Approximations for chat service systems using many-server diffusion limits*

Leon Cui  
Simon Graduate School of Business  
University of Rochester  
leon.cui@simon.rochester.edu,  

Tolga Tezcan  
Simon Graduate School of Business  
University of Rochester  
tolga.tezcan@simon.rochester.edu

We study a queueing model of customer service chat (CSC) systems. A unique feature of these queueing systems is that a single agent can serve multiple customers simultaneously. We prove the convergence of the queueing process to different diffusion processes in a many server heavy-traffic regime in three different cases. Using this result, we are able to offer approximations for the steady state performance measures such as the number of customers in the system, the abandonment probability and the sojourn time of a customer. Our numerical experiments show that proposed approximations are accurate in various cases.

Key words: customer service chat systems, queueing theory, heavy-traffic many-server regime, diffusion approximations
MSC2000 subject classification: Primary: 60K25; secondary: 90B22, 60F17
OR/MS subject classification: Primary: queues; secondary: diffusion models, limit theorems
History: Submitted: August 6, 2013

1. Introduction  
Contact centers are an integral part of all organizations since they provide a way for organizations to interact with their customers or clients. Traditionally, phone service has been the most common method of communication, but with the growth of Internet usage, customer service chat (CSC) systems have become an important tool as well. Rather than calling in to the call center, in a CSC system customers are able to access an instant messaging system built into the organization’s website to interact with customer service representatives online. From an operational point of view, the main difference between CSC systems and call centers is that while a traditional call center agent can only serve a single customer at once, a CSC agent can serve multiple customers simultaneously. CSC systems also have unique features such as screen sharing and the ability to share files and data, which are particularly useful to computer companies, software companies, and e-retailers. There are also some disadvantages to CSC systems, including a longer service time due to extra typing and reading time, an inability to access the service while traveling, and frustration caused by technological barriers Shae et al. [21]. However, with the prevalence of proficient computer users, CSC systems remain an important communication channel between the organization and customers TELUS International [22].

From a management perspective, there are two important decisions that are imperative for the effective manage these systems (which are also common to most service systems): (i) Staffing problem: find the staffing level required in the CSC system to meet a certain service level requirement,

* Research supported by NSF Grants CMMI-0954126 and CMMI-1130346.
Routing problem: manage the CSC system in real time by deciding which available agent an arriving customer should be routed to. Clearly these problems are closely related. The analysis of chat-service systems is different from that of the traditional many-server queueing systems (see Dai and He [6], Gans et al. [10], Halfin and Whitt [12]) due to the fact that agents can serve multiple customers at a time (we refer to the number of customers an agent is serving as the level of an agent). In Tezcan and Zhang [24] and [25] effective solutions for the routing problem have been identified using a fluid limit approach. They also offer solutions for the staffing problem and approximations for the system performance.

Our goal in this paper is to use the more refined diffusion limits (see Dai and He [6], Gans et al. [10], or Halfin and Whitt [12] for details) rather than the fluid limits that are employed in Tezcan and Zhang [24] and [25] to obtain improved approximations for the system performance. These approximations can be used to address the staffing problem mentioned above more effectively. We show that especially for cases when the arrival rate is close to a critical value (more specifically when it is close to the total service rate when all the agents are at the same level) diffusion limits provide more accurate approximations than fluid limits. In addition, the diffusion limits enable us to build approximations for the whole distribution of the number of customers in system whereas with fluid limits only approximations for the average number of customers in system can be obtained.

We model a CSC system as a processor sharing queue where each agent can serve up to $K$ customers. A customer is served by the same agent throughout service and the service rate of a customer depends on the level of the agent serving that customer. We also assume that customers have limited patience and they may abandon the system from the queue or service. To establish the diffusion limits of the underlying queueing system we consider a many-server heavy traffic regime. Specifically we consider a sequence of systems where the arrival rate into the system gets large, and the staffing level for each system in the sequence is such that the relationship between the arrival rate and staffing level satisfies a condition that is similar to the so-called square root staffing rule (see Dai and He [6], Gans et al. [10], Halfin and Whitt [12]). We mainly consider the “lightest-load-first” policy which routes an arriving customer to the least busy agent, see Luo and Zhang [15]. The asymptotic optimality of this policy under certain conditions is proved in Tezcan and Zhang [24]. Furthermore we assume that the service time distribution is exponential. Under these assumptions, we show that the processes for the number of customers in different levels in the system converge to three different diffusion processes in three different cases that arise based on different assumptions about the limiting behavior of the arrival rate and the number of agents.

To obtain approximations for the distribution of the number of customers in the system and the sojourn time of an actual CSC system, we first establish the steady state quantities of the diffusion limits obtained, either exactly or using approximations. Then we build approximations for the number of customers and the abandonment probability in the actual system. Next we build approximations for the the distribution of the sojourn time in steady state. Results based on Puhalskii [18] have been used in the literature to estimate the sojourn time distribution from the queue length distribution, based on a generalized Little’s Law, see for example Garnett et al. [11], Puhalskii and Reiman [19], Tezcan [23]. However, beyond the mean sojourn time, it does not seem that such an approach would work in our case. Therefore, we supply two heuristic approximations for the sojourn time distribution. The first one is based on a simplified Markov chain whose transition rates are estimated using the approximations built for the number of customers in different levels and the second one is based on a simplified routing scheme.

Our proofs are based on a novel representation of the queueing equations for the CSC systems. Three different parameter regimes lead to three different limiting processes based on state space collapse (SSC) results (see Bramson [3], Dai and Tezcan [8]). Once we establish the associated SSC result, we prove that the system equations can be written as a continuous map of certain primitive processes. We conclude the proof of convergence to a diffusion limit by establishing the convergence of these primitive processes.
systems, beginning with the seminal paper by Halfin and Whitt [12]. There is a rich literature on many-server diffusion approximations of queueing systems, with number of agents ranging from 25 to 100. It is not clear which of the three aforementioned cases is more applicable to a given CSC system. We test which approximation is more accurate in estimating the system performance for the expected number of customers in system and its variance in steady state under different scenarios. For the sojourn time, we test the two proposed methods in similar systems. We observe that both proposed procedures for estimating the sojourn time are quite accurate and make observations about how the system parameters affect the accuracy of these approximations. In general, even for systems with 25 agents, our approximations for the expected number of customers and the sojourn time in system as well as the abandonment probability are within 2% of the simulation estimates in most cases. As the system size gets large, our approximations are much more accurate with errors less than 0.5%. Our approximations for the variance of the number of customers and sojourn time are within 5-6% of the simulation estimates on average for small systems and within 2-3% for systems with 100 agents.

The rest of this paper is organized as follows: In §2, we introduce a model of the CSC system under the “lightest-load-first” routing policy. In §3, we analyze this CSC system using fluid limits. Next, we analyze the same CSC system using diffusion limits. We split our analysis into three different cases which depend on how the arrival rate compares to the total service rate. These three cases are analyzed in §4, §5, and §6. In §7, we use the diffusion limits derived in §4, §5, and §6 to develop approximations for the number of customers in the system and abandonment probability and we provide two methods for estimating the distribution of the sojourn time. We present the results of our numerical experiments in §8 and conclude in §9. Appendix C extends our results to the more sophisticated routing policies proposed in Tezcan and Zhang [24], but in that case we are unable to find closed-form expressions for the steady state of the limiting diffusion process.

**Notation** All random variables and processes are defined on a common probability space \((\Omega, \mathcal{F}, \mathcal{P})\) unless otherwise specified. The symbols \(\mathbb{Z}, \mathbb{Z}_+, \mathbb{N}, \mathbb{R}\) and \(\mathbb{R}_+\) are used to denote the sets of integers, nonnegative integers, positive integers, real numbers and nonnegative real numbers, respectively. For \(d \in \mathbb{N}\), \(\mathbb{R}^d\) denotes the \(d\)-dimensional Euclidean space; thus, \(\mathbb{R} = \mathbb{R}^1\). The space of functions \(f: \mathbb{R}_+ \to \mathbb{R}^d\) that are right-continuous on \([0, \infty)\) and have left limits in \((0, \infty)\) is denoted by \(\mathbb{D}(\mathbb{R}_+, \mathbb{R}^d)\) or simply \(\mathbb{D}^d\); similarly, with \(T > 0\), the space of functions \(f: [0, T] \to \mathbb{R}^d\) that are right-continuous on \([0, T)\) and have left limits in \((0, T]\) is denoted by \(\mathbb{D}([0, T], \mathbb{R}^d)\). For \(f \in \mathbb{D}^d\), \(f(t-)\) denotes its left limit at \(t > 0\). For a sequence of random elements \(\{X^n, n \in \mathbb{N}\}\) taking values in a metric space, we write \(X^n \Rightarrow X\) to denote the convergence of \(X^n\) to \(X\) in distribution or weak convergence. Each stochastic process whose sample paths are in \(\mathbb{D}^d\) is considered to be a \(\mathbb{D}^d\)-valued random element. The space \(\mathbb{D}^d\) is assumed to be endowed with the Skorohod \(J_1\) topology (see Ethier and Kurtz [9] or Billingsley [2]). Given \(x \in \mathbb{R}\), we set \(x^+ = \max\{x, 0\}\) and \(x^- = \min\{x, 0\}\).

1.1. Literature Review In this section, we review three separate streams of literature that are related to CSC system queueing models.

Processor sharing queues: A stream of literature that is closely related to the CSC systems that we study are processor-sharing systems, which are common in communications literature to model the flow of data traffic. Similar to CSC systems, processor-sharing systems also have the feature that a single server can handle many different jobs at once. The difference is that all the servers in the processor-sharing system work jointly to finish every arriving job; in a CSC system each arriving customer’s service is handled by a single server. For more details, the reader is referred to Altman et al. [1], Kleinrock [14], Zhang et al. [27], and the references contained therein.

Diffusion approximations: The methodology in this paper involves many-server diffusion approximations. There is a rich literature on many-server diffusion approximations of queueing systems, beginning with the seminal paper by Halfin and Whitt [12], which reduces the analysis of a
M/M/N queueing process to that of a simple diffusion process. Many subsequent papers have been devoted to extending Halfin and Whitt’s diffusion analysis. We give several examples: Puhalskii and Reiman [19] extend this analysis by considering phase-type service distributions. Garnett et al. [11] give a diffusion approximation for a M/M/N queueing process with customer abandonments. Finally, Whitt [26] studies the diffusion limit of a queueing system with limited capacity, in the special case where the service time is a mixture between an exponential distribution and a point mass at 0. For more details, the reader is referred to Dai and He [6] and Gans et al. [10]. To the best of our knowledge this is the first paper that addresses the diffusion limits for CSC systems.

**CSC systems:** The CSC system queueing model was formulated in Tezcan and Zhang [24]. In this study, fluid limits are established to approximate the behavior of the queueing system. Using these fluid limits, Tezcan formulates a LP to determine the staffing level required in the CSC system, and also designs a routing algorithm that is shown to be optimal in the steady state of the CSC system. In a parallel study, Luo and Zhang [15] also examine the staffing problem of the CSC system by investigating its fluid limit approximation with patient customers. However, they fix the routing policy and focus on both the transient and steady state behavior of the CSC system, whereas Tezcan and Zhang [24] focuses on the optimal routing policy in steady state. Finally, in Tezcan and Zhang [25], they extend their fluid analysis to incorporate general service time distributions. Whereas previous work in CSC system analysis has only used fluid approximations, our paper analyzes the system using many-server diffusion approximations. To our knowledge, this is the first study that uses such an approach.

2. Model and the asymptotic regime

In this section we first explain the queueing model we use to analyze CSC systems in §2.1. Then, we introduce the asymptotic regime that we use in §2.2.

2.1. Queueing model

We consider a system with a single class of customers that arrives according to a renewal process with rate λ, and a homogeneous pool of n agents, where each agent has the ability to serve up to K customers simultaneously. The coefficient of variation for the interarrival time is denoted by c2a. Hereafter we say that a agent is at “level k” if it is currently serving k customers, and “class k” refers to the pool of all agents that are currently serving k customers. We assume that the service time of a customer receiving service from an agent at level k is exponentially distributed with rate µk, for k ∈ {1, 2, . . . , K}. We also assume that customers have limited patience and they abandon from the queue if they wait too long or from service if their service is not completed in a reasonable time. To model this feature we assume that the abandonment time of a customer in service is exponentially distributed with rate νs, and the abandonment time from the queue is exponentially distributed with rate ν.

Arriving customers are routed according to the the lightest-load-first rule, that is, they are routed to one of the least busy agents (one of those agents serving the least number of customers). It is immaterial for our purposes which agent among the least busy agents (if there are more than one) is selected, we assume it is selected randomly for concreteness. (We use this assumption only in one of the methods to estimate the sojourn time.) If all agents are at level K, the customer joins the queue to await service. Once the customer is at the head of the queue, his service begins as soon as an agent completes a chat. Furthermore we assume that customers in the queue are served according to a FCFS discipline. Next, we define dk = k(µk + νs). The value dk represents the total departure rate from an agent at level k, including both service completions and abandonment during service. Finally, we assume that dk+1 ≥ dk, for k ∈ {1, 2, . . . , K}, that is, the total departure rate is increasing as the agent’s level increases. For further justification of this assumption, see Tezcan and Zhang [24].
Queueing model equations: Next we present the queueing equations that determine the
evolution of chat systems. Define $Z_k(t)$ to be the number of agents at level $k$ at time $t$. Let $A(t)$ be
the total number of customers arriving into the queuing system by time $t$. Next, define $Q(t)$ as the
number of customers waiting in the queue at time $t$. Let $A_k(t)$ denote the total number of arrivals routed
to class $k$, for $k \in \{0, 1, \ldots, K-1\}$ and $A_K(t)$ denote the number of arrivals that are sent
into the queue by time $t$ at the time of arrival.

To model customer departures from the system first we let $S_k$ be a Poisson process with rate 1, for
$k \in \{1, 2, \ldots, K+1\}$. Then $S_k \left( d_k \int_0^t Z_k(s)ds \right)$ represents the total number of departures from
level $k$, for $k = 1, \ldots, K$ and $S_{K+1} \left( \nu \int_0^t Q(s)ds \right)$ denotes the number of abandonments from the
queue by time $t$. We assume that $A$ and $S_k$ are mutually independent for all $k \in \{1, 2, \ldots, K+1\}$. For
notational simplicity we define

$$D_k(t) = S_k \left( d_k \int_0^t Z_k(s)ds \right), \text{ for } k = 1, \ldots, K \text{ and } D_Q(t) = S_{K+1} \left( \nu \int_0^t Q(s)ds \right).$$

Also we set $Z = (Z_0, \ldots, Z_K)$, and $A = (A_0, \ldots, A_K)$. The queueing equations are given by

\begin{align}
Z_0(t) &= Z_0(0) - A_0(t) + D_1(t) \quad (1) \\
Z_k(t) &= Z_k(0) - A_k(t) + A_{k-1}(t) - D_k(t) + D_{k+1}(t), \text{ for } k \in \{1, 2, \ldots, K-2\}, \quad (2) \\
Z_{K-1}(t) &= Z_{K-1}(0) - A_{K-1}(t) + A_{K-2}(t) - D_{K-1}(t) + L_Q(t), \quad (3) \\
Q(t) &= Q(0) + A_K(t) - D_K(t) - D_Q(t) + L_Q(t), \quad (4)
\end{align}

where $L_Q$ is defined as

$$L_Q(t) = \int_0^t 1_{\{Q(s) = 0\}} dD_K(s), \quad (5)$$
for all $t > 0$, and

$$Z_K(t) = n - \sum_{k=0}^{K-1} Z_k(t). \quad (6)$$

Furthermore $(Z, A, Q)$ must satisfy

\begin{align}
Q(t)Z_k(t) &= 0, \text{ for } k \in \{0, 1, \ldots, K-1\}, \quad (7) \\
\int_0^t Z_k(s) dA_m(s) &= 0, \text{ for } m \geq k+1 \text{ and } k \in \{0, 1, \ldots, K-1\}. \quad (8)
\end{align}

Equations (1)-(2) follow from the fact that once a customer is assigned to an agent at level $k$
for $k \in \{1, 2, \ldots, K-1\}$, the agent goes to level $k+1$. Similarly, once a customer leaves the system
from an agent at level $k$, that agent goes to level $k-1$. However, this is not true for an agent
at level $K$; an agent at level $K$ goes to level $K-1$ after a customer departure only if the queue
is empty, otherwise that agent remains at level $K$. Hence $L_Q$ defined in (5) determines when an
agent goes from level $K-1$ from $K$. Finally $(7)$ and $(8)$ follow from our routing policy. Equation
$(7)$ guarantees that customers join the queue only when all the agents are serving $K$ customers. In
addition, $(8)$ implies that customers will only be routed to a level if all the classes below that level
are empty.

We also note that

$$Z_K(t) = Z_K(0) + A_{K-1}(t) - L_Q(t). \quad (9)$$
For notational simplicity we set

\[ X(t) = Z_K(t) + Q(t). \]

By (4) and (9)

\[ X(t) = X(0) + A_{K-1}(t) + A_K(t) - D_K(t) - D_{K+1}(t). \quad (10) \]

**Another form of queueing model equations:** Before we turn our attention to the analysis of chat systems we next write equations (1)-(4) in a form more amenable to analysis. This new form plays a crucial role in our analysis. We first define \( L_k(t) \) to be the number of customers that have been routed to either to the queue or one of the agents in levels \( k+1 \) or higher by time \( t \), for \( k = -1, 0, \ldots, K-1 \) and so \( L_{-1} = A \). Using the notation defined above

\[ L_k(t) = \sum_{t=k+1}^K A_t, \quad \text{for } k = -1, 0, 1, \ldots, K-1. \]

Observe that \( A_k(t) = L_{k-1}(t) - L_k(t) \) for \( k \in \{0, 1, \ldots, K-1\} \) and \( A_K(t) = L_{K-1}(t) \) for \( t \geq 0 \). With this new definition, we can rewrite (1)-(3) as

\[
\begin{align*}
Z_0(t) &= Z_0(0) - A(t) + D_1(t) + L_0(t), \\
Z_k(t) &= Z_k(0) - 2L_{k-1}(t) + L_{k-2}(t) - D_k(t) + D_{k+1}(t) + L_k(t), \quad \text{for } k \in \{1, 2, \ldots, K-2\}, \\
Z_{K-1}(t) &= Z_{K-1}(0) - 2L_{K-2}(t) + L_{K-3}(t) - D_{K-1}(t) + L_Q(t) + L_{K-1}(t), \\
Q(t) &= Q(0) + L_{K-1}(t) - D_K(t) - D_Q(t) + L_Q(t),
\end{align*}
\]

and

\[
\begin{align*}
\int_0^t Q(s)dL_Q(s) &= 0, \\
\int_0^t Z_k(s)dL_k(s) &= 0, \quad \text{for } k \in \{0, 1, \ldots, K-1\}.
\end{align*}
\]

Clearly (11)–(14) follow from (1)–(3), respectively. (15) follows from (5) and (16) follows from (8). By (10) we have

\[ X(t) = X(0) + L_{K-2}(t) - D_K(t) - D_{K+1}(t). \quad (17) \]

The new form of queueing equations (11)–(17) will prove useful to define a continuous map based on the reflection condition (16) that we use to prove diffusion limits below.

**2.2. Asymptotic regime** We focus on the asymptotic analysis of chat systems with many servers (as \( n \rightarrow \infty \)). Our main goal is to find an approximation for the steady state performance of these systems using diffusion limits. In general, the diffusion scaling is a refinement of the fluid scaling. While the fluid scaling is used to capture the mean value of a queueing process, the diffusion scaling is used to analyze the fluctuations of that process around its (fluid-scaled) mean value. Therefore as an intermediate step, we must first establish the fluid limits. In this section we describe the details of the asymptotic regime that we consider, and then define the fluid and diffusion scalings.

We consider a sequence of chat systems indexed by \( n \), the number of agents in the system. The arrival rate in the \( n \)th system is denoted by \( \lambda^n \), and is assumed to satisfy the following condition, which is similar to the well-known square root staffing rule (see Dai and He [6], Gans et al. [10], Halfin and Whitt [12], for details):

\[ \lambda^n = \lambda n - \beta \sqrt{n}, \quad (18) \]
for $\lambda > 0$. We append “$n$” to all the quantities in the $n$th system, e.g. $Z^n_0(t)$ gives the number of idle agents at time $t$ in the $n$th system. With this notation, equations (11)-(16) obviously still hold for the $n$th system.

The diffusion limits of the underlying queueing process depends on the limiting value $\lambda$, as will be seen in our analysis that follows. More specifically we consider three different cases.

**Case I.** $d_j < \lambda < d_{j+1}$ for some $j \in \{1, 2, \ldots, K-1\}$,

**Case II.** $\lambda = d_j$ for some $j \in \{1, 2, \ldots, K-1\}$,

**Case III.** $\lambda = d_K$.

The main reason we consider three different cases is because it turns out that the diffusion limits are different and we need a different (continuous mapping) approach to prove convergence in each case. We do not consider the case when $\lambda > d_K$ because in that case the analysis is identical to the many-server asymptotic analysis of an overloaded $M/M/N + M$ queue.

**Remark 1.** As mentioned above our purpose is to build approximations for the system performance for a given number of servers $n$ and arrival rate $\lambda^n$. We will see that in each case we have a different approximation. Therefore in order to apply our results one needs to decide which case (among Cases I–III) is applicable for fixed parameters. However it is not clear which case is more applicable for a fixed $\lambda^n$ and $n$ since for certain $\beta$ values both Cases I and II (or Cases I and III) can be applicable. To illustrate, assume that $\lambda^n = 100$, $d_1 = 0.5$, $d_2 = 1.5$, and $n = 100$. Then we can take $j = 1$, $\beta = 0$ and use Case I or $j = 2$, $\lambda = 1.5$ and $\beta = 5/1.5$ and use Case II. Intuitively, the asymptotic regime in Case II (when $\lambda = d_j$) should be more applicable when $\lambda^n$ is “close” to $d_j n$ for some $j$ and Case I should be more applicable when $\lambda^n$ is in between $d_j n$ and $d_{j+1} n$, for some $j$. In what follows we first establish the diffusion limits in each case and build approximations for the actual system based on stationary distribution of the limiting process. We explore which case is more applicable afterwards using numerical results and provide simple guidelines in §8.

For the rest of this section our purpose is to define a general notion of fluid and diffusion scaling that captures all three cases, to avoid redefining these concepts repeatedly for each case. Next, we define the fluid scaled processes

\begin{align*}
\bar{A}^n(t) &= n^{-1} A^n(t), \\
\bar{Z}^n_k(t) &= n^{-1} Z^n_k(t), \text{ for } k \in \{0, 1, \ldots, K\}, \\
\bar{L}^n_k(t) &= n^{-1} L^n_k(t), \text{ for } k \in \{0, 1, \ldots, K-1\}, \\
\bar{Q}^n(t) &= n^{-1} Q^n(t), \\
A^n_k(t) &= n^{-1} A^n_k(t), \text{ for } k \in \{0, 1, \ldots, K\}, \\
D^n_k(t) &= n^{-1} D^n_k(t) = n^{-1} S_k \left( nd_k \int_0^t \bar{Z}^n_k(s) ds \right), \text{ for } k \in \{1, \ldots, K-1\}, \\
\bar{D}^n_K(t) &= n^{-1} D^n_K(t) = n^{-1} S_{K+1} \left( n\nu \int_0^t \bar{Q}^n(s) ds \right), \\
\bar{X}^n(t) &= \bar{Z}^n_K(t) + \bar{Q}^n(t),
\end{align*}

and define $Z^n = (\bar{Z}^n_0, \ldots, \bar{Z}^n_K)$, $\bar{Z}^n = (\bar{Z}^n_0, \ldots, \bar{Z}^n_K)$, $A^n = (A^n_0, \ldots, A^n_K)$, $\bar{A}^n = (\bar{A}^n_0, \ldots, \bar{A}^n_K)$, $L^n = (\bar{L}^n_0, \ldots, \bar{L}^n_K)$, $\bar{L}^n = (\bar{L}^n_0, \ldots, \bar{L}^n_K)$, $D^n = (D^n_0, \ldots, D^n_K, D^n_K)$, and $\bar{D}^n = (\bar{D}^n_0, \ldots, \bar{D}^n_K, \bar{D}^n_K)$. To obtain meaningful limits we assume that

\begin{equation}
(\bar{Z}^n(0), \bar{Q}^n(0)) \to (\bar{Z}(0), \bar{Q}(0)) \text{ a.s. as } n \to \infty,
\end{equation}

for random vectors $\bar{Z}(0) \in \mathbb{R}_+^K$ and $\bar{Q}(0) \in \mathbb{R}_+$.

As a first step towards obtaining our diffusion approximations, we show that under assumptions (18) and (28) the fluid-scaled process $(\bar{Z}^n, \bar{Q}^n, \bar{L}^n)$ are tight and we provide a set of equations that
are satisfied by the fluid limits \((\hat{Z}, \hat{Q}, \hat{L})\). We also show that the fluid limits have an invariant state, denoted by \((\hat{Z}(\infty), \hat{Q}(\infty))\) with \(\hat{Z}(\infty) = (\hat{Z}_0(\infty), \hat{Z}_1(\infty), \ldots, \hat{Z}_K(\infty))\). The invariant state will be different for each case mentioned above. Given the invariant state of the fluid limits we are able to define the diffusion scaling, assuming that the initial state of the system is at the invariant state for the fluid limit, that is, we assume

\[
(Z(0), X(0)) = (\hat{Z}(\infty), \hat{Q}(\infty)).
\] (29)

Next we define the diffusion scaling by

\[
\hat{Z}_k^n(t) = \sqrt{n}(\hat{Z}_k^n(t) - \hat{Z}_k(\infty)), \quad \text{for } k \in \{0, 1, \ldots, K\},
\] (30)

\[
\hat{Q}_n(t) = \sqrt{n}(\hat{Q}_n(t) - \hat{Q}(\infty)),
\] (31)

\[
\hat{X}_n(t) = \hat{Z}_K^n(t) + \hat{Q}_n(t),
\] (32)

and define \(\hat{Z}^n = (\hat{Z}_0^n, \ldots, \hat{Z}_K^n)\). Finally in establishing the diffusion limits, we assume that

\[
(\hat{Z}_n(0), \hat{Q}_n(0)) \Rightarrow (\hat{Z}(0), \hat{Q}(0)) \text{ as } n \to \infty,
\] (33)

where \((\hat{Z}(0), \hat{Q}(0))\) is a random vector with \((\hat{Z}_n(0), \hat{Q}_n(0))\) satisfying (7) and \(\hat{Z}(0) \geq 0\) and \(\hat{Q}(0) \geq 0\).

3. Fluid limits of CSC queueing models In this section, we first establish the fluid model equations as the limit of the fluid scaled processes. The results will be used to establish the existence of fluid invariant states \((\hat{Z}(\infty), \hat{Q}(\infty))\) below in Theorem 2. Recall that the fluid invariant states are needed in the definition of diffusion scaling (30)–(32).

**Theorem 1.** Assume (18) and (28) hold. Then \(\{\hat{Z}_n^\alpha, \hat{Q}_n^\alpha, \hat{A}_n^\alpha, \hat{L}_n^\alpha, \hat{L}_Q^n\}\) is a.s. tight. Furthermore every limit \((\hat{Z}, \hat{Q}, \hat{A}, \hat{L}, \hat{L}_Q)\) with \(\hat{Z}(t) = (\hat{Z}_1(t), \ldots, \hat{Z}_K(t)), \hat{A}(t) = (\hat{A}_0(t), \ldots, \hat{A}_{K-1}(t)), \hat{L}(t) = (\hat{L}_0(t), \ldots, \hat{L}_{K-1}(t))\), satisfies the following fluid model equations

\[
\hat{Z}_0(t) = \bar{Z}_0(t) - \bar{A}_0(t) + d_1 \int_0^t \hat{Z}_1(s) ds,
\] (34)

\[
\hat{Z}_k(t) = \bar{Z}_k(t) - \bar{A}_k(t) + \bar{A}_{k-1}(t) - d_k \int_0^t \hat{Z}_k(s) ds + d_{k+1} \int_0^t \hat{Z}_{k+1}(s) ds, \quad \text{for } k = 1, \ldots, K-2,
\] (35)

\[
\hat{Z}_{K-1}(t) = \bar{Z}_{K-1}(t) - \bar{A}_{K-1}(t) + \bar{A}_{K-2}(t) - d_{K-1} \int_0^t \hat{Z}_{K-1}(s) ds + \bar{L}_Q(t),
\] (36)

\[
\hat{Q}(t) = \bar{Q}(t) + \bar{A}_K(t) - d_K \int_0^t \hat{Z}_K(s) ds - \nu \int_0^t \bar{Q}(s) ds + \bar{L}_Q(t),
\] (37)

\[
\bar{X}(t) = \bar{X}(0) + \bar{A}_{K-1}(t) + \bar{A}_K(t) - d_k \int_0^t \hat{Z}_k(s) ds - \nu \int_0^t \bar{Q}(s) ds,
\] (38)

\[
\bar{Q}(t) \bar{Z}_k(t) = 0, \quad \text{for } k = 0, 1, \ldots, K-1,
\] (39)

\[
\bar{Z}_k(t) d\bar{A}_j(t) = 0, \quad \text{for } k = 0, 1, \ldots, K-1 \text{ and } j \geq k + 1,
\] (40)

\[
\bar{Q}(t) d\bar{L}_Q(t) = 0,
\] (41)

\[
\bar{L}_Q(t) \leq d_K \int_0^t \hat{Z}_K(s) ds,
\] (42)

\[
\bar{L}_k(t) = \sum_{t=k+1}^K \bar{A}_t(t), \quad k = -1, \ldots, K-1,
\] (43)

\[
\sum_{k=0}^K \bar{Z}_k(t) = 1,
\] (44)

\[
(Z(t), \bar{Q}(t), \bar{A}(t), \bar{L}(t), \bar{L}_Q(t)) \geq 0,
\] (45)

\[
\sum_{k=0}^K \bar{A}_k(t) = \lambda t,
\] (46)
In addition \((\bar{Z}, \bar{Q}, \bar{A}, \bar{L}, \bar{L}_Q)\) is differentiable almost everywhere.

A fluid model analysis of chat systems is provided in Tezcan and Zhang [24] under a general policy. Here our queueing model equations are slightly different and due to our relatively simpler policy, our proof is much shorter. We provide a detailed proof in Appendix A for completeness. We refer to the set of equations \((34)-(46)\) as the fluid model equations and any \((\bar{Z}, \bar{Q}, \bar{A}, \bar{L}, \bar{L}_Q)\) that satisfies \((34)-(46)\) as a fluid model solution.

**Invariant state of the fluid models:** Next we identify the invariant state of fluid model solutions. In the current context \((\bar{Z}(1), \bar{Q}(1))\) is said to be an invariant state of the fluid model if

\[
(\bar{Z}(0), \bar{Q}(0)) = (\bar{Z}(\infty), \bar{Q}(\infty))
\]

for any fluid model solution implies that

\[
(\bar{Z}(t), \bar{Q}(t)) = (\bar{Z}(\infty), \bar{Q}(\infty))
\]

for all \(t > 0\). Note that under Cases I and II

\[
\lambda = \gamma d_j + (1 - \gamma)d_{j+1},
\]

for some \(j \in \{1, 2, \ldots, K - 1\}, 0 \leq \gamma \leq 1\), and \(\lambda\) defined in \((18)\). Under Case III

\[
\lambda = d_K.
\]

**Theorem 2.** The fluid model equations \((34)-(46)\) have an invariant state denoted by \((\bar{Z}(\infty), \bar{Q}(\infty))\).

i. In Cases I and II (when \((47)\) holds) the invariant state is given by

\[
(\bar{Z}_j(\infty), \bar{Z}_{j+1}(\infty)) = (\gamma, 1 - \gamma),
\]

\(\bar{Z}_k(\infty) = 0\) for \(k \neq j, j + 1\), \(\bar{Q}(\infty) = 0\). Moreover, if \((\bar{Z}(0), \bar{Q}(0)) = (\bar{Z}(\infty), \bar{Q}(\infty))\) we have

\[
\bar{L}_k(t) = \lambda t, \text{ for } k = 0, \ldots, j - 2, \bar{L}_{j-1}(t) = \lambda t - d_j \gamma t
\]

and \(\bar{L}_k(t) = 0\) for \(k = j, j + 1, \ldots, K - 1\).

ii. In Case III (when \((48)\) holds) the invariant state is given by

\[
\bar{Z}_K(\infty) = 1,
\]

\(\bar{Z}_k(\infty) = 0\) for \(k \neq K\), \(\bar{Q}(\infty) = 0\). Moreover, if \((\bar{Z}(0), \bar{Q}(0)) = (\bar{Z}(\infty), \bar{Q}(\infty))\) we have

\[
\bar{L}_k(t) = \lambda t, \text{ for } k = 0, 1, \ldots, K - 1.
\]

**Remark 2.** Based on Theorem 7 in Tezcan and Zhang [24] the invariant states are also the steady states of the corresponding fluid models in each case (this fact is not needed for the remainder of the paper). Because we focus only on the lightest-load-first policy we are also able to show that the steady states are invariant states as well, a fact that will be needed in our analysis of diffusion limits.
Proof of Theorem 2: We mainly focus on Case I, so assume that (47) holds and $0 < \gamma < 1$ and for simplicity we assume that $j + 1 < K$. By Theorem 1, $(\bar{Z}, \bar{Q})$ is a.e. differentiable. For the remainder of the proof we only consider regular points of $t$ where the derivative $(\dot{\bar{Z}}(t), \dot{\bar{Q}}(t))$ exists. We also use the following fact in a number of places in the proof; if $(\dot{\bar{Z}}(t), \dot{\bar{Q}}(t))$ exists and $\bar{Z}_k(t) = 0$, then $\dot{\bar{Z}}_k(t) = 0$ as well, since at time $t$, $\bar{Z}_k$ attains its minimum.

We first show that in the invariant state, $\bar{Z}_k(\infty) = 0$ for $k \in \{0, 1, \ldots, j - 1\}$. To do so we define the Lyapunov function $f_1$, similar to that used in Theorem 7 of Tezcan and Zhang [24], as follows

$$f_1(t) = \sum_{k=0}^{j-1} (j-k)\bar{Z}_k(t).$$

Therefore by (34) and (35)

$$f_1(t) = -\sum_{k=0}^{j-1} \dot{A}_k(t) + \sum_{k=0}^{j} \dot{D}_k(t),$$

where

$$\dot{D}_k(t) = d_k \int_0^t \bar{Z}_k(s) ds.$$

It suffices to show that for any $t > 0$, if $f_1(t) > 0$ then $\dot{f}_1(t) \leq 0$. Note that if $f_1(t) > 0$ then $\dot{Z}_k(t) > 0$ for some $k = 0, \ldots, j - 1$. Therefore, by (34), (35), (40), (44), (46), (47) and (49), when $f_1(t) > 0$

$$\dot{f}_1(t) = -\lambda + \sum_{k=1}^{j} d_k \bar{Z}_k(t) \leq 0,$$

because $d_k$ is increasing in $k$.

We next show that $\bar{Z}_k(\infty) = 0$ for all $k \in \{j + 2, j + 3, \ldots, K\}$ by using a similar approach. First, by (38) and (40)

$$\dot{X}(t) \leq \max\{\lambda - d_K, -(d_K \wedge \nu)\bar{X}(t)\}.$$

By (47), $\lambda < d_K$, hence $\bar{X}(t) = 0$ for all $t \geq 0$ if $\bar{X}(0) = 0$.

Next we define the Lyapunov function

$$f_2(t) = \sum_{k=j+2}^{K-1} (k-j-1)\bar{Z}_k(t).$$

Because $\bar{Z}_K(0) = 0$, $L_\mathcal{Q}(t) = 0$ for all $t$ by (39), (41) and by (38) $\dot{A}_K(t) = \dot{A}_{K-1}(t) = 0$. Hence

$$f_2(t) = \sum_{k=j+1}^{K-2} \dot{A}_k(t) - \sum_{k=j+2}^{K-1} d_k \bar{Z}_k(t)$$

by (35) and (36).

We split our analysis of $f_2$ into two cases: either (i) $\bar{Z}_j(t) > 0$ or (ii) $\bar{Z}_j(t) = 0$. If $\bar{Z}_j(t) > 0$, then from (35)-(37), and (40), (50) gives

$$\dot{f}_2(t) = -\sum_{k=j+2}^{K} d_k \bar{Z}_k(t) \leq 0.$$
In the other case if \( \dot{Z}_j(t) = 0 \), then from (35)-(37), (46), and the fact that \( \dot{Z}_k(t) = 0 \) for \( k \in \{0, 1, \ldots, j - 1\} \) for all \( t \geq 0 \), we have that \( \dot{A}_j(t) = d_{j+1} \dot{Z}_{j+1}(t) + \sum_{k=j+1}^{K-1} \dot{Z}_k(t) = 1 \) by (44). Thus

\[
\dot{f}_2(t) = \lambda - \sum_{k=j+1}^{K} d_k \dot{Z}_k(t) \leq 0,
\]

where the last inequality follows from (47) and our assumption that \( d_k \) is increasing in \( k \).

It remains to show the invariant state of \( \dot{Z}_j \) and \( \dot{Z}_{j+1} \). We already showed that \( \dot{Z}_k(t) = \dot{Q}(t) = 0 \) for all \( k \neq j, j+1 \) and \( t \geq 0 \), and so by (44), \( \dot{Z}_j(t) + \dot{Z}_{j+1}(t) = 1 \). In addition by (34)-(36) \( \dot{A}_k(t) = 0 \) for \( k = 0, 1, \ldots, j - 2 \) and \( k = j + 1, \ldots, K \), for all \( t \geq 0 \). Because \( \dot{Z}_{j-1}(t) = 0 \) for all \( t \geq 0 \), we have \( \dot{A}_{j-1}(t) = d_j \dot{Z}_j(t) \) by (34)-(36) for all \( t \geq 0 \). Combining this fact with (35), (46), and (47), we see that

\[
\dot{Z}_j(t) = d_j \dot{Z}_j(t) + d_{j+1}(1 - \dot{Z}_j(t)) - (d_j \gamma + d_{j+1}(1 - \gamma)).
\]

By (51), for any \( t > 0 \), if \( \dot{Z}_j(t) > \gamma \), \( \dot{Z}_j(t) < 0 \) and if \( \dot{Z}_j(t) < \gamma \), \( \dot{Z}_j(t) > 0 \). This completes the proof for \( (\dot{Z}(\infty), \dot{Q}(\infty)) \). The expression for \( \dot{L}(\infty) \) can easily be verified by substituting the expression for \( (\dot{Z}(\infty), \dot{Q}(\infty)) \) into (34)-(37) and (43). The proof for the case \( j + 1 = K \) is similar except that instead of showing \( \dot{X}(t) = 0 \) for all \( t \geq 0 \) we only show that \( \dot{Q}(t) = 0 \) for all \( t \geq 0 \) using (37). The last step of the proof is identical afterwards.

For Case II the proof is very similar. The result in the first and second steps follow similarly. The argument leading to (51) is not needed since there is only one level. In Case III, the first step is still valid. Also, for the second step it is not difficult to see that \( \dot{Q}(t) < 0 \) if \( \dot{Q}(t) > 0 \), if (48) holds. □

4. Diffusion approximations in Case I

In this section we establish the diffusion limit \( (\dot{Z}, \dot{Q}) \) of the diffusion-scaled process \( (\dot{Z}^n, \dot{Q}^n) \) in Case I. We then find the stationary distribution of the limiting diffusion process \( (\dot{Z}, \dot{Q}) \), which we later to use to approximate the steady state of the actual CSC system.

Recall that in Case I we assume that \( \gamma \) defined in (47) satisfies \( 0 < \gamma < 1 \). Hence we have

\[
\lambda^n = \lambda n - \beta \lambda \sqrt{n}, \ n \geq 1 \text{ and } \lambda = \gamma d_j + (1 - \gamma)d_{j+1},
\]

for some \( 0 < \gamma < 1 \) and \( j = 0, \ldots, K - 1 \) and \( \beta \). We also make the following assumption about the initial state which we need to establish a state space collapse result, see Remark 3 below for more details;

\[
(\dot{Q}^n(0), \dot{Z}^n_k(0) \text{ for } k \in \mathcal{S}_1) \Rightarrow 0, \text{ in } \mathbb{R}^K \text{ as } n \to \infty.
\]

where \( \mathcal{S}_1 = \{0, 1, \ldots, K\} \setminus \{j, j + 1\} \).

**Theorem 3.** Assume that (28), (29), (33), (52), and (53) hold, and that \( 0 < \gamma < 1 \). Then \( (\dot{Z}^n, \dot{Q}^n) \Rightarrow (\dot{Z}, \dot{Q}) \) in \( \mathbb{D}^{K+2} \) as \( n \to \infty \), where \( (\dot{Z}, \dot{Q}) \) is the unique solution to the stochastic differential equation

\[
\dot{Z}_k(t) = 0, \text{ for } k \in \mathcal{S}_1, \quad (54)
\]

\[
\dot{Z}_j(t) = \dot{Z}_j(0) + \lambda \beta t - (d_{j+1} - d_j) \int_0^t \dot{Z}_j(s)ds + B(t), \quad (55)
\]

\[
\dot{Z}_{j+1}(t) = -\dot{Z}_j(t), \quad (56)
\]

\[
\dot{Q}(t) = 0, \quad (57)
\]

and \( B \) is a driftless Brownian motion with variance \( \lambda(1 + c_\alpha^2) \).
Remark 3. The result (54) in general is known as a state space collapse result, which shows that the total number of agents in those levels in $S_1$ goes to zero. Before we present the proof, we next provide the stationary distribution of the limiting process. Because $Z_j$ is an Ornstein-Uhlenbeck process the following result is well-known (see Browne and Whitt [4] for details).

Theorem 4. The diffusion process $\hat{Z}_j$ has a stationary distribution which is normal with mean $-\frac{\lambda \beta}{d_j + 1 + d_j}$ and variance $\frac{\lambda (1 + \gamma^2)}{2(d_j + 1 - d_j)}$.

Proof of Theorem 3: Assume that (28), (29), (33), (52), and (53) hold, and that $0 < \gamma < 1$. We next write the queueing equations for the diffusion scaled processes defined in (30)-(32). Let

$$A^n(t) = A^n(t) - \frac{\lambda^n}{n} t,$$

and $\hat{A}^n(t) = \sqrt{n} A^n(t)$, $\hat{S}_k^n(t) = \sqrt{n} \hat{S}_k^n(t)$ for $k \in \{1, 2, \ldots, K + 1\}$ where

$$\hat{S}_{K+1}(t) = \hat{D}^n(t) - \nu \int_0^t \hat{Q}^n(s) ds,$$

$$\hat{S}_k^n(t) = \hat{D}_k^n(t) - d_k \int_0^t \hat{Z}_k^n(s) ds, \text{ for } k \in \{1, 2, \ldots, K\}$$

Based on Theorem 2, we define

$$\hat{L}_k^n(t) = \sqrt{n} (\hat{L}_k^n(t) - \lambda t), \text{ for } k \in \{-1, 0, 1, \ldots, j - 2\},$$

$$\hat{L}_{j-1}(t) = \sqrt{n} (\hat{L}_{j-1}(t) - (\lambda t - d_j \gamma t))(= \sqrt{n} (\hat{L}_{j-1}(t) - d_{j+1}(1 - \gamma) t))$$

$$\hat{L}_k^n(t) = \sqrt{n} \hat{L}_k^n(t), \text{ for } k \in \{j, \ldots, K - 1\},$$

$$\hat{L}_{K+1}(t) = \sqrt{n} (\hat{L}_{K+1}(t) - d_K \hat{Z}_K(\infty) t).$$

By (11)–(17) we have

$$\hat{Z}_0^n(t) = \hat{Z}_0^n(0) + d_1 \int_0^t \hat{Z}_1^n(s) ds + \hat{L}_0^n(t) - \hat{A}^n(t) + \hat{S}_1^n(t) - \sqrt{n} \left( \frac{\lambda^n}{n} - \lambda \right) t,$$

$$\hat{Z}_k^n(t) = \hat{Z}_k^n(0) - 2 \hat{L}_{k+1}(t) + \hat{Z}_k^n(t) + d_k \int_0^t \hat{Z}_k^n(s) ds + d_{k+1} \int_0^t \hat{Z}_{k+1}^n(s) ds$$

$$- \hat{S}_k^n(t) + \hat{S}_{k+1}^n(t) + \hat{L}_k^n(t), \text{ for } k \in \{1, 2, \ldots, K - 1\},$$

$$\hat{Z}_{K-1}(t) = \hat{Z}_{K-1}(0) - 2 \hat{L}_{K-2}(t) + \hat{Z}_{K-1}(t) - d_{K-1} \int_0^t \hat{Z}_{K-1}(s) ds + \hat{L}_{K-1}(t) - \hat{S}_{K-1}(t) + \hat{L}_K^n(t),$$

$$\hat{Q}^n(t) = Q^n(0) + \hat{L}_k^n(t) - d_K \int_0^t \hat{Z}_K^n(s) ds - \nu \int_0^t \hat{Q}^n(s) ds + \hat{L}_K^n(t) - \hat{S}_K^n(t) - \hat{S}_{K+1}(t),$$

where $\hat{Z}_K^n(t) = - \sum_{k=0}^{K-1} \hat{Z}_k^n(t)$. We first show the convergence of $(\hat{A}^n, \hat{S}_j^n, \hat{S}_{j+1}^n)$. By Donsker’s theorem,

$$\hat{A}^n \Rightarrow \hat{A} \text{ in } D \text{ as } n \rightarrow \infty,$$

where $\hat{A}$ is a Brownian motion with drift 0 and variance $\lambda c_a^2$. Next we conclude that

$$(\hat{S}_j^n, \hat{S}_{j+1}^n) \Rightarrow (\hat{S}_j, \hat{S}_{j+1}) \text{ in } D^2 \text{ as } n \rightarrow \infty,$$

where $\hat{S}_j$ and $\hat{S}_{j+1}$ are independent driftless Brownian motions with variances $d_j \gamma$ and $d_{j+1}(1 - \gamma)$ respectively. This follows from Theorem 2, Donsker’s theorem, and (29).
Below we prove that
\[ (\hat{Q}^n, \hat{Z}_k^n \text{ for } k \in S_1) \Rightarrow 0 \text{ in } \mathbb{D}^K \text{ as } n \to \infty. \] (64)

Now we focus on \( \hat{Z}_j^n \) and \( \hat{Z}_{j+1}^n \): From Theorem 2 and (29) we know that \( \hat{Z}_j(t) = \gamma \) for all \( t \geq 0 \) a.s. Therefore, we conclude that \( (\hat{L}_j^n, \hat{L}_{j+1}^n) \to (0, 0) \) a.s. u.o.c. as \( n \to \infty \) by (16).

By (58), (59), and (64), progressing recursively starting from \( k = 0 \), we can find a process \( \epsilon^n \equiv (\epsilon^n_0, \ldots, \epsilon^n_{j-1}) \) such that \( \epsilon^n_k \Rightarrow 0 \) in \( \mathbb{D} \) as \( n \to \infty \) for \( k \in \{-1, 0, 1, \ldots, j-1\} \), and such that
\[
\hat{L}_k^n(t) = \hat{A}_k^n(t) + \sqrt{n} \left( \frac{\lambda^n}{n} - \lambda \right) t + \epsilon^n_k(t) \quad \text{for } k \in \{0, 1, \ldots, j-2\},
\]
and
\[
\hat{L}_{j-1}^n(t) = \hat{A}_j^n(t) - d_j \int_0^t \hat{Z}_j^n(s) dS(s),
\]
\[
\hat{Z}_j^n(t) = \hat{Z}_j^n(0) - \sqrt{n}t \left( \frac{\lambda^n}{n} - \lambda \right) - \hat{A}_j^n(t) + \hat{S}_j^n(t) + \epsilon^n_j(t) + \int_0^t \hat{Z}_j^n(s) ds,
\]
\[
\hat{Z}_{j+1}^n(t) = -\hat{Z}_j^n(t) + \epsilon^n_{j+1}(t),
\]
for some process \( (\epsilon^n_j, \epsilon^n_{j+1}) \Rightarrow 0 \) in \( \mathbb{D}^2 \) as \( n \to \infty \), and for \( \lambda \) defined in (47).

If \( j = K - 1 \), then by (14) and (64)
\[
\hat{L}_Q^n(t) + \hat{L}_{K-1}^n(t) = d_K \int_0^t \hat{Z}_K^n(s) ds + \hat{S}_K^n(t) + \epsilon^n_K(t).
\]
Hence (66) and (67) also hold in this case with \( j = K - 1 \) and \( j + 1 = K \).

From (66) and (67), it is enough to focus our attention on the convergence of \( \hat{Z}_j^n \). Define \( x^n \) by
\[
x^n(t) \equiv \hat{Z}_j^n(0) - \sqrt{n}t \left( \frac{\lambda^n}{n} - \lambda \right) - \hat{A}_j^n(t) + \hat{S}_j^n(t) + \hat{S}_{j+1}^n(t) + \epsilon^n_j(t).
\]
By (47), (62), and (63), \( x^n \Rightarrow x \) in \( \mathbb{D} \) as \( n \to \infty \), where
\[
x(t) \equiv \hat{Z}_j(0) + \lambda \beta t + B(t),
\]
for all \( t \geq 0 \) and \( B \) a driftless Brownian motion with variance \( \lambda(1 + \epsilon_0^2) \). Then \( \hat{Z}_j^n = \Phi(x^n) \) where \( \Phi \) is the continuous mapping defined in Lemma 9 of Dai et al. [7]. Thus we apply the continuous mapping theorem to conclude that \( \hat{Z}_j^n \Rightarrow \hat{Z}_j \) in \( \mathbb{D} \) as \( n \to \infty \), for \( \hat{Z}_j \) that follows (55).

We conclude the proof by showing (64). We prove only that \( \hat{Z}_0^n \Rightarrow 0 \) in \( \mathbb{D} \) as \( n \to \infty \); the convergence of \( (\hat{Q}^n, \hat{Z}_k^n \text{ for } k \in S_1 \setminus \{0\}) \) follow from a similar argument and their proofs are omitted. We follow an argument similar to that in Reiman [20]. Formally, we show that \( P\left( \|\hat{Z}_0^n(t)\|_\tau > \epsilon \right) \to 0 \) as \( n \to \infty \) for any \( \epsilon > 0 \) and for \( \tau = 1 \).
If $\hat{Z}_0^n(t) > \epsilon$ for some $\epsilon > 0$ and for some $t > 0$, then there exists $\tau_n > 0$ and $\tau'_n > 0$ such that
\[
\tau_n = \inf\{t \geq 0 : \hat{Z}_0^n(t) > \epsilon\},
\tau'_n = \sup\{t < \tau_n : \hat{Z}_0^n(t) \leq \epsilon/2\}.
\]
Note that by (53), we can assume that $\tau_n > 0$. Next, from (16) and (58),
\[
\hat{Z}_0^n(\tau_n) - \hat{Z}_0^n(\tau'_n) \leq - (\hat{A}_n(\tau_n) - \hat{A}_n(\tau'_n)) + (\hat{S}_1^n(\tau_n) - \hat{S}_1^n(\tau'_n)) - \sqrt{n}\left(\frac{\lambda_n}{n} - d_1\right)(\tau_n - \tau'_n).
\]
Hence, by definition of $\tau_n$ and $\tau'_n$, we can find $s$ and $t$ such that
\[
-(\hat{A}_n(t) - \hat{A}_n(s)) + (\hat{S}_1^n(t) - \hat{S}_1^n(s)) - \sqrt{n}\left(\frac{\lambda_n}{n} - d_1\right)(t - s) \geq \frac{\epsilon}{2},
\]
where $0 \leq s \leq t \leq 1$. Thus we have
\[
P\left(\sup_{0 \leq t \leq 1} \hat{Z}_0^n(t) > \epsilon\right) \leq P\left(\sup_{0 \leq s \leq t \leq 1} -(\hat{A}_n(t) - \hat{A}_n(s)) + (\hat{S}_1^n(t) - \hat{S}_1^n(s)) - \sqrt{n}\left(\frac{\lambda_n}{n} - d_1\right)(t - s) > \frac{\epsilon}{2}\right)
\]
\[
\leq P\left(\sup_{0 \leq s \leq t \leq 1} -(\hat{A}_n(t) - \hat{A}_n(s)) - \frac{\sqrt{n}}{2}\left(\frac{\lambda_n}{n} - d_1\right)(t - s) > \frac{\epsilon}{4}\right)
+ P\left(\sup_{0 \leq s \leq t \leq 1} (\hat{S}_1^n(t) - \hat{S}_1^n(s)) - \frac{\sqrt{n}}{2}\left(\frac{\lambda_n}{n} - d_1\right)(t - s) > \frac{\epsilon}{4}\right).
\]
All that remains is to show that for any $\eta > 0$, we can find $N$ such that for all $n > N$,
\[
P\left(\sup_{0 \leq s \leq t \leq 1} -(\hat{A}_n(t) - \hat{A}_n(s)) - \frac{\sqrt{n}}{2}\left(\frac{\lambda_n}{n} - d_1\right)(t - s) > \frac{\epsilon}{4}\right) < \eta. \tag{68}
\]
\[
P\left(\sup_{0 \leq s \leq t \leq 1} (\hat{S}_1^n(t) - \hat{S}_1^n(s)) - \frac{\sqrt{n}}{2}\left(\frac{\lambda_n}{n} - d_1\right)(t - s) > \frac{\epsilon}{4}\right) < \eta. \tag{69}
\]
We only prove (68); the proof of (69) is omitted since it follows by an identical argument.

For $x \in \mathbb{D}$, we define the mapping $(f_1^\delta, f_2^\delta)(x) : \mathbb{D} \to \mathbb{R}^2$ by
\[
f_1^\delta(x) = \sup_{0 \leq s < t < 1, |t - s| \leq \delta} \{-x(t) - x(s)\},
f_2^\delta(x) = \sup_{0 \leq s < t < 1, |t - s| > \delta} \{-x(t) - x(s)\},
\]
for $\delta > 0$. The mapping $(f_1^\delta, f_2^\delta)$ is obviously continuous, hence $(f_1^\delta, f_2^\delta)(\hat{A}_n) \Rightarrow (f_1^\delta, f_2^\delta)(\hat{A})$ in $\mathbb{R}^2$ as $n \to \infty$, by the continuous mapping theorem and (62). Since $f_1^\delta(\hat{A}) \Rightarrow 0$ as $\delta \to 0$, we can fix $\delta > 0$ such that
\[
P\left(f_1^\delta(\hat{A}) > \frac{\epsilon}{4}\right) < \frac{\eta}{4}. \tag{70}
\]
Next, note that
\[
f_2^\delta(\hat{A}) \leq 2 \sup_{0 \leq t \leq 1} \{\hat{A}(t)\}.
\]
It is well known that the supremum of a Brownian motion is stochastically bounded (see Karatzas and Shreve [13] for details), in the sense that
\[
P\left(\sup_{0 \leq t \leq 1} \hat{A}(t) > x\right) \to 0, \text{ as } x \to \infty.
\]
Thus we can find $M < \infty$ such that

$$P \left( f_2^j(\hat{A}) > \frac{\epsilon}{4} + M \right) < \frac{\eta}{4}. \quad (71)$$

Finally pick $N$ such that for $n > N,$

$$\frac{\sqrt{n}}{2} \delta (\frac{\lambda^n}{n} - d_i) > M, \quad (72)$$

$$P \left( f_i^n(\hat{A}^n) > x \right) - P \left( f_i^j(\hat{A}) > x \right) < \frac{\eta}{4}, \text{ for some } x \geq 0, \quad (73)$$

$$P \left( f_i^n(\hat{A}^n) > x \right) - P \left( f_i^j(\hat{A}) > x \right) < \frac{\eta}{4}, \text{ for some } x \geq 0. \quad (74)$$

Then for $n > N,$ by (70) and (71)-(74),

$$P \left( \sup_{0 \leq s \leq t \leq 1} (-\hat{A}^n(t) - \hat{A}^n(s)) - \frac{\sqrt{n}}{2} \left( \frac{\lambda^n}{n} - d_1 \right) (t - s) > \frac{\epsilon}{4} \right) \leq P \left( f_2^j(\hat{A}^n) > \frac{\epsilon}{4} \right) + P \left( f_2^j(\hat{A}) > \frac{\epsilon}{4} + M \right) \leq \eta,$$

completing the proof. $\Box$

5. Diffusion approximations in Case II In this section, we consider Case II where

$$\lambda^n = \lambda n - \beta \lambda \sqrt{n}, n \geq 1 \text{ and } \lambda = d_j, \quad (75)$$

for some $j \in \{1, 2, \ldots, K - 1\}$. Using the queueing equations (11)-(14), we first prove the convergence of the diffusion scaled processes, under assumption (75). The closed-form solutions for steady state of the diffusion limits could not be obtained, hence in §5.2 we give a heuristic method to estimate the steady state of the diffusion limit.

5.1. Diffusion analysis In establishing the diffusion limit we use the continuous mapping approach with a novel map that we define next. Given $x \in \mathbb{D}^2$, we wish to define a map $(\Phi, \Psi)(x) : \mathbb{D}^2 \to \mathbb{D}^3$. For $x = (x_1, x_2) \in \mathbb{D}^2$, $r = (r_1, r_2) \in \mathbb{R}^2$ with $r_1 > 0$, and $m = \{m_{ij}\}$ a square matrix of dimension $2 \times 2$, we define $(\Phi, \Psi)(x)$ to be

$$(\Phi, \Psi)(x) \equiv (z, l),$$

where $z = (z_1, z_2) \in \mathbb{D}^2$, $l \in \mathbb{D}$, and $(z, l)$ satisfies

$$z_1(t) = x_1(t) + m_{1,1} \int_0^t z_1(s)ds + m_{1,2} \int_0^t z_2(s)ds + r_1l(t), \quad (76)$$

$$z_2(t) = x_2(t) + m_{2,1} \int_0^t z_1(s)ds + m_{2,2} \int_0^t z_2(s)ds + r_2l(t), \quad (77)$$

$$l(0) = 0, \text{ } l \text{ is non-decreasing, and } z_1(t)dl(t) = 0. \quad (78)$$

The following theorem establishes the existence and the continuity of the map $(\Phi, \Psi)$.

**Theorem 5.** For each $x \in \mathbb{D}^2$, there exists a unique $(z, l) \in \mathbb{D}^3$ satisfying (76)-(78). Furthermore the map $(\Phi, \Psi)$ is continuous when the domain $\mathbb{D}^2$ and the range $\mathbb{D}^3$ are equipped with the Skorohod $J_1$ topology.

See Appendix B for a proof.

For notational simplicity we set $\mathcal{S}_2 \equiv \{0, 1, \ldots, K\} \setminus \{j - 1, j, j + 1\}$. We also assume that

$$(\hat{Q}^n(0), \hat{Z}_k^n(0) \text{ for } k \in \mathcal{S}_2) \Rightarrow 0 \text{ in } \mathbb{R}^{K-1} \text{ as } n \to \infty. \quad (79)$$

We have the following asymptotic result.
THEOREM 6. Assume that (29), (33), (75), and (79) hold. Then \( \hat{Z}_n, \hat{Q}_n \Rightarrow (\hat{Z}, \hat{Q}) \) in \( \mathbb{D}^{K+2} \) as \( n \to \infty \), where \( (\hat{Z}, \hat{Q}) \) is the unique solution to the stochastic differential equations

\[
\dot{Z}_{j-1}(t) = \dot{Z}_{j-1}(0) + \lambda \beta t + d_{j-1} \int_0^t \dot{Z}_{j-1}(s) ds + B(t) + \dot{L}_{j-1}(t) \tag{80}
\]

\[
\dot{Z}_j(t) = \dot{Z}_j(0) - \lambda \beta t - (d_{j-1} + d_{j+1}) \int_0^t \dot{Z}_{j-1}(s) ds - (d_j + d_{j+1}) \int_0^t \dot{Z}_{j+1}(s) ds - B(t) - 2 \dot{L}_{j-1}(t) \tag{81}
\]

\[
\dot{Z}_{j-1}(t) d\dot{L}_{j-1}(t) = 0 \tag{82}
\]

\[
\dot{Z}_{j+1}(t) = -\dot{Z}_{j-1}(t) - \dot{Z}_j(t), \tag{83}
\]

\[
\hat{Q}(t) = 0, \tag{84}
\]

\[
\hat{Z}_k(t) = 0 \text{ for } k \in \mathcal{S}_2, \tag{85}
\]

and \( B \) is a driftless Brownian motion with variance \( \lambda (1 + c_n^2) \).

Proof of Theorem 6: Assume that (29), (33), (75), and (79) hold. By (29), (33), and Theorem 2, \( \dot{Z}_j(t) = 1 \) for all \( t \geq 0 \). Based on Theorem 2, we define

\[
\hat{L}_k^n(t) = \sqrt{n} \left( \hat{L}_k^n(t) - \lambda t \right), \text{ for } k \in \{-1, 0, 1, \ldots, j - 2\},
\]

\[
\hat{L}_k^n(t) = \sqrt{n} \hat{L}_k^n(t) \text{ for } k \in \{j - 1, \ldots, K - 1\}.
\]

Therefore \( \hat{L}_k^n \to 0 \text{ a.s. u.o.c. as } n \to \infty \) by (16). Also

\[
\hat{Z}_{j-1}(t) d\hat{L}_{j-1}(t) = 0
\]

by (16). In addition, we have

\[
(\hat{A}_n^n, \hat{S}_j^n) \Rightarrow (\hat{A}, \hat{S}_j), \tag{86}
\]

in \( \mathbb{D}^2 \) as \( n \to \infty \), where \( \hat{A} \) and \( \hat{S}_j \) are independent driftless Brownian motions with variances \( \lambda c_n^2 \) and \( \lambda \), respectively. As in (62) and (63), this again follows from Theorem 2, Donsker’s theorem, and (29). Also, again by (29), (33), and Theorem 2 we have

\[
\hat{S}_k^n \Rightarrow 0 \tag{87}
\]

in \( \mathbb{D} \) as \( n \to \infty \) for \( k \neq j \).

From an identical argument as in the proof of (64), we conclude that

\[
(\hat{Q}_n^n, \hat{Z}_k^n \text{ for } k \in \mathcal{S}_2) \Rightarrow 0 \text{ in } \mathbb{D}^{K-1} \text{ as } n \to \infty. \tag{88}
\]

Next by (58)-(61), (87) and (88), we can find a process \( e^n \equiv (e_0^n, \ldots, e_{j-2}^n) \) such that \( e_k^n \to 0 \) in \( \mathbb{D} \) as \( n \to \infty \) for \( k \in \{0, 1, \ldots, j-2\} \), and

\[
\hat{L}_k^n(t) = \hat{A}_n^n(t) + \sqrt{n} \left( \frac{\lambda n}{n} - \lambda \right) t + e_k^n(t) \text{ for } k \in \{0, 1, \ldots, j-2\},
\]

and

\[
\hat{L}_{j-2}^n(t) = \hat{A}_n^n(t) - d_{j-1} \int_0^t \hat{Z}_{j-1}^n(s) ds + \sqrt{n} \left( \frac{\lambda n}{n} - \lambda \right) t + e_{j-2}^n(t).
\]
Then by (58)-(60) and (87), \((\hat{Z}_{j-1}^n, \hat{Z}_j^n, \hat{Z}_{j+1}^n)\) satisfy
\[
\begin{align*}
\hat{Z}_{j-1}^n(t) &= \hat{Z}_{j-1}^n(0) - \sqrt{n} \left( \frac{\lambda^n}{n} - \lambda \right) t - \hat{A}_j^n(t) + \hat{S}_j^n(t) + \epsilon_{j-1}^n(t) \\
&+ d_{j-1} \int_0^t \hat{Z}_{j-1}^n(s) ds + d_j \int_0^t \hat{Z}_j^n(s) ds + \hat{L}_{j-1}^n(t), \\
\hat{Z}_j^n(t) &= \hat{Z}_j^n(0) + \sqrt{n} \left( \frac{\lambda^n}{n} - \lambda \right) t + \hat{A}_j^n(t) - \hat{S}_j^n(t) + \hat{S}_{j+1}^n(t) + \epsilon_j^n(t) \\
&- d_{j-1} \int_0^t \hat{Z}_{j-1}^n(s) ds - d_j \int_0^t \hat{Z}_j^n(s) ds + d_{j+1} \int_0^t \hat{Z}_{j+1}^n(s) ds - 2\hat{L}_{j-1}^n(t), \\
\hat{Z}_{j+1}^n(t) &= -\hat{Z}_j^n(t) - \hat{Z}_{j-1}^n(t) + \epsilon_{j+1}^n(t),
\end{align*}
\]  
for some process \((\epsilon_{j-1}^n, \epsilon_j^n, \epsilon_{j+1}^n) \Rightarrow 0\) in \(\mathbb{D}^3\) as \(n \to \infty\).

From (65), (88), and (89)--(91), it is enough only to focus on the convergence of \((\hat{Z}_{j-1}^n, \hat{Z}_j^n)\). Let \(x^n = (x_{j-1}^n, x_j^n)\) be defined by
\[
\begin{align*}
x_{j-1}^n(t) &= \hat{Z}_{j-1}^n(0) - \sqrt{n} \left( \frac{\lambda^n}{n} - \lambda \right) t - \hat{A}_j^n(t) + \hat{S}_j^n(t) + \epsilon_{j-1}^n(t), \\
x_j^n(t) &= \hat{Z}_j^n(0) + \sqrt{n} \left( \frac{\lambda^n}{n} - \lambda \right) t + \hat{A}_j^n(t) - \hat{S}_j^n(t) + \hat{S}_{j+1}^n(t) + \epsilon_j^n(t),
\end{align*}
\]
Note that \(x^n \Rightarrow x\) in \(\mathbb{D}^2\) as \(n \to \infty\) by (75) and (86), where \(x \equiv (x_{j-1}, x_j)\) satisfies
\[
\begin{align*}
x_{j-1}(t) &= \hat{Z}_{j-1}(0) + \lambda \beta t + B(t), \\
x_j(t) &= \hat{Z}_j(0) - \lambda \beta t - B(t),
\end{align*}
\]
and \(B\) is a driftless Brownian motion with variance \(\lambda(1 + \epsilon^2_\alpha)\). Next, observe that
\[
(\hat{L}_{j-1}^n, \hat{Z}_k^n \text{ for } k \in \{j-1, j\}) = (\Psi(x^n), \Phi(x^n)),
\]
where \((\Psi, \Phi)\) are the continuous mappings defined in Theorem 5. Thus we apply the continuous mapping theorem to conclude that
\[
(\hat{L}_{j-1}^n, \hat{Z}_k^n \text{ for } k \in \{j-1, j\}) \Rightarrow (\hat{L}_{j-1}, \hat{Z}_k \text{ for } k \in \{j-1, j\}),
\]
in \(\mathbb{D}^3\) as \(n \to \infty\), for \((\hat{L}_{j-1}, \hat{Z}_k \text{ for } k \in \{j-1, j\})\) satisfying (80) and (81). \(\square\)

5.2. An approximation for the stationary distribution of the diffusion limit

Closed form solutions for the steady state of the diffusion limits established in Theorem 6 could not be obtained. In this section we offer approximations whose validity we verify using numerical results in later sections. The idea behind obtaining an approximation is to reduce the dimension of the diffusion process from 2 to 1. We achieve this by assuming that the system cannot have agents at levels \(j-1\) and \(j+1\) at the same time. This can be done either by enforcing this constraint on the diffusion process or more intuitively by slightly modifying the routing policy as we explain next.

Assume that if an agent in level \(j\) finishes service and there are agents in level \(j+1\), a customer from an agent in level \(j+1\) can be transferred to the agent finishing service. Hence both agents end up in level \(j\). It is easy to show that under such a preemptive policy there cannot be agents in levels \(j-1\) and \(j+1\) at the same time. (Such a preemption procedure is usually not applicable in the chat service applications for various reasons, see Tezcan and Zhang [24].)

To explain our approximation under the preemptive policy, let \(V^n(t)\) to be the number of people in the \(n\)th queueing system at time \(t\) and \(\hat{V}\) denote its fluid limit. It can be shown similar to
Theorem 1 that in the steady state of the fluid limit we have \( \hat{V}(t) = j \). Define \( \hat{V}^n(t) = \sqrt{n}(\hat{V}(t) - j) \). Note that by (58)-(61)

\[
\hat{V}^n(t) = \hat{V}^n(0) - \lambda \beta t - \sum_{k=1}^K d_j \int_0^t \hat{Z}_k(s)ds + \hat{A}(t) - \sum_{k=1}^K \hat{S}_j(t). \tag{92}
\]

Using an argument similar to the proof of (64) we can show under the preemptive policy that

\[
\begin{align*}
\hat{Z}^n_{j-1}(t) &= -(\hat{V}^n(t))^– + \epsilon^n_{j-1}(t), \\
\hat{Z}^n_j(t) &= (\hat{V}^n(t))^– - (\hat{V}^n(t))^+ + \epsilon^n_j(t), \\
\hat{Z}^n_{j+1}(t) &= (\hat{V}^n(t))^+ + \epsilon^n_{j+1}(t), \\
\hat{Q}^n_{j+1}(t) &= \epsilon^n_{K+1}(t), \\
\hat{Z}^n_k(t) &= \epsilon^n_k(t) \quad \text{for all } k \in S_2,
\end{align*}
\]

where \( \epsilon^n_k \to 0 \) as \( n \to \infty \) u.o.c. for \( k = 0, \ldots, K + 1 \). In words, agents can only be at levels \( j - 1, j \) and \( j + 1 \) and there cannot be any agents in levels \( j - 1 \) and \( j + 1 \) at the same time. Hence using (92) and (93)–(97) and passing to the limit we have \( \hat{V}^n \Rightarrow \hat{V} \), as \( n \to \infty \), where

\[
\hat{V}(t) = \hat{V}(0) - \lambda \beta t - (d_j - d_{j-1}) \int_0^t (\hat{V}(s))^– ds - (d_{j+1} - d_j) \int_0^t (\hat{V}(s))^+ ds + B(t), \tag{98}
\]

and \( B \) is a driftless Brownian motion with variance \( \lambda(1 + \epsilon^2_2) \). The stationary distribution of \( \hat{V} \) can be obtained in closed form and we use this to approximate the steady state of the original system (with no preemptions). We conclude the section by giving the stationary distribution of the limiting diffusion process \( \hat{V} \).

**Theorem 7.** Assume \( \hat{V} \) follows (98). Define the parameters

\[
\begin{align*}
\alpha_1 &= d_j - d_{j-1}, \quad \alpha_2 = d_{j+1} - d_j, \quad \mu_i = -\lambda \beta / \alpha_i, \quad i = 1, 2, \\
\sigma^2 &= \lambda(1 + c^2_2), \quad \sigma^2_i = \sigma^2 / 2 \alpha_i, \quad i = 1, 2, \\
c_1 &= \psi(-\mu_1 / \sigma_1), \quad c_2 = 1 - \psi(-\mu_2 / \sigma_2), \quad c_3 = \frac{1}{\sigma_1} \phi(-\mu_1 / \sigma_1), \quad c_4 = \frac{1}{\sigma_2} \phi(-\mu_2 / \sigma_2), \\
a_1 &= \frac{c_4}{c_2 c_3 + c_4}, \quad a_2 = \frac{c_3}{c_2 c_3 + c_1 c_4},
\end{align*}
\]

where \( \phi \) and \( \psi \) denote the pdf and cdf of the standard normal distribution respectively. Then the diffusion process \( \hat{V} \) admits a piecewise normal stationary distribution, with the density \( g \) given by

\[
g(z) = \begin{cases} 
\frac{1}{\sigma_1} \phi \left( \frac{z - \mu_1}{\sigma_1} \right) & \text{if } z \leq 0, \\
\frac{1}{\sigma_2} \phi \left( \frac{z - \mu_2}{\sigma_2} \right) & \text{if } z > 0.
\end{cases} \tag{99}
\]

This follows from the fact that \( \hat{V} \) is a piecewise Ornstein-Uhlenbeck process, which has a piecewise normal stationary distribution (see Browne and Whitt [4] for details).

6. **Diffusion approximations in Case III** In this section we focus on Case III where

\[
\lambda^n = \lambda n - \beta \lambda \sqrt{n}, \quad n \geq 1 \quad \text{and} \quad \lambda = d_K,
\]

and establish the diffusion limits of the chat systems. We need to make the following assumptions about the initial state of the system, similar to the corresponding assumptions (53) and (79) in Cases I and II. Let \( S_3 = \{0, 1, \ldots, K\} \setminus \{K - 1, K\} \) and assume that

\[
\hat{Z}^n_k(0) \Rightarrow 0 \quad \text{in } \mathbb{R}^K \quad \text{as } n \to \infty, \quad \text{for } k \in S_3.
\]

We have the following result for \( (\hat{Z}^n, \hat{X}^n) \) (recall definition (32)).
THEOREM 8. Assume that (29), (33), (100), and (101) hold. Then \((\hat{Z}^n, \hat{X}^n) \Rightarrow (\tilde{Z}, \tilde{X})\) in \(\mathbb{D}^{K+2}\) as \(n \to \infty\), where \((\tilde{Z}, \tilde{X})\) is the unique solution to the stochastic differential equations

\[
\begin{align*}
\hat{Z}_k(t) &= 0 \quad \text{for} \quad k \in S_3, \\
\hat{Z}_{K-1}(t) &= -\hat{X}^-(t), \\
\hat{X}(t) &= \hat{X}(0) - \lambda \beta t - (d_K - d_{K-1}) \int_0^t \hat{X}^-(s) ds - \nu \int_0^t \hat{X}^+(s) ds - B(t),
\end{align*}
\]

where \(B\) is a driftless Brownian motion with variance \(\lambda(1 + \epsilon_a^2)\).

Before we present the proof, we next provide the stationary distribution of the limiting process. Note that \(\hat{X}\) is a diffusion process similar to (98) with different coefficients. Hence the stationary distribution of \(\hat{X}\) can be found as in Theorem 7 by setting \(\alpha_1 = d_K - d_{K-1}\), and \(\alpha_2 = \nu\).

Proof of Theorem 8: Assume that (29), (33), (100), and (101) hold. By (29), (33), and Theorem 2, \(\hat{X}(t) = 1\) for all \(t \geq 0\). Based on Theorem 2, we define

\[
\hat{L}_k^n(t) = \sqrt{n} \left( \hat{L}_k^n(t) - \lambda t \right), \quad \text{for} \quad k \in \{-1, 0, 1, \ldots, K-1\}.
\]

As in the proof of (62), we apply Donsker’s theorem, Theorem 2, and (29) to conclude that

\[
(\hat{A}^n, \hat{S}_K^n) \Rightarrow (\tilde{A}, \tilde{S}_K),
\]

in \(\mathbb{D}^2\) as \(n \to \infty\), where \(\hat{A}\) and \(\hat{S}_K\) are independent driftless Brownian motions with variances \(\lambda \epsilon_a^2\) and \(\lambda\), respectively. Also, again by (29), (33), and Theorem 2 we have

\[
\hat{S}_k^n \Rightarrow 0
\]
in \(\mathbb{D}\) as \(n \to \infty\) for \(k = 1, \ldots, K-1\) and \(k = K+1\).

Next, from an identical argument as in the proof of (64), we have that

\[
\hat{Z}_k^n \Rightarrow 0 \quad \text{for} \quad k \in S_3.
\]

Observe that by (58)-(61) and (106) there exists some process \(\epsilon^n \equiv (\epsilon_0^n, \ldots, \epsilon_{K-2}^n)\) such that \(\epsilon^n \Rightarrow 0\) as \(n \to \infty\) in \(\mathbb{D}^{K-1}\) and such that

\[
\hat{L}_k^n(t) = \hat{A}^n(t) + \sqrt{n} \left( \frac{\lambda_n}{n} - \lambda \right) t + \epsilon_k^n(t) \quad \text{for} \quad k \in \{0, 1, \ldots, K-3\},
\]

and

\[
\hat{L}_{K-2}^n(t) = \hat{A}^n(t) - d_{K-1} \int_0^t \hat{Z}_{K-1}(s) ds + \sqrt{n} \left( \frac{\lambda_n}{n} - \lambda \right) t + \epsilon_{K-2}^n(t).
\]

By (32),

\[
\hat{X}^n(t) = \frac{X^n(t) - n}{\sqrt{n}}.
\]

Then by (17), (107), and (108)

\[
\hat{X}^n(t) = \hat{X}^n(0) + \sqrt{n} t \left( \frac{\lambda_n}{n} - \lambda \right) - d_{K-1} \int_0^t \hat{Z}_{K-1}(s) ds - d_K \int_0^t \hat{Z}_K(s) ds - \nu \int_0^t \hat{Q}^n(s) ds + \hat{A}^n(t) - \hat{S}_K(t) + \epsilon_K^n(t),
\]
for some process \( \epsilon_k^n \to 0 \) in \( \mathbb{D} \) as \( n \to \infty \). By definition of \( \hat{X}^n \), we have \( \hat{Q}^n = (\hat{X}^n)^+ \), \( \hat{Z}_K^n = (\hat{X}^n)^- \), and thus from (7), (65) and (106) we have \( \hat{Z}_{K-1}^n = -(\hat{X}^n)^- + \epsilon_{K-1}^n(t) \), for some \( \epsilon_{K-1}^n \to 0 \) in \( \mathbb{D} \) as \( n \to \infty \). Therefore, from (106), it suffices to focus only on the convergence of \( \hat{X}^n \). Define \( x^n \) by

\[
x^n(t) = \hat{X}^n(0) + \sqrt{nt} \left( \frac{\lambda^n}{n} - \lambda \right) + \hat{A}^n(t) - \hat{S}_K^n(t) + \epsilon_K^n(t).
\]

Note that \( x^n \Rightarrow x \) in \( \mathbb{D} \) by Donsker’s theorem, (100), and (105), where

\[
x(t) = \hat{X}(0) - \lambda \beta t - B(t),
\]

and \( B \) is a driftless Brownian motion with variance \( \lambda(1 + c_2^2) \). Observe that \( \hat{X}^n = \Phi(x^n) \) where \( \Phi \) is the continuous mapping defined in Theorem 4.1 of Pang et al. [17]. Because \( \Phi \) is continuous, we apply the continuous mapping theorem to conclude that \( \hat{X}^n \) converges weakly to \( \hat{X} \), which satisfies (104). \( \square \)

7. Approximations for performance measures in steady state In this section we build approximations for three performance metrics in chat systems; number in system, abandonment probability and sojourn time in steady state, based on the diffusion limits we established. Our goal is to build approximations for the performance measures given fixed arrival rate \( \lambda^n \) and number of servers \( n \). The approximations we obtain for the number in system are somewhat standard and are obtained by “reversing” the diffusion scaling in (30)–(32) and using the stationary distribution of the diffusion limits. Once the approximations for the number in system are obtained, it is also not very difficult to obtain approximations for the abandonment probability. However, obtaining approximations for the sojourn times requires more work as explained in the introduction. In §7.3 we offer two approximations based on our estimates of number of agents in each level in steady state.

7.1. Number in system First we focus on estimating the number in system. Fix \( \lambda^n \) and \( n \) and let \( Z_j^n(\infty) \) and \( V_j^n(\infty) \) denote the number of customers in level \( j \) and in system in the \( n \)-server system in steady state, respectively. Also let \( Q^n(\infty) \) denote the number of customers in queue in steady state. We consider two estimation procedures, referred to as Procedures 1 and 2. Procedure 1 is derived from Theorem 4 and Procedure 2 from Theorem 7. Given \( f = (d_1, \ldots, d_K) \) and \( d_1 \leq \lambda \leq d_K \), we define \( j(\lambda, d) \) and \( \gamma(\lambda, d) \) to be the unique values of \( j \) and \( \gamma \) that satisfy (47), where \( \gamma \) is restricted to be between 0 and 1. For notational simplicity set \( j = j(\lambda^n, d) \) for the rest of this section.

First we describe Procedure 1. Let \( \lambda \equiv \lambda^n/n \). Then the first equation in (52) holds with \( \beta = 0 \) and let \( \gamma(\lambda, d) \) be defined as above (we drop \( \lambda, d \) for notational simplicity). From Theorem 4, by taking \( \beta = 0 \), an approximation for the steady state is given by \( E [Z_j^n(\infty)] = \gamma n \) and \( E [Z_{j+1}^n(\infty)] = (1 - \gamma)n \) and \( E Z_k^n(\infty) = 0 \) for all \( k \in \mathcal{S}_1 \) and \( E Q^n(\infty) = 0 \). For the variance we have \( Var (Z_j^n(\infty)) = Var (Z_{j+1}^n(\infty)) = \frac{\lambda^n(1 + c_2^2)}{2(d_{j+1} - d_j)} \) and \( Var (Z_j^n(\infty)) = 0 \) for all \( k \in \mathcal{S}_1 \) and \( Var (Q^n(\infty)) = 0 \). Hence we have the following approximations in Procedure 1.

\[
E[V^n(\infty)] = j\gamma n + (j + 1)(1 - \gamma)n
\]

and

\[
Var (V^n(\infty)) = \frac{\lambda^n(1 + c_2^2)}{2(d_{j+1} - d_j)}.
\]

We note in passing that the approximations obtained for the expected number of customers, as well as for \( E [Z_j^n(\infty)] \) and \( E [Z_{j+1}^n(\infty)] \), under Procedure 1 agree with the corresponding approximations obtained from the fluid limits in [25].
Next we explain the details of Procedure 2. Let $j$ be such that $nd_j$ is closest to $\lambda^n$ and let $\lambda \equiv d_j$. For the rest of this section we assume that $j < K$ but if $j = K$ the approximations can be obtained from Theorem 8 in a similar way. From (18) and (75), the value of $\beta$ for a fixed $\lambda$ and $d$ is given by

$$\beta = \sqrt{n} \left( 1 - \frac{\lambda^n}{nd_j} \right).$$

Let $V(\infty)$ denote a random variable with the stationary distribution $g$ defined in (99). We use the approximation

$$E[V^n(\infty)] = \sqrt{n} E[V(\infty)] + jn.$$

Similarly, the variance of the number of customers in steady state, $Var(V^n(\infty))$, is calculated by

$$Var(V^n(\infty)) = n Var(V(\infty)).$$

Approximations for $Z^n_j$’s can be obtained using (93)–(95).

To illustrate which procedure should work better in a given parameter setting, we note that in Case I, the arrival rate is assumed to be in between two $d_j$’s and in Case II the assumption is that the arrival rate is fairly “close” to $d_j$. Since we take $\beta = 0$ in (18) in Procedure 1, we expect that when $\gamma(\lambda, d)$ used in Procedure 1 is far away from the boundaries, 0 and 1, Procedure 1 should perform better. When it is close to 0 or 1 Procedure 2 should perform better.

### 7.2. Abandonment probability

We wish to estimate two quantities: the probability that an arriving customer abandons from service, $p_s$, and from the queue $p_q$, in steady state. The approximations are based on the fact that the patience times are exponential. We have

$$p_q = \frac{\nu E[Q^n(\infty)]}{\lambda^n},$$

$$p_s = \frac{\sum_{k=1}^{K} k \nu_s E[Z^n_k(\infty)]}{\lambda^n}.$$

Equation (109) follows from the fact that per unit time there will be $\nu E[Q^n(\infty)]$ abandonments from the queue out of $\lambda^n$ arriving customers. The expression (110) follows from an identical reasoning.

### 7.3. Sojourn times

Now we focus on the sojourn time. In particular, we are interested in a method to find higher moments of the sojourn time since the expected sojourn time can be obtained from a simple application of Little’s Law and our estimate for $E[V^n(\infty)]$. The approximations for the distribution of the sojourn time is a little bit more involved. For the sake of brevity we will mainly focus on the case when all the agents are expected to be in two levels $j$ and $j + 1$, for example see the approximation in Case I above. The approximations for the other cases can be obtained similarly. We illustrate the accuracy of our approximations suggested in this section in §8.3. Our approximations to follow are based on estimated $E[Z^n_j(\infty)]$’s in steady state, hence any method to estimate $E[Z^n_j(\infty)]$’s can be used. To this end let $E[Z^n_j(\infty)]$ and $E[Z^n_{j+1}(\infty)]$ denote the expected number of agents in level $j$ and $j + 1$ in steady state.

**Method 1:** The main idea behind the first method is to assume that the number of agents in each level at all times is equal to their corresponding expected number in steady state. To model the journey of a customer in the system we define a continuous-time Markov chain with states $\{j, j + 1, out\}$ and transition rates between two states $m$ and $n$ denoted by $\theta_{m,n}$, where

$$\theta_{j,j+1} = \frac{\lambda^n - d_j E[Z^n_j(\infty)]}{E[Z^n_j(\infty)]}.$$
A diagram of the Markov chain can be found in Figure 1. The intuition behind the Markov chain is as follows: We “tag” a customer that enters the system at time 0 and begins service with an agent that is either at level \( j \) or level \( j + 1 \). Then we model the customer’s sojourn in the system as a continuous time Markov chain which is in state \( j \) if the tagged customer’s agent is at level \( j \), state \( j + 1 \) if the tagged customer’s agent is at level \( j + 1 \), and in state “out” if the customer has completed service by time \( t \). (Obviously “out” is an absorbing state.)

We assume that customers are routed randomly to an agent that is either at level \( j \) or level \( j + 1 \) at arrival. In particular, we assume that the probability that a customer begins service in class \( j \) is \( \frac{d_j}{\lambda^n} \) and the probability that a customer begins service in class \( j + 1 \) is \( \frac{d_{j+1}}{\lambda^n} \). This assumption ensures that the arrival rates into class \( j \) and class \( j + 1 \) are equal to the total departure rates from class \( j \) and class \( j + 1 \) respectively. The customer’s sojourn time is given by the first passage time from entry into the system to the state “out”.

The rate \( \theta_{j,\text{out}} \) is as given in (112) since an agent at level \( j \) completes service at rate \( d_j \), and with probability \( 1/j \) the customer who completes service is the tagged customer, because service times are exponential. A similar argument gives us the expressions for \( \theta_{j+1,j} \) and \( \theta_{j+1,\text{out}} \). Finally, \( \theta_{j,j+1} \) is defined as in (111) since the arrival rate into level \( j \) is given by \( \lambda^n - d_j E[Z^n_j(\infty)] \), and with probability \( 1/E[Z^n_j(\infty)] \) the arrival is routed to the agent that is serving the tagged customer (assuming that an agent in that level is selected randomly for a newly arriving customer). Closed-form solutions for the higher moments of the sojourn time cannot be obtained in general. However it can easily estimated by simulating the described Markov chain.

**Method 2:** We now provide another approximation that would yield a simpler approximation than the first method and yield a closed-form expression of the sojourn time. The intuition behind our next method is as follows: We wish to divide our chat system into two separate systems in such a way that the steady state is similar to the original system, and such that it is possible to analyze each system separately. A similar approximation is used in Tezcan and Zhang [25] based on fluid approximations.

We divide our chat system into two separate systems, where the first system (denoted as \( S^j \)) consists only of agents that are allowed to serve at most \( j \) customers, and the second system (denoted as \( S^{j+1} \)) only of agents that are allowed to serve at most \( j + 1 \) customers. We assume that there are \( E[Z^n_j(\infty)] \) agents in \( S^j \), and \( E[Z^n_{j+1}(\infty)] \) agents in \( S^{j+1} \). We also assume that arrivals are randomly routed at the time of arrival into \( S^j \) with probability \( p \) or \( S^{j+1} \) with probability \( 1 - p \), and we set \( p = \frac{d_j Z^n_j(\infty)}{\lambda^n} \). By Theorem 2, in steady state there are \( E[Z^n_j(\infty)] \) agents at level
j and $E[Z_j^{n+1}(\infty)]$ agents at level $j + 1$ in this new system. By dividing our original system into $S^j$ and $S^j+1$ and by selecting our parameters carefully, the steady-state number of agents in the new system at each level matches the number of agents predicted by Theorem 2 for the original system. Therefore, we expect that this method yields accurate approximations.

Define $Y_j$ and $Y_{j+1}$ to be independent exponential random variables with rates $(\mu_j + \nu_s)$ and $(\mu_{j+1} + \nu_s)$ respectively, representing the sojourn times experienced by a customer in $S^j$ and $S^{j+1}$, and let $X$ be a Bernoulli random variable, independent from $Y_j$ and $Y_{j+1}$, indicating whether the customer enters $S^j$ or $S^{j+1}$, that is,

$$X = \begin{cases} 1 & \text{with probability } \frac{d_j E[Z_j^n(\infty)]}{\lambda^n}, \\ 0 & \text{with probability } \frac{d_{j+1} E[Z_{j+1}^n(\infty)]}{\lambda^n}. \end{cases}$$

The sojourn time of a customer (denoted by $T$) is then given by

$$T = XY_j + (1 - X)Y_{j+1}.$$

We use $T$ to approximate the sojourn time of a customer in the original system. By conditioning on the value of $X$, we can find the mean and variance of $T$. (Recall that $p = \frac{d_j E[Z_j^n(\infty)]}{\lambda^n}$.) Then

$$E[T] = \frac{p}{\mu_j} + \frac{1 - p}{\mu_{j+1}},$$
$$Var(T) = \frac{(1 - p^2)(\mu_j^2 - 2(1 - p)p\mu_j\mu_{j+1} + (2 - p)p\mu_{j+1}^2)}{\mu_j^2}\mu_{j+1}^2.$$

8. Numerical Results In this section we test the approximations offered in the previous section for the number of customers in system, abandonment probability and the sojourn time using numerical experiments. We consider systems with 25 to 100 agents with varying arrival rates. We wish to investigate two issues in our numerical tests: (i) determine the effect of the number of agents on the quality of our approximations, (ii) determine when it is more appropriate to use either Procedure I or II, based on different parameters. We focus mainly on approximations for Cases I and II. Case III can be considered as a special case of Case II (or Procedure II).

8.1. Details of simulation experiments To perform our numerical tests, we use 3 different staffing levels: $n = 100$, $n = 50$, $n = 25$, and 2 sets of parameters for $d$: case (i), $d^{(1)} = \{3, 5, 6, 7, 6\}$, case (ii), $d^{(2)} = \{3, 5, 7, 8, 5\}$. We set $\nu = \nu_s = 0.2$. For each combination of $n$ and $d$ (for a total of 6 scenarios), we simulate the system using arrival rates denoted by $\lambda^n$, where $\lambda^n = n\hat{\lambda}$. In case (i) we take $\hat{\lambda} \in \{5.01, 5.2, 5.4, 5.5, 5.6, 5.8, 5.99\}$, and in case (ii) we take $\hat{\lambda} \in \{5.01, 5.5, 6, 6.25, 6.5, 7, 7.49\}$. In all the experiments we use Poisson arrivals. From our choice of parameters, we note that $j(\lambda, d) = 2$ for all combinations of $\hat{\lambda}$ and $d$. We choose the values of $d$ in cases (i) and (ii) to investigate the impact of increasing $d_{j+1} - d_j$ on the accuracy of our approximations, and we choose the values of $\hat{\lambda}$ to test how the accuracy of our approximations are affected by different values of $\gamma(\hat{\lambda}, d)$. In both cases, $\gamma(\hat{\lambda}, d)$ varies from 0 to 1.

All simulations were run for 20000 time units with a 2000 time unit warmup period. We do not report the confidence intervals of the simulated number in system (obtained from batch sampling), since they are all less than 1% of the average value in general. To test the accuracy of our approximations proposed in §7, we compare the simulated mean and variance of the number in system with the predicted values and report the percentage difference between the two. For each performance measure, we present

$$\% \text{ error} = \frac{\text{predicted value} - \text{simulated value}}{\text{simulated value}}.$$
8.2. Number in system  The results of our numerical results for the number of customers in system are reported in Figure 2 to Figure 5. We present the results for the expected value in Figures 2 and 3 and the variance of the number of customers in system in Figures 4 and 5. For each value of \( n \) and \( d \), we plot the percentage error defined in (117) over different arrival rates (see Figure 2 - Figure 5). From Figure 2 and Figure 3, we notice that both procedures perform well in estimating the \textit{average} number of customers in the system, with a maximum average error of around 2%.

In Table 1 (which is intended to serve as a general summary of all the simulation results in this section), for each set of values of \( d \), we find the minimum absolute error between Procedure 1 and Procedure 2 for each \( \gamma(\lambda, d) \), then take the average over all the arrival rates and present this average error over the size of the chat system. This gives an idea of how our best approximation performs relative to system size.

As expected, Procedure 1 outperforms Procedure 2 in predicting the \textit{variance} of the number in system for moderate values of \( \gamma(\lambda, d) \), but when \( \gamma(\lambda, d) \) is more extreme, Procedure 2 outperforms Procedure 1 (see Figure 4 and Figure 5). Thus, we use Procedure 1 to calculate the variance in the former case and Procedure 2 in the latter case. Using this rule, our approximations perform well in predicting the variance. At most, the average error is 10.5% (see Table 1). The maximum percentage error is 38% (see Figure 5(a)) for a small number of agents, although for a larger system the maximum percentage error drops to 7% (see Figure 5(c)). Finally, as a general rule of thumb, we recommend using Procedure 1 when \( 0.3 \leq \gamma(\lambda, d) \leq 0.7 \) under Procedure 1, and recommend using Procedure 2 otherwise. Because the approximations under Procedure 1 for the expected number of customers in system agree with that of the fluid approximations, our numerical results also show that approximations based on diffusion limits are more accurate than those based on fluid limits when \( \gamma(\lambda, d) \sim 0 \) or 1.
While both procedures usually work well in estimating system performance, an exception occurs in the extreme case when $\tilde{\lambda}$ is small, and $n$ is small. If $\tilde{\lambda}$ and $d_{j+2} - d_{j+1}$ is small, the variance becomes extremely large and the rate at which our approximations converge to the true variance becomes quite slow, making the approximations inaccurate.
for smaller systems. To illustrate this observation, we plot the error in variance of the number in system, for the chat system with parameters $d^{(3)} = \{3, 5, 7.5, 7.6\}$, $\lambda \in \{5.01, 5.5, 6, 6.25, 6.5, 7, 7.49\}$, and with staffing levels $n = 100$, $n = 50$, and $n = 25$, for a total of three different scenarios. We note that for all values of $\lambda$, $j(\lambda, d^{(3)}) = 2$. Our results are given in Figure 6. Focusing on the case when $\lambda = 7.49$, we see that for $n = 25$ the error for Procedure 2 is around 70% (see Figure 6(a)), but as $n = 100$, the error drops to 10% (see Figure 6(c)).

![Figure 6. % error in variance of number in system for $d = d_3$.](image)

8.3. Sojourn time In this section we test the two proposed methods in §7.3 for estimating the sojourn time. In both methods we use the estimates of $E[Z^n_0(\infty)]$ and $E[Z^n_{j+1}(\infty)]$ obtained using Procedure 1 in §7.1. To determine the accuracy of our estimates, we focus only on the variance of the sojourn time since the expected sojourn time is easily determined from Little’s Law and from our approximations for the average number of customers in system (whose accuracies are already verified). To estimate the variance of the sojourn times with method 1, for each fixed value of $\lambda$ and $d$, we simulate the Markov chain defined in Figure 1 for 100,000 iterations. The expected sojourn time is given by the average first passage time to state ‘out’, and its variance is determined from these samples. In method 2, we calculate the variance of the sojourn time numerically from (116) for each fixed value of $\lambda$ and $d$.

The results are reported in Figure 7 and Figure 8. For each value of $n$ and $d$ introduced in §8.2, we plot the percentage error in the variance of the sojourn time, defined by (117) over different arrival rates. In Table 1, for each set of parameters of $d$, we find the minimum absolute error between Methods 1 and 2 for each $\lambda$, then take the average of all the errors over all arrival rates and present this average error versus the size of the chat system.

Overall, both methods perform well in predicting the variance of the sojourn time. Thus, in general we remark that both methods are equally as appropriate to use to analyze the sojourn time in all cases. From Table 1, we see that the average error is at most around 5% for $n = 25$, while the maximum error in variance is around 15% when $n = 25$ (see Figure 7(a)). When $n = 100$, the average error drops below 3%. Finally as a sidenote, we observe that method 1 outperforms method 2 as the quantity $d_{j+1} - d_j$ increases (see, for example, Figure 7(c) and Figure 8(c)). This is mainly because method 1 accounts for the fact that the tagged customer’s agent may change levels during the customer’s service time, while method 2 fails to account for this fact. Thus, as the value $d_{j+1} - d_j$ increases, method 2 becomes more inaccurate.
8.4. Abandonment probability  In this section we test the accuracies of our approximations for the abandonment probability. Once the number of customers in different levels is estimated accurately, we expect very little error in approximating abandonment probability, since (109) and (110) hold for chat systems in steady state and our diffusion approximations for the total number of customers in system are quite accurate. We next report the results when the abandonment probabilities are estimated using (110) where the expected number of agents in each level \( E[Z^n_j(\infty)] \) and \( E[Z^n_{j+1}(\infty)] \) is estimated using Procedure 2. The results when Procedure 1 is used is very similar.

We plot the percentage error between our predicted abandonment probability and the actual abandonment probability against the arrival rate in Figure 9 and Figure 10. Overall, our approximations for the abandonment probability is very accurate. In most cases the error is well below 1%, even for as few as 25 agents, while the maximum error is less than 3% (see Figure 9(c)). Finally, to give an idea of the overall accuracy, we show the average absolute percentage error for different staffing levels in Table 1, see row “Abandonment probability”. In almost all cases the average error is well below 1%.

<table>
<thead>
<tr>
<th>Performance measure</th>
<th>( d^{(1)} )</th>
<th>( d^{(2)} )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( n=25 )</td>
<td>( n=50 )</td>
</tr>
<tr>
<td>Mean of number in system</td>
<td>1.06 0.53 0.55</td>
<td>0.38 0.47 0.21</td>
</tr>
<tr>
<td>Variance of number in system</td>
<td>4.3 4.62 3.33</td>
<td>10.5 4.38 1.74</td>
</tr>
<tr>
<td>Sojourn time variance</td>
<td>5.31 3.89 2.9</td>
<td>3.48 2.47 1.51</td>
</tr>
<tr>
<td>Abandonment probability</td>
<td>1.01 0.43 0.42</td>
<td>0.709 0.43 0.26</td>
</tr>
</tbody>
</table>

Table 1. Average absolute percentage error
9. Conclusion This paper derives diffusion approximations to estimate the performance of chat service systems with impatient customers in steady state. We focus on the lightest-load-first routing policy. We establish the diffusion limits describing the asymptotic behavior of the diffusion-scaled number of customers in the system in a many-server heavy traffic regime. Under three different cases we obtain different limits and propose approximations for the number of customers in the system and the abandonment probability in steady state based on these limits. In addition, we give two heuristic methods to estimate the sojourn time of a customer based on our diffusion approximations. We justify the validity of our approximations with numerical experiments. Our numerical experiments suggest that in most cases our approximations work well, even for small systems with 25 agents.

Moving forward, we believe there are several important future research directions. First it may be possible to extend our CSC model to incorporate general or phase-type service time distributions and obtain estimates on system performance. On the technical side, other routing policies proposed in Tezcan and Zhang [25] can also be studied. We study a special case in Appendix C. Another interesting extension of our model is to introduce heterogeneity in either the customers or in the chat agents. It will be useful to determine how to effectively route customers to available agents and to determine the resulting system performance under such a policy if customers have different service requirements and chat agents can serve customers at different rates.

References


Appendix

A. Proof of Theorem 1 Assume that (18) and (28) hold. We first show that every fluid model solution \((Z, Q, \bar{A}, \bar{L}, \bar{L}_Q)\) is differentiable almost everywhere, by proving it is Lipschitz. We observe that \(\bar{A}\) and \(d_k \int_0^t \bar{Z}_k(s) ds\) are Lipschitz for all \(k \in \{0, 1, \ldots, K\}\), where the latter statement follows from the fact that \(0 \leq \bar{Z}_k(t) \leq 1\) for all \(k \in \{0, 1, \ldots, K\}\) and \(t \geq 0\). The process \(\bar{A}_k\) is also Lipschitz, since it is increasing and bounded above in any interval \([t_1, t_2]\) by \(\bar{A}(t_2) - \bar{A}(t_1)\). Finally, \(\bar{L}\) and \(\bar{L}_Q\) are Lipschitz by (43) and (42), respectively. Since summation preserves the Lipschitz property, we conclude that \((\bar{Z}, \bar{Q})\) is Lipschitz.

Next we prove tightness. We first note that by the functional strong law of large numbers (FSLLN), \(\bar{A}^n \to \bar{A}\) u.o.c. a.s. as \(n \to \infty\), where \(\bar{A}(t) \equiv \lambda t\). For the rest of the proof we consider only the sample paths in which this convergence holds, and denote the collection of all such sample paths as \(\Omega'\). Next we show that \(\{\bar{Z}^n, \bar{Q}^n, \bar{A}^n, \bar{L}^n, \bar{L}_Q^n\}\) is tight. Since \(\bar{A}^n \to \bar{A}\) u.o.c. a.s. as \(n \to \infty\), we conclude that \(\bar{A}^n\) is tight. Next, observe that for \(0 < t_1 < t_2\), we can bound \(\bar{A}^n_k\) by

\[
\bar{A}^n(t_2) - \bar{A}^n(t_1) \geq \bar{A}^n_k(t_2) - \bar{A}^n_k(t_1) \geq 0, \text{ for all } k \in \{0, 1, \ldots, K\}.
\]

Hence \(\bar{A}^n\) is tight and there exists a subsequence along which \(\bar{A}^n\) converges u.o.c. to some continuous function \(\bar{A}\). By a similar argument we conclude that \((\bar{D}^n, \bar{L}^n, \bar{L}_Q^n)\) is tight as well, hence \((\bar{Z}^n, \bar{Q}^n)\) is tight.

We next show that every fluid limit satisfies (34)-(46). For the remainder of the proof we fix a sample path \(\omega \in \Omega'\), and choose a subsequence, again denoted by \(\{n\}\), such that \((\bar{Z}^n, \bar{Q}^n, \bar{A}^n, \bar{L}^n, \bar{L}_Q^n) \to (\bar{Z}, \bar{Q}, \bar{A}, \bar{L}, \bar{L}_Q)\) u.o.c. Under fluid scaling defined in (19)-(27), equations (1)-(8) become

\[
\begin{align*}
\bar{Z}^n_0(t) &= \bar{Z}^n_0(0) + \bar{S}^n_1(t) - \bar{A}^n_0(t) + d_1 \int_0^t \bar{Z}^n_1(s) ds, \\
\bar{Z}^n_k(t) &= \bar{Z}^n_k(0) + \bar{S}^n_{k+1}(t) - \bar{A}^n_k(t) + \bar{A}^n_{k-1}(t) - d_k \int_0^t \bar{Z}^n_k(s) ds \\
&\quad + d_{k+1} \int_0^t \bar{Z}^n_{k+1}(s) ds, \text{ for } k \in \{1, 2, \ldots, K-2\}, \\
\bar{Z}^n_{K-1}(t) &= \bar{Z}^n_{K-1}(0) + \bar{S}^n_{K-1}(t) - \bar{A}^n_{K-1}(t) + \bar{A}^n_{K-2}(t) - d_{K-1} \int_0^t \bar{Z}^n_{K-1}(s) ds + \bar{L}^n_Q(t), \\
\bar{Q}^n(t) &= \bar{Q}^n(0) - \bar{S}^n_K(t) - \bar{S}^n_{K+1}(t) + \bar{A}^n_K(t) - d_K \int_0^t \bar{Z}^n_K(s) ds - \nu \int_0^t \bar{Q}^n(s) ds + \bar{L}^n_Q(t), \\
\bar{X}^n(t) &= \bar{X}^n(0) + \bar{A}^n_{K-1}(t) + \bar{A}^n_K(t) - d_K \int_0^t \bar{Z}^n_K(s) ds - \nu \int_0^t \bar{Q}^n(s) ds - \bar{S}^n_K(t) - \bar{S}^n_{K+1}(t),
\end{align*}
\]

where

\[
\bar{L}^n_Q(t) = \int_0^t 1_{\{\bar{Q}^n(s) = 0\}} d\bar{D}^n_K(s).
\]

Also, by (16)

\[
\begin{align*}
\bar{Q}^n(t) \bar{Z}^n_k(t) &= 0, \text{ for } k \in \{0, 1, \ldots, K-1\}, \\
\int_0^t \bar{Z}^n_k(s) d\bar{A}^n_j(s) &= 0, \text{ for } k = 0, 1, \ldots, K \text{ and } j \geq k+1.
\end{align*}
\]
For any \(\epsilon > 0, T > 0\), (because of the assumed convergence and by FSLLN) there exists \(n\) large such that
\[
\left\| (\tilde{Z}^n, \tilde{Q}^n, \tilde{A}^n, \tilde{L}^n, \tilde{L}_Q^n)(t) - (\bar{Z}, \bar{Q}, \bar{A}, \bar{L}, \bar{L}_Q)(t) \right\|_T < \epsilon, \tag{121}
\]
\[
\left\| \tilde{S}^n_k(t) \right\|_T < \epsilon, \quad \text{for all } k \in \{1, 2, \ldots, K\}. \tag{122}
\]

Next, we show that \((\tilde{Z}, \tilde{Q}, \tilde{A}, \tilde{L}, \tilde{L}_Q)\) satisfies (36). Equations (34), (35), (37), and (38) follow by an identical argument and their proofs will be omitted. Define \(R\) to be the right side of (36). Then for \(T > 0\) and for \(n\) large, by (3)
\[
\left\| \tilde{Z}_{K-1}(t) - R(t) \right\|_T \leq \epsilon + \left\| \tilde{Z}^n_{K-1}(t) - R(t) \right\|_T \leq 2\epsilon + |\tilde{Z}_{K-1}(0) - \tilde{Z}^n_{K-1}(0)| + \left\| A_{K-1}(t) - A^n_{K-1}(t) \right\|_T + \left\| \bar{A}_{K-2}(t) - \bar{A}^n_{K-2}(t) \right\|_T \\
+ \left\| d_{K-1} \int_0^t \tilde{Z}_{K-1}(s)ds - d_{K-1} \int_0^t \tilde{Z}^n_{K-1}(s)ds \right\|_T + \left\| \bar{L}_Q(t) - \bar{L}^n_Q(t) \right\|_T + \left\| \tilde{S}^n_{K-1}(t) \right\|_T \\
\leq 7\epsilon,
\]
where the last inequality follows from (121) and (122).

We next show that \((\tilde{Z}, \tilde{Q}, \tilde{A}, \tilde{L}, \tilde{L}_Q)\) satisfies (40). Similarly (39) follows from (41) and (119). Fix \(k\) and \(j\) \(k + 1\), and assume (121) holds. Then for some \(T > 0\) if \(\tilde{Z}_k(t) > 0\) then \(\tilde{Z}_k(s) > 0\) for \(|s - t| < \delta\) and for some \(\delta > 0\), by continuity. Then by (121), \(\tilde{Z}_k^*(s) > 0\) for \(|s - t| < \delta\). By (120), \(\tilde{A}_j(t + \delta) - \tilde{A}_j(t - \delta) = 0\), and therefore by (121) again, \(\tilde{A}_j(t + \delta) - \tilde{A}_j(t - \delta) < \epsilon\) giving the desired result. (42) follows from (5) and (43)-(46) follow from the definition of the fluid scaling. Q.E.D.

**B. Proof of Theorem 5** We use an approach similar to that in Lemma 9 in Dai et al. [7]. We begin by assuming that at least one solution \((z, l)\) to (76)-(78) exists for any \(x \in \mathbb{D}^2\), and we first prove uniqueness of the solution. For \(x \in \mathbb{D}^2\), let \((z, l)\) be any solution such that \((z, l)\) satisfies (76)-(78), and similarly for \(\bar{x} \in \mathbb{D}^2\), let \((\bar{z}, \bar{l})\) be any solution to (76)-(78) associated with \(\bar{x}\). Next observe that for any \(a, b \in \mathbb{D}\), we have
\[
\left| \sup_{0 \leq s \leq t} a(s)^+ - \sup_{0 \leq s \leq t} b(s)^+ \right| \leq 2 \sup_{0 \leq s \leq t} |a(s)^+ - b(s)^+| \leq 2 \sup_{0 \leq s \leq t} |a(s) - b(s)|. \tag{123}
\]

Then for any \(t \in [0, T]\), by (76)-(78), (123), and Theorem 6.1 of Chen and Yao [5], we have
\[
|z_1(t) - \bar{z}_1(t)| \leq |x_1(t) - \bar{x}_1(t)| + |m_{1,1}| \int_0^t |z_1(s) - \bar{z}_1(s)|ds + |m_{1,2}| \int_0^t |z_2(s) - \bar{z}_2(s)|ds + \tag{124}
\]
\[
2 \sup_{0 \leq u \leq T} \left| m_{1,1} \int_0^u (z_1(s) - \bar{z}_1(s))ds + m_{1,2} \int_0^u (z_2(s) - \bar{z}_2(s))ds \right|
\]
\[
|z_2(t) - \bar{z}_2(t)| \leq |x_2(t) - \bar{x}_2(t)| + |m_{2,1}| \int_0^t |z_1(s) - \bar{z}_1(s)|ds + |m_{2,2}| \int_0^t |z_2(s) - \bar{z}_2(s)|ds + \tag{125}
\]
\[
2 \sup_{0 \leq u \leq T} \left| m_{1,1} \int_0^u (z_1(s) - \bar{z}_1(s))ds + m_{1,2} \int_0^u (z_2(s) - \bar{z}_2(s))ds \right|
\]
Applying the uniform norm to (124) and (125) yields
\[
\|z(t) - \bar{z}(t)\|_T \leq \|x(t) - \bar{x}(t)\|_T + c_2 \int_0^T \|z(u) - \bar{z}(u)\|_s, \quad \text{for } t \in [0, T],
\]
and for some finite \(c_2 \geq 0\). Thus by Corollary 11.2 of Mandelbaum et al. [16],
\[
\|z(t) - \bar{z}(t)\|_T \leq \|x(t) - \bar{x}(t)\|_T \exp(c_2T). \tag{126}
\]
This shows the uniqueness of the solution \( z \) given \( x \in \mathbb{D}^2 \) since \( T \) is arbitrary. From Theorem 6.1 of Chen and Yao [5], it easily follows that \( l \) is a unique solution as well.

Next we show the existence of a solution to (76)-(78) for a given \( x \in \mathbb{D}^2 \). Fix \( T > 0 \), and let \( z^n_1(t) = 0 \), \( z^n_2(t) = 0 \), and \( l^n(t) = \sup_{0 \leq s \leq t} (-x_1(t))^+ \) for \( t \in [0, T] \). Given \( (x_1, x_2, z^n_1, z^n_2) \), let \( (z^{n+1}_1, z^{n+1}_2, l^n) \) be defined by

\[
\begin{align*}
z^{n+1}_1(t) &= x_1(t) + m_{1,1} \int_0^t z^n_1(s)ds + m_{1,2} \int_0^t z^n_2(s)ds + r_1 l^n(t), \quad (127) \\
z^{n+1}_2(t) &= x_2(t) + m_{2,1} \int_0^t z^{n+1}_1(s)ds + m_{2,2} \int_0^t z^n_2(s)ds + r_2 l^n(t), \quad (128) \\
l^n(t) &= \sup_{0 \leq s \leq t} (-x_1(t) - m_{1,1} \int_0^t z^n_1(s)ds - m_{1,3} \int_0^t z^n_2(s)ds)^+, \quad (129)
\end{align*}
\]

for \( t \in [0, T] \). Equation (129) is similar to the one-dimensional reflection mapping from Theorem 6.1 of Chen and Yao [5]. We show that the sequence \( (z^n, l^n) \) generated by (127)-(129) is Cauchy under the uniform topology, which proves the existence of a solution \( (z, l) \in \mathbb{D}^3 \) to (76)-(78), given \( x \in \mathbb{D}^2 \).

By (123), (127), (128) and (129), for \( t \in [0, T] \) we have

\[
|z^{n+1}_1(t) - z^n_1(t)| \leq |m_{1,1}| \int_0^t |z^n_1(s) - z^{n-1}_1(s)|ds + |m_{1,2}| \int_0^t |z^n_2(s) - z^{n-1}_2(s)|ds + 2 \sup_{0 \leq s \leq t} \left| -m_{1,1} \int_0^t (z^n_1(s) - z^{n-1}_1(s))ds - m_{1,3} \int_0^t (z^n_2(s) - z^{n-1}_2(s))ds \right|. \quad (130)
\]

Similarly, we can write

\[
|z^{n+1}_2(t) - z^n_2(t)| \leq |m_{2,1}| \int_0^t |z^{n+1}_1(s) - z^n_1(s)|ds + |m_{2,2}| \int_0^t |z^n_2(s) - z^{n-1}_2(s)|ds + 2 \sup_{0 \leq s \leq t} \left| -m_{2,1} \int_0^t (z^n_1(s) - z^{n-1}_1(s))ds - m_{1,3} \int_0^t (z^n_2(s) - z^{n-1}_2(s))ds \right|. \quad (131)
\]

Setting \( Z^{n+1}(t) = \|z^{n+1}(s) - z^n(s)\|_t \), we apply the uniform norm to both sides of (130) and (131) to get

\[
Z^{n+1}(t) \leq c_1 \int_0^t Z^n(s)ds, \quad \text{for } t \in [0, T],
\]

and for some \( c_1 \geq 0 \). Thus by Lemma 11.3 of Mandelbaum et al. [16],

\[
Z^{n+1}(t) \leq c_1 \frac{T^n}{n!} \sup_{0 \leq s \leq t} Z^n(s), \quad \text{for } t \in [0, T]. \quad (132)
\]

Because \( \mathbb{D}^2[0, T] \) is complete when endowed with the uniform norm, by (132), \( \{z^n, n \in \mathbb{N}\} \) is a Cauchy sequence, and has a limit \( z \) that is in \( \mathbb{D}^2[0, T] \). Then it follows from (129) that \( \{l^n, n \in \mathbb{N}\} \) also has a limit \( l \) in \( \mathbb{D}[0, T] \) as well. One can easily check that \( (z, l) \) satisfies (76)-(78). Thus, given that \( x \in \mathbb{D}^2 \), there exists a solution \( (z, l) \in \mathbb{D}^3 \) such that (76)-(78) are satisfied, since \( T \) is arbitrary. Therefore by (126) and (132), we establish that the mapping \( (\Phi, \Psi) \) exists. The continuity of the mapping in the uniform norm follows from (126).

Next we prove the continuity of \( (\Phi, \Psi) \) provided that \( \mathbb{D}^2[0, T] \) is endowed with the Skorohod \( J_1 \) topology. Consider a sequence \( \{x^n\} \) and \( x \in \mathbb{D}^2[0, T] \) such that \( x^n \to x \) as \( n \to \infty \). Also let \( (z, l) = (\Phi(x), \Psi(x)) \) and \( (z^n, l^n) = (\Phi(x^n), \Psi(x^n)) \). There exists \( M > 0 \) such that

\[
\|(z, l)(t)\|_T < M \quad (133)
\]
because $z \in \mathbb{D}^2[0,T]$ and $l \in \mathbb{D}[0,T]$. Let $\Lambda$ be the set of strictly increasing functions $\lambda: \mathbb{R}_+ \to \mathbb{R}_+$ with $\lambda(0) = 0$, $\lim_{t \to \infty} \lambda(t) = \infty$, and

$$\theta(\lambda) = \sup_{0 \leq s \leq t} \left| \log \frac{\lambda(t) - \lambda(s)}{t - s} \right| < \infty.$$ 

By Proposition 3.5.3 of Ethier and Kurtz [9] that there exists $\{\lambda^n\} \subset \Lambda$ such that

$$\lim_{n \to \infty} \theta(\lambda^n) = 0,$$

and for each $T > 0$

$$\lim_{n \to \infty} \|x^n(t) - x(\lambda^n(t))\|_T = 0. \quad (134)$$

As in the proof of Lemma 8 in Dai et al. [7], each $\lambda^n \in \Lambda$ is Lipschitz in $t$, hence differentiable a.e. Also, when $\lambda^n$ is differentiable at time $t$, its derivative $\lambda^n(t)$ satisfies

$$\left| \dot{\lambda}^n(t) - 1 \right| \leq \theta(\lambda^n). \quad (135)$$

Note that

$$\int_0^{\lambda^n(t)} z_i(s) ds = \int_0^t z_i(\lambda^n(s)) \dot{\lambda}^n(s) ds.$$ 

By (76) and (77)

$$z_i(\lambda^n(t)) = x_i(\lambda^n(t)) + m_{i,1} \int_0^{\lambda^n(t)} z_1(s) ds + m_{i,2} \int_0^{\lambda^n(t)} z_2(s) ds + \lambda^n(t)$$

Then similar to (124), for $m \geq m_{ij}$ and $m \geq r_i$, for $i, j = 1, 2$, by (136), for $i = 1, 2$

$$|z^n_i(t) - z_i(\lambda^n(t))| \leq |x^n_i(t) - x(\lambda^n(t))| + m \sum_{j=1,2} \int_0^u |z^n_j(s) - z_j(\lambda^n(s))| ds$$

$$+ 3m \sum_{j=1,2} \sup_{0 \leq u \leq t} \int_0^u |z^0_j(s) - z_j(\lambda^n(s))| 1 - \dot{\lambda}^n(s) | ds. \quad (137)$$

Note that by (133), (135), and the dominated convergence theorem we have that

$$\sup_{0 \leq u \leq t} \int_0^u |z_j(\lambda^n(s))| 1 - \dot{\lambda}^n(s) | ds \to 0 \text{ as } n \to \infty.$$ 

It follows from this, (134), (137) and Corollary 11.2 in Mandelbaum et al. [16] that

$$\|z^n_i(t) - z_i(\lambda^n(t))\|_T \to 0 \text{ as } n \to \infty \text{ for } i = 1, 2,$$

which proves continuity of $\Phi$ in the $J_1$ topology. The continuity of $\Psi$ then easily follows from (76) and (77). Q.E.D.
C. Priority-switching policy  

So far we have only considered the lightest-load-first policy. It was proved in Tezcan and Zhang [24] that this policy in general is not asymptotically optimal and priorities of levels may need to be switched to improve performance. Our goal in this section is to show that the general approach we used for the analysis of the lightest-load-first policy can be extended to such policies. First we assume that

$$\frac{\lambda^n}{n} = \gamma d_j + (1 - \gamma)d_{j+2},$$

(138)

for some \( j \in \{1, 2, \ldots, K - 2\} \), and for \( 0 \leq \gamma \leq 1 \). We assume that the priorities of classes \( j \) and \( j + 1 \) are switched. That is, once there are no agents that are at level \( j - 1 \) or below, incoming customers are sent to class \( j + 1 \) instead of \( j \). Here, our goal is to establish the diffusion limit \((\hat{Z}, \hat{X})\) under this priority-switching policy. More details of this policy can be found in Tezcan and Zhang [24]. We establish the fluid limit as a first step in §C.1, and use the result to establish the diffusion limit in §C.2.

For simplicity we make the assumption that \( 1 \leq j \leq K - 3 \) in this section. We first define \( S_4 \equiv \{0, 1, \ldots, K - 1\} \setminus \{j, j + 1\} \), and \( S_4' \equiv S_4 \setminus \{j + 2\} \). Our queueing equations again follow (1)-(7). Our queueing equations must also satisfy

$$Z_k^n(t) dA_m^n(t) = 0 \quad \text{for} \quad k \in S_4 \quad \text{and} \quad m \in \{k + 1, \ldots, K\},$$

(139)

$$Z_j^n(t) dA_m^n(t) = 0, \quad \text{for} \quad m \in \{j + 2, \ldots, K\},$$

(140)

$$Z_{j+1}^n(t) dA_m^n(t) = 0 \quad \text{for} \quad m = j \quad \text{or} \quad m \in \{j + 2, \ldots, K\}.$$  

(141)

Equations (139)-(141) represent our priority policy. By (139), if there are any agents available at level \( k \in S_4 \), then customers cannot be sent to any higher level. Similarly in (140), if any agents are available at level \( j \), then by our priority policy, incoming customers cannot be sent to any class \( m \), where \( m \in \{j + 2, \ldots, K\} \). Finally, (141) can be interpreted in a similar fashion. For any \( k \in \{1, 2, \ldots, K - 1\} \), \( L_k^n(t) \) represents the number of customers that have been sent either to the queue or to any class with a lower priority than class \( k \) by time \( t \) in the \( n \)th system. To make the queueing equations more amenable to analysis, we use the fact that

$$A_k^n(t) = L_{k-1}^n(t) - L_k^n(t) \quad \text{for} \quad k \in S_4',$$

(142)

$$A_j^n(t) = L_{j+1}^n(t) - L_j^n(t),$$

(143)

$$A_{j+1}^n(t) = L_{j-1}^n(t) - L_{j+1}^n(t),$$

(144)

$$A_{j+2}^n(t) = L_j^n(t) - L_{j+2}^n(t).$$

(145)

Thus by applying (142)-(145) to (1)-(5), we have

$$Z_0^n(t) = Z_0^n(0) - A_0^n(t) + S_1^n \left( d_1 \int_0^t Z_1^n(s) ds \right) + L_0^n(t),$$

(146)

$$Z_k^n(t) = Z_k^n(0) - 2L_{k-1}^n(t) + L_{k-2}^n(t) - S_k^n \left( d_k \int_0^t Z_k^n(s) ds \right)$$

+ \( S_{k+1} \left( d_{k+1} \int_0^t Z_{k+1}^n(s) ds \right) + L_k^n(t) \) \( \text{for} \quad k \in S_4' \setminus \{0, j + 3\} \),

(147)

$$Z_j^n(t) = Z_j^n(0) + L_{j-2}^n(t) - L_{j-1}^n(t) - L_{j+1}^n(t) - S_j^n \left( d_j \int_0^t Z_j^n(s) ds \right)$$

+ \( S_{j+1} \left( d_{j+1} \int_0^t Z_{j+1}^n(s) ds \right) + L_j^n(t),$$

(148)

$$Z_{j+1}^n(t) = Z_{j+1}^n(0) - L_{j-1}^n(t) - L_j^n(t) - S_{j+1} \left( d_{j+1} \int_0^t Z_{j+1}^n(s) ds \right)$$

+ \( S_{j+2} \left( d_{j+2} \int_0^t Z_{j+2}^n(s) ds \right) + 2L_{j+1}^n(t),$$

(149)
\[ Z^n_{j+2}(t) = Z^n_{j+2}(0) + L^n_{j-1}(t) - L^n_j(t) - L^n_{j+1}(t) - S^n_{j+2} \left( d_{j+2} \int_0^t Z^n_{j+2}(s) ds \right) \]
\[ + S^n_{j+3} \left( d_{j+3} \int_0^t Z^n_{j+3}(s) ds \right) + L^n_{j+2}(t), \]
\[ Z^n_{j+3}(t) = Z^n_{j+3}(0) - 2L^n_{j+2}(t) + L^n_j(t) - S^n_{j+3} \left( d_{j+3} \int_0^t Z^n_{j+3}(s) ds \right) \]
\[ + S^n_{j+4} \left( d_{j+4} \int_0^t Z^n_{j+4}(s) ds \right) + L^n_{j+3}(t), \]
\[ X^n(t) = X^n(0) + L^n_{K-2}(t) - S_K \left( d_K \int_0^t Z^n_K ds \right), \]

where \( Z^n_K \) is defined by (6).

**C.1. Fluid analysis** We first establish the convergence of the fluid scaled process. Under the fluid scaling, as defined in (19)-(27), we have the following result.

**Theorem 9.** Assume (18) and (28) hold. Then \( (Z^n, X^n, A^n, L^n, \bar{L}_Q^n) \) is a.s. tight. Furthermore, every limit \( (\bar{Z}, \bar{X}, \bar{A}, \bar{L}, \bar{L}_Q) \) satisfies (34)-(39), (41)-(42), (44)-(46)

\[ \bar{Z}_k(t)dA_m(t) = 0 \text{ for } k \in \mathcal{S}_4 \text{ and for } m \in \{k+1, \ldots, K\}, \]
\[ \bar{Z}_j(t)dA_m(t) = 0, \text{ for } m \in \{j+2, \ldots, K\}, \]
\[ \bar{Z}_{j+1}(t)dA_m(t) = 0 \text{ for } m = j \text{ or } m \in \{j+2, \ldots, K\}. \]

In addition \( (\bar{Z}, \bar{X}, \bar{A}, \bar{L}, \bar{L}_Q) \) is differentiable almost everywhere.

The proof is identical to the proof of Theorem 1. Next, we establish an invariant state of our fluid-scaled process \( (\bar{Z}, \bar{X}) \), under condition (138).

**Theorem 10.** Assume that (138) holds, and let \( (\bar{Z}, \bar{X}) \) be a solution of the fluid model satisfying (34)-(39), (41)-(42), and (153)-(155). Then \( (\bar{Z}, \bar{X}) \) has an invariant state, denoted by \( (\bar{Z}(\infty), \bar{X}(\infty)) \), given by \( \bar{Z}_k(\infty) = 0 \) for \( k \in \mathcal{S}'_4 \), \( \bar{X}(\infty) = 0 \), \( \bar{Z}_j(\infty) = \gamma \), \( \bar{Z}_{j+1}(\infty) = 0 \), \( \bar{Z}_{j+2}(\infty) = 1 - \gamma \).

Moreover, if \( (\bar{Z}(0), \bar{Q}(0)) = (\bar{Z}(\infty), \bar{Q}(\infty)) \) we have

\[ \bar{L}_k(t) = \lambda t \text{ for } k \in \{0, 1, \ldots, j-2\}, \quad \bar{L}_{j-1}(t) = \lambda t - d_j \gamma t \]

and \( \bar{L}_k(t) = 0 \) for \( k \in \{j, \ldots, K\} \) for all \( t \geq 0 \).

**Proof of Theorem 10:** The proof is similar to that of Theorem 2. Assume that (138) holds. By Theorem 9, the process \( (\bar{Z}, \bar{X}) \) is a.e. differentiable. For the remainder of the proof, we again only consider regular points of \( t \) where the derivative \( (\bar{Z}, \bar{X}) \) exists. As in the proof of Theorem 2, we conclude that

\[ \bar{Z}_k(\infty) = 0 \text{ for } k \in \mathcal{S}'_4. \]

It remains to show the invariant state of \( (\bar{Z}_j, \bar{Z}_{j+1}, \bar{Z}_{j+2}) \). We first show that if \( \bar{Z}_k(0) \geq \gamma \), then for all \( t > 0 \),

\[ \bar{Z}_j(t) \geq \gamma. \]

To prove this, we show that for any \( t > 0 \), if \( \bar{Z}_j(t) < \gamma \), then \( \dot{\bar{Z}}_j(t) \geq 0 \). We consider two cases: either (i) \( \bar{Z}_{j+1}(t) = 0 \) or (ii) \( \bar{Z}_{j+1}(t) > 0 \). If \( \bar{Z}_{j+1}(t) = 0 \), then from (45), (46), (148), (153), (154), and (157), we have

\[ \dot{\bar{Z}}_j(t) = \frac{1}{2}(-\lambda + d_j \bar{Z}_j(t) + d_{j+2} \bar{Z}_{j+2}(t)) \geq 0, \]
where the last inequality follows from our assumption that \( \bar{Z}_j(t) \leq \gamma \), the fact that \( d_k \) is increasing in \( k \), (45), and (138). In the other case if \( \bar{Z}_{j+1}(t) > 0 \), then from (45), (46), (149), (153), (154), and (157), we have

\[
\hat{Z}_j(t) = d_{j+1} \bar{Z}_{j+1}(t) > 0.
\]

This concludes the proof of (158). Next, we show the invariant state of \( \bar{Z}_{j+1} \). If \( \bar{Z}_{j+1}(t) > 0 \) for some \( t > 0 \), then by (45), (46), (149), (154), (155), and (157),

\[
\hat{Z}_{j+1}(t) = -\lambda + d_j \bar{Z}_j(t) + d_{j+2} \bar{Z}_{j+2}(t) - d_{j+1} \bar{Z}_{j+1}(t) \leq 0, \tag{159}
\]

where the last inequality follows from (158), (138), and the fact that \( d_k \) is increasing in \( k \). This proves that \( \bar{Z}_{j+1} = 0 \).

From (44), (45), (157), and (159), we have \( \bar{Z}_{j+2}(t) = 1 - \bar{Z}_j(t) \). Thus it only remains to show the invariant state of \( \bar{Z}_j \). For some \( t > 0 \), if \( \bar{Z}_j(t) > 0 \), then from (46), (148), (153), (154), (157), and (159), we have

\[
\hat{Z}_j(t) = \frac{1}{2} (-\lambda + d_j \bar{Z}_j(t) + d_{j+2} \bar{Z}_{j+2}). \tag{160}
\]

By (160), for any \( t > 0 \), if \( \bar{Z}_j(t) > \gamma \), \( \bar{Z}_j(t) < 0 \) and if \( \bar{Z}_j(t) < \gamma \), \( \bar{Z}_j(t) > 0 \). This completes the proof for \( (Z(\infty), X(\infty)) \). The expression for \( L \) can easily be verified by substituting the expression for \( (Z(\infty), X(\infty)) \) into (34)-(37), and by using (142)-(145). \( \Box \)

### C.2. Diffusion analysis

We now prove the convergence of the diffusion-scaled process \( (\hat{Z}^n, \hat{X}^n) \). In this section we also assume that

\[
(\hat{Q}^n(0), \hat{Z}^n_k(0) \text{ for } k \in \mathcal{S}_1') \Rightarrow 0, \text{ in } \mathbb{R}^{K-2} \text{ as } n \to \infty, \tag{161}
\]

and that \( 0 < \gamma < 1 \) in (138). Then under diffusion scaling as in (30)-(32), we have the following result.

**Theorem 11.** Assume that (29), (33), (138), (161) hold, and that \( 0 < \gamma < 1 \). Then \( (\hat{Z}^n, \hat{X}^n) \Rightarrow (\hat{Z}, \hat{X}) \) in \( D^{K+1} \) as \( n \to \infty \), where \((\hat{Z}, \hat{X})\) is the unique solution to the stochastic differential equation

\[
\begin{align*}
\hat{X}(t) &= 0, \\
\hat{Z}_k(t) &= 0 \text{ for } k \in \mathcal{S}_1', \\
\hat{Z}_j(t) &= -\hat{Z}_{j+1}(t) - \hat{Z}_{j+2}(t), \\
\hat{Z}_{j+1}(t) &= \hat{Z}_{j+1}(0) + \beta \lambda t - (d_j + d_{j+1}) \int_0^t \hat{Z}_{j+1}(s) ds + (d_{j+2} - d_j) \int_0^t \hat{Z}_{j+2}(s) ds \\
- B(t) + 2\hat{L}_{j+1}(t), \\
\hat{Z}_{j+2}(t) &= \hat{Z}_{j+2}(0) - \beta \lambda t + d_j \int_0^t \hat{Z}_{j+1}(s) ds + (d_j - d_{j+2}) \int_0^t \hat{Z}_{j+2}(s) ds + B(t) - \hat{L}_{j+1}(t), \\
\hat{Z}_{j-1}(t) d\hat{L}_{j-1}(t) &= 0, \tag{167}
\end{align*}
\]

where \( B \) is a driftless Brownian motion with variance \( \lambda(1+\gamma^2) \).

**Proof of Theorem 11:** Assume that (29), (33), (138), and (161) hold. Based on Theorem 10, we define

\[
\begin{align*}
\hat{L}_k^n(t) &= \sqrt{n} (\hat{L}_k^n(t) - \lambda t), \text{ for } k \in \{-1, 0, 1, \ldots, j-2\}, \\
\hat{L}_{j-1}^n(t) &= \sqrt{n} (\hat{L}_{j-1}^n(t) - (\lambda t - d_j \gamma t)) = \sqrt{n} (\hat{L}_{j-1}^n(t) - d_{j+1}(1-\gamma)t)) \\
\hat{L}_k^n(t) &= \sqrt{n} \hat{L}_k^n(t) \text{ for } k \in \{j, \ldots, K-1\}, \\
\hat{L}_0^n(t) &= \sqrt{n} (\hat{L}_0^Q(t) - d_K \bar{Z}_K(\infty) t). 
\end{align*}
\]
By applying diffusion scaling to (146)-(152), we have

\[ \hat{Z}_0^n(t) = \hat{Z}_0^n(0) - \sqrt{n} \left( \frac{\lambda_n}{n} - \lambda \right) t + d_1 \int_0^t \hat{Z}_1^n(s)ds + \hat{L}_0^n(t) - \hat{A}_n(t) + \hat{S}_n^n(t), \]

(168)

\[ \hat{Z}_k^n(t) = \hat{Z}_k^n(0) - 2\hat{L}^n_{k-1}(t) + \hat{L}_k^n(t) - d_k \int_0^t \hat{Z}_k^n(s)ds + d_{k+1} \int_0^t \hat{Z}_{k+1}^n(s)ds - \hat{S}_k^n(t) + \hat{S}_k^n(t+1) + \hat{L}_k^n(t), \]

for \( k \in S'_3 \setminus \{0, j+3\} \),

(169)

\[ \hat{Z}_j^n(t) = \hat{Z}_j^n(0) + \hat{L}_j^n(t) - \hat{L}_{j-1}^n(t) - \hat{L}_{j+1}^n(t) - d_j \int_0^t \hat{Z}_j^n(s)ds + d_{j+1} \int_0^t \hat{Z}_{j+1}^n(s)ds \]

(170)

\[ + \hat{L}_j^n(t) - \hat{S}_j^n(t) + \hat{S}_j^n(t), \]

(171)

\[ \hat{Z}_{j+1}^n(t) = \hat{Z}_{j+1}^n(0) - \hat{L}_{j-1}^n(t) - \hat{L}_j^n(t) - d_{j+1} \int_0^t \hat{Z}_{j+1}^n(s)ds \]

(172)

\[ + d_{j+2} \int_0^t \hat{Z}_{j+2}^n(s)ds + 2\hat{L}_{j+1}^n(t) - \hat{S}_{j+1}^n(t) + \hat{S}_{j+2}^n(t), \]

(173)

\[ \hat{Z}_{j+2}^n(t) = \hat{Z}_{j+2}^n(0) + \hat{L}_j^n(t) - \hat{L}_{j-1}^n(t) - d_{j+2} \int_0^t \hat{Z}_{j+2}^n(s)ds \]

(174)

\[ + d_{j+3} \int_0^t \hat{Z}_{j+3}^n(s)ds + \hat{L}_{j+2}^n(t) - \hat{S}_{j+2}^n(t) + \hat{S}_{j+3}^n(t), \]

(175)

\[ \hat{Z}_{j+3}^n(t) = \hat{Z}_{j+3}^n(0) - 2\hat{L}_{j+2}^n(t) + \hat{L}_{j+2}^n(t) - d_{j+3} \int_0^t \hat{Z}_{j+3}^n(s)ds + d_{j+4} \int_0^t \hat{Z}_{j+4}^n(s)ds - \hat{S}_{j+3}^n(t) + \hat{S}_{j+4}^n(t), \]

(176)

\[ \hat{A}_n \Rightarrow \hat{A} \text{ in } \mathbb{D} \text{ as } n \to \infty, \]

(177)

\[ (\hat{S}_{j+1}^n, \hat{S}_{j+2}^n) \Rightarrow (\hat{S}_{j+1}^0, \hat{S}_{j+2}^0) \text{ in } \mathbb{D}^3 \text{ as } n \to \infty, \]

where \( \hat{A}, \hat{S}_j, \hat{S}_{j+1} \) are independent driftless Brownian motions with variances \( \lambda \sigma^2, d_j \gamma, \) and \( d_{j+1}(1-\gamma) \) respectively. Next, from an identical proof as that of (62) and (63), we have

\[ (\hat{X}_n^0, \hat{Z}_0^n) \Rightarrow 0 \text{ as } n \to \infty, \text{ for } k \in S'_4. \]

(178)

It remains to establish the convergence of \((\hat{Z}_j^n, \hat{Z}_{j+1}^n, \hat{Z}_{j+2}^n)\). Note that by Theorem 10 and (29), \((\hat{Z}_j(t), \hat{Z}_{j+2}(t)) = (\gamma, 1-\gamma)\) for all \( t \geq 0 \). Hence \((\hat{L}_j^n, \hat{L}_{j+2}^n) \rightarrow (0, 0)\) a.s. u.o.c. as \( n \to \infty \) by (16). Next, note that by (168)-(175) and (178), there exists some process \( e^n \equiv (e_0^n, \ldots, e_{j-1}^n) \) such that \( e^n \Rightarrow 0 \) as \( n \to \infty \) in \( \mathbb{D}^j \) and such that

\[ \hat{L}_k^n(t) = \hat{A}_n(t) + \sqrt{n} \left( \frac{\lambda_n}{n} - \lambda \right) t + e_k^n(t) \text{ for } k \in \{0,1,\ldots,j-2\}, \]

and

\[ \hat{L}_{j-1}^n(t) = \hat{A}_n(t) - d_j \int_0^t \hat{Z}_j^n(s)ds + \sqrt{n} \left( \frac{\lambda_n}{n} - \lambda \right) t - \hat{S}_j^n(t) + e_{j-1}^n(t). \]
Next, in light of (65) and (178), it will suffice only to focus on the convergence of \((\hat{Z}_{j+1}^n, \hat{Z}_{j+2}^n)\). By (168)-(175), \((\hat{Z}_{j+1}^n, \hat{Z}_{j+2}^n)\) follow the equations

\[
\hat{Z}_{j+1}^n(t) = \hat{Z}_{j+1}^n(0) - \sqrt{n}t \left( \frac{\lambda}{n} - \lambda \right) - (d_{j+1} + d_j) \int_0^t \hat{Z}_{j+1}^n(s)ds + (d_{j+2} - d_j) \int_0^t \hat{Z}_{j+2}^n(s)ds - \hat{A}^n(t) + \hat{S}_j^n(t) - \hat{S}_{j+1}^n(t) + \hat{S}_{j+2}^n(t) + 2\hat{L}_{j+1}^n(t) + \epsilon_{j+1}^n(t),
\]

\[
\hat{Z}_{j+2}^n(t) = \hat{Z}_{j+2}^n(0) + \sqrt{n}t \left( \frac{\lambda}{n} - \lambda \right) + d_j \int_0^t \hat{Z}_{j+1}^n(s)ds - (d_{j+2} - d_j) \int_0^t \hat{Z}_{j+2}^n(s)ds + \hat{A}^n(t) - \hat{S}_j^n(t) - \hat{S}_{j+1}^n(t) - \hat{L}_{j+1}^n(t) + \epsilon_{j+2}^n(t),
\]

for some process \((\epsilon_{j+1}^n, \epsilon_{j+2}^n) \Rightarrow 0\) in \(\mathbb{D}^2\) as \(n \to \infty\). Define \(x^n \equiv (x_{j+1}^n, x_{j+2}^n)\) by

\[
x_{j+1}^n(t) = \hat{Z}_{j+1}^n(0) - \sqrt{n}t \left( \frac{\lambda}{n} - \lambda \right) - \hat{A}^n(t) + \hat{S}_j^n(t) - \hat{S}_{j+1}^n(t) + \hat{S}_{j+2}^n(t) + \epsilon_{j+1}^n(t),
\]

\[
x_{j+2}^n(t) = \hat{Z}_{j+2}^n(0) + \sqrt{n}t \left( \frac{\lambda}{n} - \lambda \right) + \hat{A}^n(t) - \hat{S}_j^n(t) - \hat{S}_{j+1}^n(t) + \epsilon_{j+2}^n(t).
\]

Note that \(x^n \Rightarrow x\) in \(\mathbb{D}^2\) by Donsker’s theorem, (138), (176), and (177), where

\[
x_{j+1}(t) = \hat{Z}_{j+1}(0) + \beta \lambda t - B(t),
\]

\[
x_{j+2}(t) = \hat{Z}_{j+2}(0) - \beta \lambda t + B(t),
\]

and \(B\) is a driftless Brownian motion with variance \(2\lambda\). Observe that \(((\hat{Z}_{j+1}^n, \hat{Z}_{j+2}^n, \hat{L}_{j+1}^n)) = (\Phi, \Psi)(x^n)\) where \((\Phi, \Psi)\) is the continuous mapping defined in Theorem 5. Since \((\Phi, \Psi)\) is continuous, we apply the continuous mapping theorem to conclude that \((\hat{Z}_{j+1}^n, \hat{Z}_{j+2}^n, \hat{L}_{j+1}^n)\) converges weakly to \((\hat{Z}_{j+1}, \hat{Z}_{j+2}, \hat{L}_{j+1})\), which satisfies (165) and (166). \(\square\)