$$\mathbb{E}[\delta_{(v_l^n,u_l^n)}(\mathcal{A}(\mu_i(t-s),(t-s)))] = 1 - \frac{G^c(s_l^n+\mu_i(t-\tau_l^n))}{G^c(s_l^n)}\frac{F^c(w_l^n+t-\tau_l^n)}{F^c(w_l^n)},$$

where the equality follows from the fact that the remaining service time v_l^n follows distribution function $1-\frac{G^c(\tau_l^n+x)}{G^c(\tau_l^n)}$ and the remaining patience time during service u_l^n follows distribution function $1-\frac{F^c(w_l^n+y)}{F^c(w_l^n)}$. It then follows follows from (EC.58) and Kolmogorov's strong law of large numbers that for any $\epsilon>0$,

$$\begin{aligned}
&\int_{t_0}^t \frac{1}{n}\Phi^{i-1}\left(\mathcal{L}_i^n(s)\right)\left(\mathcal{A}(\mu_i(t-s),(t-s))\right)dM_{i,i-1}^n(s)\\
&\le \frac{1}{n}\sum_{l=1}^{(i-1)M_{i,i-1}^n(t_0,t)}\left(1-\frac{G^c(s_l^n+\mu_i(t-\tau_l^n))}{G^c(s_l^n)}\frac{F^c(w_l^n+t-\tau_l^n)}{F^c(w_l^n)}\right)+\epsilon
\end{aligned} \tag{EC.59}$$

for all large n. By our assumption the time spent during service w_l^n is bounded. Thus, s_l^n is also bounded since $s_l^n \le \mu_{\max} w_l^n$. So there exists $M_0>0$ such that $w_l^n, s_l^n \le M_0$ for all l and n.

Let $0=a_0<a_1<\ldots<a_J=M_0$ and $0=b_0<b_1<\ldots<b_K=M_0$ be two partitions of the interval $[0,M_0]$. Among all the $(i-1)M_{i,i-1}^n(t_0,t)$ randomly selected customers from level i to level $i-1$ on the time interval $(t_0,t]$, let $M_{i,i-1}^{n,jk}(t_0,t)$ denote the number of customers whose total amount of service and time spent in service satisfy $(s_l^n,w_l^n)\in(a_j,a_{j+1}]\times(b_k,b_{k+1}]$. We have, $(i-1)M_{i,i-1}^n(t_0,t) = \sum_{j=1}^J\sum_{k=1}^K M_{i,i-1}^{n,jk}(t_0,t)$. And the limit of the fluid scaled process of $M_{i,i-1}^{n,jk}(t_0,t)$ also exists and let $\bar{M}_{i,i-1}^{jk}(t_0,t)$ denote its limit. Also

$$(i-1)\bar{M}_{i,i-1}(t_0,t) = \sum_{j=0}^{J-1}\sum_{k=0}^{K-1}\bar{M}_{i,i-1}^{jk}(t_0,t). \tag{EC.60}$$

We can also see from the above discussion that the right-hand side of (EC.59) satisfies

$$\begin{aligned}
&\frac{1}{n}\sum_{l=1}^{(i-1)M_{i,i-1}^n(t_0,t)}\left(1-\frac{G^c(s_l^n+\mu_i(t-\tau_l^n))}{G^c(s_l^n)}\frac{F^c(w_l^n+t-\tau_l^n)}{F^c(w_l^n)}\right)\\
&=\frac{1}{n}\sum_{j=0}^{J-1}\sum_{k=0}^{K-1}\sum_{l=1}^{M_{i,i-1}^{n,jk}(t_0,t)}\left(1-\frac{G^c(s_l^n+\mu_i(t-\tau_l^n))}{G^c(s_l^n)}\frac{F^c(w_l^n+t-\tau_l^n)}{F^c(w_l^n)}\right)\mathbb{1}_{\{(s_l^n,w_l^n)\in(a_j,a_{j+1}]\times(b_k,b_{k+1}]\}}\\
&\le\frac{1}{n}\sum_{j=0}^{J-1}\sum_{k=0}^{K-1}\sum_{l=1}^{M_{i,i-1}^{n,jk}(t_0,t)}\left(1-\frac{G^c(a_{j+1}+\mu_i(t-\tau_l^n))}{G^c(a_j)}\frac{F^c(b_{k+1}+t-\tau_l^n)}{F^c(b_k)}\right).
\end{aligned} \tag{EC.61}$$

Again by the same argument we used to prove (EC.35), we have

$$\lim_{n\to\infty}\frac{1}{n}\sum_{l=1}^{M_{i,i-1}^{n,jk}(t_0,t)}\left(1-\frac{G^c(a_{j+1}+\mu_i(t-\tau_{i,j}^n))F^c(b_{k+1}+t-\tau_{i,j}^n)}{G^c(a_j)F^c(b_k)}\right)$$

$$= \int_{t_0}^{t} \left(1 - \frac{G^c(a_{j+1} + \mu_i(t-s))F^c(b_{k+1}+t-s)}{G^c(a_j)F^c(b_k)}\right) d\bar{M}^{jk}_{i,i-1}(s).$$

Combing the above limit with (EC.59) and (EC.61) yields

$$\begin{aligned} &\int_{t_0}^{t} \frac{1}{n}\Phi^{i-1}\left(\mathcal{L}^n_i(s)\right)\left(\mathcal{A}(\mu_i(t-s),(t-s))\right) dM^n_{i,i-1}(s) \\ &\leq \sum_{j=0}^{J-1}\sum_{k=0}^{K-1} \int_{t_0}^{t} \left(1 - \frac{G^c(a_{j+1} + \mu_i(t-s))F^c(b_{k+1}+t-s)}{G^c(a_j)F^c(b_k)}\right) d\bar{M}^{jk}_{i,i-1}(s) + 2\epsilon \end{aligned} \tag{EC.62}$$

for all large n.

Now we consider the second term on the right-hand side of (EC.57). Using the same partitions $\{a_j\}$ and $\{b_k\}$, we use $M^{n,jk}_{i,i+1}(t_0,t)$ to denote the number of customers with $(s^n_l, w^n_l) \in (a_j, a_{j+1}] \times (b_k, b_{k+1}]$. Then $iM^n_{i,i+1}(t_0,t) = \sum_{j=0}^{J-1}\sum_{k=0}^{K-1} M^{n,jk}_{i,i+1}(t_0,t)$. Also it's limit $\bar{M}^{jk}_{i,i+1}(t_0,t)$ as $n \to \infty$ exists. Then

$$i\bar{M}_{i,i+1}(t_0,t) = \sum_{j=0}^{J-1}\sum_{k=0}^{K-1} \bar{M}^{jk}_{i,i+1}(t_0,t). \tag{EC.63}$$

Using the same argument we used to prove (EC.62), we also have

$$\begin{aligned} &\int_{t_0}^{t} \frac{1}{n}\Phi^{i}\left(\mathcal{L}^n_i(s)\right)\left(\mathcal{A}(\mu_i(t-s),(t-s))\right) dM^n_{i,i+1}(s) \\ &\leq \sum_{j=0}^{J-1}\sum_{k=0}^{K-1} \int_{t_0}^{t} \left(1 - \frac{G^c(a_{j+1} + \mu_i(t-s))F^c(b_{k+1}+t-s)}{G^c(a_j)F^c(b_k)}\right) d\bar{M}^{jk}_{i,i+1}(s) + 2\epsilon \end{aligned} \tag{EC.64}$$

for all large n. Since the ϵ is arbitrary, we have by (EC.57), (EC.62) and (EC.64) that

$$\begin{aligned} \bar{Y}_{t_0}(t) \geq &-\sum_{j=0}^{J-1}\sum_{k=0}^{K-1} \int_{t_0}^{t} \left(1 - \frac{G^c(a_{j+1} + \mu_i(t-s))F^c(b_{k+1}+t-s)}{G^c(a_j)F^c(b_k)}\right) d\bar{M}^{jk}_{i,i-1}(s) \\ &-\sum_{j=0}^{J-1}\sum_{k=0}^{K-1} \int_{t_0}^{t} \left(1 - \frac{G^c(a_{j+1} + \mu_i(t-s))F^c(b_{k+1}+t-s)}{G^c(a_j)F^c(b_k)}\right) d\bar{M}^{jk}_{i,i+1}(s). \end{aligned} \tag{EC.65}$$

Taking derivative at t_0 yields

$$\frac{d}{dt}\bar{Y}_{t_0}(t)\Big|_{t=t_0} \geq -\sum_{j=0}^{J-1}\sum_{k=0}^{K-1} \left(1 - \frac{G^c(a_{j+1})F^c(b_{k+1})}{G^c(a_j)F^c(b_k)}\right) \dot{\bar{M}}^{jk}_{i,i-1}(t_0) - \sum_{j=0}^{J-1}\sum_{k=0}^{K-1} \left(1 - \frac{G^c(a_{j+1})F^c(b_{k+1})}{G^c(a_j)F^c(b_k)}\right) \dot{\bar{M}}^{jk}_{i,i+1}(t_0).$$

Also because the partitions $\{a_j\}$ and $\{b_k\}$ are arbitrary, we can assume without loss of generality that $1 - \frac{G^c(a_{j+1})F^c(b_{k+1})}{G^c(a_j)F^c(b_k)} \leq \delta$ for given $\delta > 0$. Then the above inequality implies

$$\frac{d}{dt}\bar{Y}_{t_0}(t)\Big|_{t=t_0} \geq -\delta \sum_{j=0}^{J-1}\sum_{k=0}^{K-1} \left(\dot{\bar{M}}^{jk}_{i,i-1}(t_0) + \dot{\bar{M}}^{jk}_{i,i+1}(t_0)\right) = -\delta\left((i-1)\dot{\bar{M}}_{i,i-1}(t_0) + i\dot{\bar{M}}_{i,i+1}(t_0)\right), \tag{EC.66}$$

where the last equality follows from (EC.60) and (EC.63).

It remains to consider the other direction. Similar to (EC.57), by (EC.54)

$$\bar{Y}^n_{t_0}(t) \le \int_{t_0}^t \frac{1}{n}\Phi^{i-1}\left(\mathcal{L}^n_{i-1}(s)\right)\left(\mathcal{A}(\mu_i(t-s),(t-s))\right)dM^n_{i-1,i}(s) + \int_{t_0}^t \frac{1}{n}\Phi^{i}\left(\mathcal{L}^n_{i+1}(s)\right)\left(\mathcal{A}(\mu_i(t-s),(t-s))\right)dM^n_{i+1,i}(s).$$

Then we can apply exactly the same argument to obtain

$$\frac{d}{dt}\bar{Y}_{t_0}(t)\Big|_{t=t_0} \le \delta\big((i-1)\dot{\bar{M}}_{i-1,i}(t_0) + i\dot{\bar{M}}_{i+1,i}(t_0)\big).$$

Combing the above inequality with (EC.66) immediately yields $\frac{d}{dt}\bar{Y}_{t_0}(t)|_{t=t_0} = 0$ because δ is arbitrary, proving the result. □

Lemma EC.8. *Consider the fluid limit in* (EC.34)*. If $\bar{Z}_i(t) = 0$ for all $t \in [t_0, t_1]$, for $t_0 < t_1$, then $\bar{S}_i(t)$ is absolutely continuous on $[t_0, t_1]$.*

Proof. We need to show that (see 7.17 in Rudin (1987)) for any $\epsilon > 0$, there exists a $\delta > 0$ such that for any m and any disjoint collection of segments $(\alpha_1, \beta_1), \ldots, (\alpha_m, \beta_m)$ in $[t_0, t_1]$ such that $\sum_m (\beta_m - \alpha_m) < \delta$, we have $\sum_m |\bar{S}_i(\beta_m) - \bar{S}_i(\alpha_m)| < \epsilon$.

We use the notation introduced in the proof of Lemma EC.7 throughout. Since $\bar{Z}_i(t_0) = 0$, we have $\bar{\mathcal{L}}_i(t_0) = \mathbf{0}$ from (18) and (EC.34). Then by (EC.55)

$$\bar{S}_i(t) = \bar{S}_i(t_0) + \int_{t_0}^t \left[1 - G^c(\mu_i(t-s))F^c(t-s)\right] d\bar{A}_i(s) + \bar{Y}_{t_0}(t) \quad \text{for all } t \in [t_0, t_1], \tag{EC.67}$$

where $\bar{Y}_{t_0}(t)$ is the fluid limit of (EC.54). Due to the reason that G and F are absolutely continuous, we have

$$\frac{d}{dt}[1 - G^c(\mu_i t)F^c(t)] = \mu_i g(\mu_i t)F^c(t) + G^c(\mu_i t)f(t), \tag{EC.68}$$

where g and f are the probability density functions of G and F, respectively. Hence by changing the order of integration, we get

$$\begin{aligned} &\int_{t_0}^t \left[1 - G^c(\mu_i(t-s))F^c(t-s)\right] d\bar{A}_i(s) \\ &= \int_{t_0}^t \int_{t_0}^x \left[\mu_i g(\mu_i(x-s))F^c(x-s) + G^c(\mu_i(x-s))f(x-s)\right] d\bar{A}_i(s)dx. \end{aligned} \tag{EC.69}$$

The above implies that (EC.69) is absolutely continuous.

Since $\bar{Z}_i(t) = 0$ for all $t \in [t_0, t_1]$, we can find that (EC.67) still holds after replacing t_0 and t by α_m and β_m, respectively. Thus,

$$\bar{S}_i(\beta_m) = \bar{S}_i(\alpha_m) + \int_{\alpha_m}^{\beta_m} \left[1 - G^c(\mu_i(\beta_m - s))F^c(\beta_m - s)\right] d\bar{A}_i(s) + \bar{Y}_{\alpha_m}(\beta_m).$$

The above implies

$$|\bar{S}_i(\beta_m) - \bar{S}_i(\alpha_m)| \le \int_{\alpha_m}^{\beta_m} [1 - G^c(\mu_i(\beta_m - s))F^c(\beta_m - s)]\, d\bar{A}_i(s) + |\bar{Y}_{\alpha_m}(\beta_m)|. \tag{EC.70}$$

We first study the first term on the right-hand side of the above inequality. By (EC.69),

$$\begin{aligned}
&\int_{\alpha_m}^{\beta_m} [1 - G^c(\mu_i(\beta_m - s))F^c(\beta_m - s)]\, d\bar{A}_i(s) \\
&= \int_{\alpha_m}^{\beta_m} \int_{\alpha_m}^{x} [\mu_i g(\mu_i(x-s))F^c(x-s) + G^c(\mu_i(x-s))f(x-s)] d\bar{A}_i(s)dx \\
&\le \int_{\alpha_m}^{\beta_m} \int_{t_0}^{x} [\mu_i g(\mu_i(x-s))F^c(x-s) + G^c(\mu_i(x-s))f(x-s)] d\bar{A}_i(s)dx,
\end{aligned} \tag{EC.71}$$

where the last inequality holds since $t_0 \le \alpha_m$. One can find that (EC.71) is just the difference of the absolutely continuous function (EC.69). So we can see from (EC.70) and (EC.71) that if $|\bar{Y}_{\alpha_m}(\beta_m)|$ can also be bounded by a difference of a certain absolutely continuous function then the absolutely continuity of $\bar{S}_i(\cdot)$ will immediately follow.

We use an argument similar to (EC.71) to analyze the last term in (EC.70). To simplify the notation, we use $\bar{X}^n_{1,t_0}(t)$ to denote the absolute value of the first term on the right-hand side of (EC.54), i.e.,

$$\bar{X}^n_{1,t_0}(t) := \int_{t_0}^{t} \frac{1}{n} \Phi^{i-1}\left(\mathcal{L}^n_i(s)\right) \left(\mathcal{A}(\mu_i(t-s), (t-s))\right) dM^n_{i,i-1}(s). \tag{EC.72}$$

Similarly, we use $\bar{X}^n_{2,t_0}(t)$, $\bar{X}^n_{3,t_0}(t)$ and $\bar{X}^n_{4,t_0}(t)$ to denote the absolute values of the last three terms on the right-hand side of (EC.54). Then by (EC.54), by replacing t_0 and t by α_m and β_m, respectively,

$$|\bar{Y}_{\alpha_m}(\beta_m)| = \lim_{n\to\infty} |\bar{Y}^n_{\alpha_m}(\beta_m)| \le \limsup_{n\to\infty} \sum_{l=1}^{4} \bar{X}^n_{l,\alpha_m}(\beta_m). \tag{EC.73}$$

The above inequality provides an upper bound to the $|\bar{Y}_{\alpha_m}(\beta_m)|$ in (EC.70). So we just need to study $\bar{X}^n_{l,\alpha_m}(\beta_m)$, $l = 1, \ldots, 4$, one by one. As the arguments are same, we mainly focus on $\bar{X}^n_{1,\alpha_m}(\beta_m)$.

In view of (EC.62), we denote

$$\bar{X}_{1,t_0}(t) = \sum_{j=0}^{J-1} \sum_{k=0}^{K-1} \int_{t_0}^{t} \left(1 - \frac{G^c(a_{j+1} + \mu_i(t-s))F^c(b_{k+1} + t - s)}{G^c(a_j)F^c(b_k)}\right) d\bar{M}^{jk}_{i,i-1}(s). \tag{EC.74}$$

Similar to (EC.68),

$$\begin{aligned}
&\frac{d}{dt}\left[1 - \frac{G^c(a_{j+1} + \mu_i t)F^c(b_{k+1} + t)}{G^c(a_j)F^c(b_k)}\right] \\
&= \frac{1}{G^c(a_j)F^c(b_k)} \left[\mu_i g(a_{j+1} + \mu_i t)F^c(b_{k+1} + t) + G^c(a_{j+1} + \mu_i t)f(b_{k+1} + t)\right].
\end{aligned}$$

Then similar to (EC.69), we have

$$\begin{aligned}\bar{X}_{1,t_0}(t) = &\sum_{j=0}^{J-1}\sum_{k=0}^{K-1}\int_{t_0}^{t}\left(1-\frac{G^c(a_{j+1})F^c(b_{k+1})}{G^c(a_j)F^c(b_k)}\right)d\bar{M}_{i,i-1}^{jk}(s) \\ &+\sum_{j=0}^{J-1}\sum_{k=0}^{K-1}\int_{t_0}^{t}\int_{t_0}^{x}\frac{1}{G^c(a_j)F^c(b_k)}\Big[\mu_i g(a_{j+1}+\mu_i(x-s))F^c(b_{k+1}+x-s) \\ &\qquad +G^c(a_{j+1}+\mu_i(x-s))f(b_{k+1}+x-s)\Big]d\bar{M}_{i,i-1}^{jk}(s)dx,\end{aligned} \tag{EC.75}$$

where the equality also follows from changing the order of integration. Clearly, the second term on the right-hand side of the above equation is absolutely continuous. With regards to (EC.72) and (EC.74), we have proven in (EC.62) that $\limsup_{n\to\infty}\bar{X}_{1,t_0}^n(t)\le\bar{X}_{1,t_0}(t)$. Replacing t_0 and t by α_m and β_m yields

$$\limsup_{n\to\infty}\bar{X}_{1,\alpha_m}^n(\beta_m)\le\bar{X}_{1,\alpha_m}(\beta_m). \tag{EC.76}$$

Moreover, it can be seen from (EC.75) that

$$\begin{aligned}\bar{X}_{1,\alpha_m}(\beta_m)\le &\sum_{j=0}^{J-1}\sum_{k=0}^{K-1}\int_{\alpha_m}^{\beta_m}\left(1-\frac{G^c(a_{j+1})F^c(b_{k+1})}{G^c(a_j)F^c(b_k)}\right)d\bar{M}_{i,i-1}^{jk}(s) \\ &+\sum_{j=0}^{J-1}\sum_{k=0}^{K-1}\int_{\alpha_m}^{\beta_m}\int_{t_0}^{x}\frac{1}{G^c(a_j)F^c(b_k)}\Big[\mu_i g(a_{j+1}+\mu_i(x-s))F^c(b_{k+1}+x-s) \\ &\qquad +G^c(a_{j+1}+\mu_i(x-s))f(b_{k+1}+x-s)\Big]d\bar{M}_{i,i-1}^{jk}(s)dx,\end{aligned} \tag{EC.77}$$

where the inequality follows due to the same reason as (EC.71), i.e., $t_0\le\alpha_m$. Same as (EC.71), the second term on the right-hand side of the above inequality is also the difference of an absolutely continuous function (the second term on the right-hand side of (EC.75)). Now we consider the first term on the right-hand side of (EC.77). Since G and F are absolutely continuous, we can choose $\{a_j\}$ and $\{b_k\}$ such that (see the definition above (EC.60))

$$1-\frac{G^c(a_{j+1})F^c(b_{k+1})}{G^c(a_j)F^c(b_k)}\le\frac{\epsilon}{8(i-1)\bar{M}_{i-1,i}(t_0,t_1)+1}$$

for all a_j's and b_k's. Considering the disjoint segments $(\alpha_1,\beta_1),\ldots,(\alpha_m,\beta_m)$ in $[t_0,t_1]$ yields

$$\begin{aligned}&\sum_{m}\sum_{j=0}^{J-1}\sum_{k=0}^{K-1}\int_{\alpha_m}^{\beta_m}\left(1-\frac{G^c(a_{j+1})F^c(b_{k+1})}{G^c(a_j)F^c(b_k)}\right)d\bar{M}_{i,i-1}^{jk}(s) \\ \le&\sum_{j=0}^{J-1}\sum_{k=0}^{K-1}\int_{t_0}^{t_1}\left(1-\frac{G^c(a_{j+1})F^c(b_{k+1})}{G^c(a_j)F^c(b_k)}\right)d\bar{M}_{i,i-1}^{jk}(s) \\ \le&\frac{\epsilon}{8(i-1)\bar{M}_{i,i-1}(t_0,t)+1}\sum_{j=0}^{J-1}\sum_{k=0}^{K-1}\bar{M}_{i,i-1}^{jk}(t_0,t) \\ <&\frac{\epsilon}{8},\end{aligned} \tag{EC.78}$$

where the last inequality follows from (EC.60).

In view of (EC.76) and (EC.77), we can obtain similar upper bounds for $\limsup_{n\to\infty} \bar{X}^n_{l,\alpha_m}(\beta_m)$, $l = 2, 3, 4$. Three similar inequalities like (EC.78) can also be obtained. By (EC.71), (EC.73), (EC.77) and (EC.78), we can conclude from (EC.70) that for any $\epsilon > 0$ there exists a $\delta > 0$ such that when $\sum_m (\beta_m - \alpha_m) < \delta$ there will be $\sum_m |\bar{S}_i(\beta_m) - \bar{S}_i(\alpha_m)| < \epsilon$. This proves the result. □

EC.6. Detailed Results of Simulation Experiments

In this section we provide the details of the results of the simulation experiments in §7 and compare them with our approximations. The results are presented in Tables EC.2–EC.7 for the experiments in §7.2, where our policy π reduces to the lightest-load-first policy. The results presented in Tables EC.8–EC.10 correspond to the experiments in §7.3. In this case, we present the both results under the policy π and the lightest-load-first policy as they are different in experiments with inefficient levels.

Each table includes the simulation results for a combination of service and patience times that are presented in Table 2. Each of these tables provides the results for the expected number of agents at each level in columns $\mathbb{E}Z_1$ through $\mathbb{E}Z_6$ for levels 1-6, the total abandonment rate under column Ab. Rate in our simulation experiments. And, we also provide the results for expected time in system: in the column $\mathbb{E}[W]$ for the expected time in system, in the column $\mathbb{E}[W|C]$ for customers whose service is completed, in the column $\mathbb{E}[W|A]$ for customers who abandoned the system, and in the last column stdev(W) for the standard deviation of time in system. We present our approximations for the associated quantities in the rows of "Approx", and we also present 95% confidence intervals found using the batch-means technique, whenever applicable.

System		$\mathbb{E}Z_1$	$\mathbb{E}Z_2$	$\mathbb{E}Z_3$	$\mathbb{E}Z_4$	$\mathbb{E}Z_5$	$\mathbb{E}Z_6$	Ab. Rate	$\mathbb{E}[W]$	$\mathbb{E}[W\|S]$	$\mathbb{E}[W\|A]$	stdev(W)
1_1	Sim.	0.885 ±0.010	11.039 ±0.058	12.368 ±0.060	0.691 ±0.032	0 ±0	0 ±0	62.780 ±0.173	0.2236 ±0.0002	0.2235 ±0.0004	0.2240 ±0.0005	0.2239 ±0.0004
	Approx.	0	12.5	12.5	0	0	0	62.5	0.2222	0.2220	0.2229	0.2229
2_1	Sim.	0.888 ±0.009	23.284 ±0.164	25.502 ±0.152	0.317 ±0.034	0 ±0	0 ±0	125.160 ±0.371	0.2228 ±0.0002	0.2227 ±0.0004	0.2232 ±0.0005	0.2230 ±0.0004
	Approx.	0	25	25	0	0	0	125	0.2222	0.2220	0.2229	0.2229
3_1	Sim.	0.843 ±0.014	48.311 ±0.383	50.779 ±0.389	0.063 ±0.016	0 ±0	0 ±0	249.871 ±0.664	0.2224 ±0.0002	0.2223 ±0.0004	0.2228 ±0.0005	0.2226 ±0.0004
	Approx.	0	50	50	0	0	0	250	0.2222	0.2220	0.2229	0.2229

Table EC.2 Results for combination I_1: Exponential service and patience times

System		$\mathbb{E}Z_1$	$\mathbb{E}Z_2$	$\mathbb{E}Z_3$	$\mathbb{E}Z_4$	$\mathbb{E}Z_5$	$\mathbb{E}Z_6$	Ab. Rate	$\mathbb{E}[W]$	$\mathbb{E}[W\|S]$	$\mathbb{E}[W\|A]$	stdev(W)
1_1	Sim.	0.333 ±0.005	6.320 ±0.047	15.462 ±0.049	2.862 ±0.061	0.018 ±0.003	0 ±0	70.814 ±0.204	0.2523 ±0.0001	0.2826 ±0.0002	0.1623 ±0.0005	0.1347 ±0.0001
	Approx.	0	4.390	20.610	0	0	0	70.610	0.2511	0.2811	0.1613	0.1340
2_1	Sim.	0.264 ±0.004	11.284 ±0.115	35.204 ±0.127	3.245 ±0.131	0 ±0	0 ±0	141.267 ±0.392	0.2516 ±0.0001	0.2819 ±0.0002	0.1617 ±0.0005	0.1341 ±0.0001
	Approx.	0	8.780	41.220	0	0	0	141.220	0.2511	0.2811	0.1613	0.1340
3_1	Sim.	0.215 ±0.005	20.217 ±0.281	76.432 ±0.199	3.135 ±0.177	0 ±0	0 ±0	282.181 ±0.732	0.2513 ±0.0001	0.2814 ±0.0001	0.1614 ±0.0005	0.1337 ±0.0001
	Approx.	0	17.560	82.440	0	0	0	282.440	0.2511	0.2811	0.1613	0.1340

Table EC.3 Results for combination II_1: Log-normal service and exponential patience times

System		$\mathbb{E}Z_1$	$\mathbb{E}Z_2$	$\mathbb{E}Z_3$	$\mathbb{E}Z_4$	$\mathbb{E}Z_5$	$\mathbb{E}Z_6$	Ab. Rate	$\mathbb{E}[W]$	$\mathbb{E}[W\|S]$	$\mathbb{E}[W\|A]$	stdev(W)
1_1	Sim.	0.321 ±0.006	6.123 ±0.064	15.521 ±0.069	3.009 ±0.058	0.021 ±0.003	0 ±0	43.010 ±0.112	0.2536 ±0.0002	0.2363 ±0.0003	0.3496 ±0.0007	0.2056 ±0.0005
	Approx.	0	3.962	21.038	0	0	0	42.731	0.2526	0.2352	0.3490	0.2049
2_1	Sim.	0.251 ±0.004	10.795 ±0.141	35.435 ±0.135	3.516 ±0.121	0 ±0	0 ±0	85.671 ±0.222	0.2530 ±0.0002	0.2358 ±0.0003	0.3491 ±0.0006	0.2052 ±0.0005
	Approx.	0	7.923	42.077	0	0	0	85.461	0.2526	0.2352	0.3490	0.2049
3_1	Sim.	0.201 ±0.006	19.084 ±0.317	77.132 ±0.248	3.582 ±0.189	0 ±0	0 ±0	171.005 ±0.464	0.2527 ±0.0002	0.2355 ±0.0003	0.3487 ±0.0006	0.2048 ±0.0005
	Approx.	0	15.846	84.154	0	0	0	170.922	0.2526	0.2352	0.3490	0.2049

Table EC.4 Results for combination III_1: Log-normal service and patience times

System		$\mathbb{E}Z_1$	$\mathbb{E}Z_2$	$\mathbb{E}Z_3$	$\mathbb{E}Z_4$	$\mathbb{E}Z_5$	$\mathbb{E}Z_6$	Ab. Rate	$\mathbb{E}[W]$	$\mathbb{E}[W\|S]$	$\mathbb{E}[W\|A]$	stdev(W)
1_2	Sim.	2.889 ±0.042	12.295 ±0.119	7.963 ±0.109	0.734 ±0.067	0.012 ±0.003	0 ±0	28.179 ±0.113	0.1504 ±0.0005	0.1502 ±0.0004	0.1536 ±0.0015	0.1535 ±0.0004
	Approx.	0	24.997	0	0	0	0	25.001	0.1333	0.1333	0.1333	0.1334
2_2	Sim.	4.478 ±0.078	31.986 ±0.345	13.233 ±0.351	0.226 ±0.059	0 ±0	0 ±0	54.523 ±0.241	0.1455 ±0.0005	0.1454 ±0.0005	0.1476 ±0.0016	01475 ±0.0005
	Approx.	0	49.995	0.005	0	0	0	50.003	0.1333	0.1333	0.1333	0.1334
3_2	Sim.	7.163 ±0.173	75.286 ±0.577	17.484 ±0.690	0.008 ±0.006	0 ±0	0 ±0	105.146 ±0.374	0.1403 ±0.0004	0.1402 ±0.0004	0.1415 ±0.0013	0.1413 ±0.0004
	Approx.	0	99.989	0.011	0	0	0	100.005	0.1333	0.1333	0.1333	0.1334

Table EC.5 Results for combination I_2: Exponential service and patience times

System		$\mathbb{E}Z_1$	$\mathbb{E}Z_2$	$\mathbb{E}Z_3$	$\mathbb{E}Z_4$	$\mathbb{E}Z_5$	$\mathbb{E}Z_6$	Ab. Rate	$\mathbb{E}[W]$	$\mathbb{E}[W\|S]$	$\mathbb{E}[W\|A]$	stdev(W)
1_2	Sim.	2.040 ±0.038	11.794 ±0.106	9.626 ±0.051	1.420 ±0.105	0.051 ±0.011	0 ±0	30.203 ±0.132	0.1613 ±0.0006	0.1666 ±0.0006	0.1014 ±0.0006	0.0815 ±0.0004
	Approx.	0	18.911	6.089	0	0	0	28.045	0.1496	0.1544	0.0934	0.0747
2_2	Sim.	2.752 ±0.069	27.981 ±0.312	18.521 ±0.272	0.701 ±0.123	0.002 ±0	0 ±0	58.502 ±0.263	0.1562 ±0.0005	0.1612 ±0.0005	0.0973 ±0.0006	0.0770 ±0.0003
	Approx.	0	37.822	12.178	0	0	0	56.089	0.1496	0.1544	0.0934	0.0747
3_2	Sim.	3.381 ±0.112	63.618 ±0.761	32.829 ±0.808	0.147 ±0.056	0 ±0	0 ±0	114.836 ±0.470	0.1532 ±0.0005	0.1581 ±0.0005	0.0950 ±0.0004	0.0746 ±0.0003
	Approx.	0	75.643	24.357	0	0	0	112.178	0.1496	0.1544	0.0934	0.0747

Table EC.6 Results for combination II_2: Log-normal service and exponential patience times

System		$\mathbb{E}Z_1$	$\mathbb{E}Z_2$	$\mathbb{E}Z_3$	$\mathbb{E}Z_4$	$\mathbb{E}Z_5$	$\mathbb{E}Z_6$	Ab. Rate	$\mathbb{E}[W]$	$\mathbb{E}[W\|S]$	$\mathbb{E}[W\|A]$	stdev(W)
1_2	Sim.	1.335 ±0.024	9.421 ±0.086	11.189 ±0.104	2.809 ±0.121	0.196 ±0.044	0.006 ±0.003	7.943 ±0.062	0.1761 ±0.0006	0.1708 ±0.0005	0.4210 ±0.0030	0.1673 ±0.0009
	Approx.	0	12.500	12.500	0	0	0	7.009	0.1667	0.1620	0.4115	0.1588
2_2	Sim.	1.522 ±0.033	21.266 ±0.250	24.749 ±0.233	2.420 ±0.274	0.020 ±0.012	0 ±0	14.762 ±0.110	0.1709 ±0.0005	0.1661 ±0.0005	0.4120 ±0.0028	0.1614 ±0.0008
	Approx.	0	25.001	24.999	0	0	0	14.018	0.1667	0.1620	0.4115	0.1588
3_2	Sim.	1.501 ±0.066	46.295 ±0.652	50.900 ±0.557	1.293 ±0.237	0 ±0	0 ±0	28.391 ±0.191	0.1681 ±0.0005	0.1635 ±0.0005	0.4074 ±0.0031	0.1582 ±0.0007
	Approx.	0	50.002	49.998	0	0	0	28.035	0.1667	0.1620	0.4115	0.1588

Table EC.7 Results for combination III_2: Log-normal service and patience times

System		Policy	$\mathbb{E}Z_1$	$\mathbb{E}Z_2$	$\mathbb{E}Z_3$	$\mathbb{E}Z_4$	$\mathbb{E}Z_5$	$\mathbb{E}Z_6$	Ab. Rate	$\mathbb{E}[W]$	$\mathbb{E}[W\|S]$	$\mathbb{E}[W\|A]$	stdev(W)
1_1	Sim.	lightest-load	0.525 ±0.007	7.484 ±0.063	14.740 ±0.060	2.234 ±0.058	0.007 ±0.002	0 ±0	68.602 ±0.205	0.2444 ±0.0003	0.2441 ±0.0004	0.2452 ±0.0004	0.2448 ±0.0002
		π	1.469 ±0.007	16.875 ±0.030	1.038 ±0.006	1.488 ±0.005	4.102 ±0.033	0 ±0	64.728 ±0.177	0.2306 ±0.0003	0.2300 ±0.0004	0.2325 ±0.0003	0.2325 ±0.0004
	Approx.		0	20.492	0	0	4.508	0	63.525	0.2259	0.2249	0.2290	0.2290
2_1	Sim.	lightest-load	0.425 ±0.008	13.710 ±0.184	33.368 ±0.161	2.493 ±0.103	0 ±0	0 ±0	137.786 ±0.434	0.2446 ±0.0005	0.2452 ±0.0005	0.2460 ±0.0004	0.2459 ±0.0005
		π	1.733 ±0.009	36.396 ±0.066	1.455 ±0.006	2.273 ±0.009	8.127 ±.059	0 ±0	128.485 ±0.377	0.2288 ±0.0002	0.2281 ±0.0004	0.2312 ±0.0005	0.2309 ±0.0005
	Approx.		0	40.983	0	0	9.017	0	127.050	0.2259	0.2249	0.2290	0.2290
3_1	Sim.	lightest-load	0.333 ±0.009	24.774 ±0.440	72.501 ±0.365	2.390 ±0.155	0 ±0	0 ±0	276.616 ±0.910	0.2464 ±0.0004	0.2462 ±0.0005	0.2469 ±0.0004	0.2467 ±0.0005
		π	1.921 ±0.016	76.233 ±0.128	2.067 ±0.008	3.447 ±0.019	16.322 ±0.128	0 ±0	255.762 ±0.819	0.2271 ±0.0002	0.2270 ±0.0003	0.2302 ±0.0003	0.2300 ±0.0003
	Approx.		0	81.967	0	0	18.033	0	254.099	0.2259	0.2249	0.2290	0.2290

Table EC.8 Results for combination I_1 with inefficient levels: Exponential service and patience times

System		Policy	$\mathbb{E}Z_1$	$\mathbb{E}Z_2$	$\mathbb{E}Z_3$	$\mathbb{E}Z_4$	$\mathbb{E}Z_5$	$\mathbb{E}Z_6$	Ab. Rate	$\mathbb{E}[W]$	$\mathbb{E}[W\|S]$	$\mathbb{E}[W\|A]$	stdev(W)
1_1	Sim.	lightest-load	0.133 ±0.004	2.982 ±0.040	14.753 ±0.054	6.991 ±0.058	0.139 ±0.013	0 ±0	78.943 ±0.221	0.2811 ±0.0001	0.3197 ±0.0002	0.1824 ±0.0004	0.1529 ±0.0001
		π	0.886 ±0.007	14.558 ±0.031	1.111 ±0.003	1.808 ±0.006	6.623 ±0.034	0 ±0	73.605 ±0.160	0.2622 ±0.0002	0.2943 ±0.0001	0.1715 ±0.0004	0.1452 ±0.0001
	Approx.		0	17.497	0	0	7.503	0	72.508	0.2578	0.2886	0.1692	0.1442
2_1	Sim.	lightest-load	0.068 ±0.002	3.584 ±0.071	33.537 ±0.122	12.790 ±0.159	0.020 ±0.007	0 ±0	158.973 ±0.418	0.2831 ±0.0001	0.3223 ±0.0001	0.1836 ±0.0003	0.1539 ±0.0002
		π	0.992 ±0.006	31.239 ±0.070	1.541 ±0.05	2.712 ±0.011	13.508 ±0.072	0 ±0	146.302 ±0.393	0.2606 ±0.0001	0.2923 ±0.0002	0.1706 ±0.0003	0.1447 ±0.0001
	Approx.		0	34.995	0	0	15.005	0	145.016	0.2578	0.2886	0.1692	0.1442
3_1	Sim.	lightest-load	0.032 ±0.001	3.794 ±0.124	72.591 ±0.302	23.583 ±0.342	0 ±0	0 ±0	319.506 ±0.863	0.2844 ±0.0001	0.3241 ±0.0002	0.1845 ±0.0004	0.1546 ±0.0001
		π	1.067 ±0.009	65.178 ±0.145	2.160 ±0.012	4.024 ±0.210	27.566 ±0.137	0 ±0	291.505 ±0.778	0.2596 ±0.0001	0.2909 ±0.0002	0.1701 ±0.1697	0.1443 ±0.0001
	Approx.		0	69.990	0	0	30.010	0	290.031	0.2578	0.2886	0.1692	0.1442

Table EC.9 Results for combination II_1 with inefficient levels: Log-normal service and exponential patience times

Overloaded systems: Next we provide the results of simulation experiments for overloaded systems. Our goal is to show that the distribution of service time also has significant impact on the steady-state behavior of the queue length of CSC systems. We assume that the patience time

System		Policy	$\mathbb{E}Z_1$	$\mathbb{E}Z_2$	$\mathbb{E}Z_3$	$\mathbb{E}Z_4$	$\mathbb{E}Z_5$	$\mathbb{E}Z_6$	Ab. Rate	$\mathbb{E}[W]$	$\mathbb{E}[W\|S]$	$\mathbb{E}[W\|A]$	stdev(W)
1_1	Sim.	lightest-load	0.140 ±0.004	3.060 ±0.051	14.850 ±0.057	6.819 ±0.072	0.128 ±0.011	0 ±0	50.625 ±0.122	0.2801 ±0.0002	0.2607 ±0.0003	0.3686 ±0.0006	0.2236 ±0.0005
		π	0.867 ±0.006	14.521 ±0.038	1.126 ±0.003	1.805 ±0.006	6.666 ±0.038	0 ±0	45.789 ±0.115	0.2627 ±0.0002	0.2441 ±0.0003	0.3583 ±0.0006	0.2136 ±0.0005
	Approx.		0	17.404	0	0	7.596	0	44.711	0.2588	0.2400	0.3577	0.2123
2_1	Sim.	lightest-load	0.074 ±0.002	3.758 ±0.080	33.821 ±0.157	12.330 ±0.191	0.017 ±0.005	0 ±0	102.258 ±0.238	0.2819 ±0.0003	0.2624 ±0.0003	0.3696 ±0.0003	0.2246 ±0.0005
		π	0.969 ±0.009	31.126 ±0.075	1.559 ±0.007	2.703 ±0.010	13.633 ±0.075	0 ±0	90.736 ±0.265	0.2613 ±0.0002	0.2427 ±0.0003	0.3579 ±0.0007	0.2130 ±0.0001
	Approx.		0	34.808	0	0	15.192	0	89.423	0.2588	0.2400	0.3577	0.2123
3_1	Sim.	lightest-load	0.036 ±0.002	4.058 ±0.157	73.426 ±0.320	22.479 ±0.385	0 ±0	0 ±0	205.994 ±0.496	0.2832 ±0.0002	0.2636 ±0.0003	0.3705 ±0.0006	0.2254 ±0.0006
		π	1.040 ±0.010	64.920 ±0.175	2.172 ±0.018	4.008 ±0.007	27.855 ±0.176	0 ±0	180.476 ±0.579	0.2604 ±0.0003	0.2418 ±0.0002	0.3575 ±0.0007	0.2126 ±0.0006
	Approx.		0	69.616	0	0	30.384	0	178.845	0.2588	0.2400	0.3577	0.2123

Table EC.10 Results for combination III_1 with inefficient levels: Log-normal service and patience times

distributions for waiting and during service are both log-normal with mean and variance equal to 1. We compare the simulation results with two service time distributions expo(1) and ln(1,1). And we use the service rate $\mu = \{4, 3.8, 3.3, 3, 2.75, 2.5\}$ same as the first experiment set in §7.1. The parameters of each experiment along with the simulation results (with 95% confidence intervals in parentheses) are presented in Table EC.11. In these experiments we consider two different pairs of values of λ and N such that the systems are overloaded. Our approximations for the probability of abandonment (see Table EC.11(a)) and the queue length (see Table EC.11(b)) are clearly very accurate. We note that the difference of the queue length for systems with expo(1) and ln(1,1) service time distributions is around 20 when $(\lambda, N) = (1100, 50)$. This shows the impact of service time distributions on system performance. The impact of patience time distributions for waiting and during service can be verified in a similar way. This consists with the approximations in §6.

Service Time	λ	N	$P^{Ab}{}_{\text{sim}}$	$P^{Ab}{}_{\text{approx}}$	Rel. Error (%)
expo(1)	550	25	0.3175(±0.0011)	0.3181	0.19
ln(1,1)	550	25	0.3243(±0.0012)	0.3250	0.22
expo(1)	1100	50	0.3175(±0.0012)	0.3181	0.19
ln(1,1)	1100	50	0.3242(±0.0012)	0.3250	0.25

(a) Relative error for P^{Ab} (in %)

Service Time	λ	N	Q_{sim}	Q_{approx}	Rel. Error (%)
expo(1)	550	25	140.9464(±5.8438)	139.8579	0.77
ln(1,1)	550	25	151.5935(±4.5686)	149.6885	1.26
expo(1)	1100	50	283.7241(±11.0558)	279.7157	1.41
ln(1,1)	1100	50	303.7658(±9.0692)	299.3770	1.44

(b) Relative error for Q (in %)

Table EC.11 Comparison of simulation results and approximations for overloaded systems