Authors are encouraged to submit new papers to INFORMS journals by means of a style file template, which includes the journal title. However, use of a template does not certify that the paper has been accepted for publication in the named journal. INFORMS journal templates are for the exclusive purpose of submitting to an INFORMS journal and should not be used to distribute the papers in print or online or to submit the papers to another publication.

Asymptotic Analysis of Multi-Queue Service Systems with Dynamic Customer Choice

Yichuan Ding
Sauder School of Business, University of British Columbia, Vancouver, British Columbia V6T 1Z2, Canada,
daniel.ding@sauder.ubc.ca, www.stanford.edu/ y7ding

Mahesh Nagarajan
Sauder School of Business, University of British Columbia, Vancouver, British Columbia V6T 1Z2, Canada,
mahesh.nagarajan@sauder.ubc.ca

Zhe George Zhang
Department of Decision Sciences, Western Washington University, Bellingham, WA 98225, george.zhang@wwu.edu

This paper introduces a multiclass queuing model to study a stochastic service system with multiple servers and dynamic consumer choice. We consider service providers of heterogeneous quality distributed at different locations with heterogeneous customers arriving randomly to the system. Based on their own valuation of the service, the quality differentiation of the service providers, and the real-time queue length information, an arriving customer chooses a service provider to obtain service. Many real life queueing systems can be accommodated by this model. We study this system using both fluid and diffusion approximations. We prove that the fluid limit process has a unique equilibrium, and the diffusion limit process is a reflected multi-dimensional Ornstein-Uhlenbeck process centered at that equilibrium. Our model provides a useful tool for predicting the performance, evaluating the efficiency, and optimizing the system parameters for a multi-queue service systems with dynamic customer choice.

Key words: Nonlinear Complementary Problem, Fluid Approximation, Diffusion Approximation, Reflected Ornstein-Uhlenbeck Process

1. Introduction Consider a stochastic service system with multiple service providers (SPs) of heterogeneous characteristics in different locations. Customers seeking service arrive randomly to this system. Customers are heterogeneous in the value they obtain from each service provider and their disutility from the delay before being served. Customers can observe the real-time queue lengths of the entire system and then decide which service provider to seek service from, i.e., which queue to join. Customers may balk (before joining a queue) or renege (after joining a queue). The stochastic properties of such a system are not well understood and remain a well known open problem. In general, the existence of a system equilibrium and properties of the queue length process such as the stationary distribution are unknown. This paper attempts to address some of these issues. The model developed in this paper not only is of theoretical interest, but also fits many practical situations. We will review some relevant models studied in the literature.

We study a service system with $J$ parallel SPs, indexed by $j = 1, 2, \ldots, J$. These SPs are heterogeneous in service type and rate. A customer either chooses a queue to join (a specific SP) or balks based on his/her expected utility which depends on the server characteristics as well as the state of
the system at the arrival instant. After receiving service, customers leave the system permanently. We also consider the reneging behavior of customers in a later section in this paper. Previous studies on queueing systems with customer choice have certain limitations. Assumptions about the number of servers (usually single server) or congestion information (usually non-real-time) or consumer types (usually single class and homogeneous) or server types (usually homogeneous) are over-simplified for practical stochastic service systems. Under these assumptions, an exact but somewhat limited analysis is mathematically tractable. To the best of our knowledge, common approximations such as fluid and diffusion approximation schemes that work well in queueing models with simple customer behavior such as reneging or retrials under different information scenarios have not been applied to the model in this paper. Understanding the steady state equilibrium behavior of multi-queue systems with customer choice (referred to as MQSCC hereafter) via well-known system approximations is one of the main objectives of this paper.

Several practical situations studied in the operations research literature fit the MQSCC model. The first example is the wait lists for kidney transplantation for patients with end-stage renal disease (ESRD). Patients register on a transplant list. Although the wait lists in the United States are long (patient deaths due to renal failure while waiting for an organ is not uncommon), many kidneys that are medically fit for patients (usually from senior donors) are routinely turned down by patients and are discarded since patients have preferences for kidneys from younger donors. One policy that was proposed and tested by Su and Zenios [51] incorporates patient choice into the allocation framework so that wastage is minimized. Under this policy, kidneys are partitioned into \( M \) types by their quality. Arriving patients choose a certain type of kidney and wait in the corresponding queue. Thus, the waitlist virtually consists of \( M \) SPs, each of whom provides a unique type of service (kidneys). This multi-queue system was analyzed in [51] where an optimal allocation scheme for this system was derived. However, their model assumes that patients choose queues by using exogenous predictions of long-term average waiting time, whereas in reality patients choose queues based on real-time waiting time information as well as SP characteristics. The analysis in our paper helps understand the realities of this system and can be used to derive operational controls.

The second example is related to an impetus in some health care systems in North America where real time emergency room wait times in specific geographic areas are available online. For example, the web-site edwaittimes.ca, shows real time wait times for 5 major hospitals with emergency rooms in the Greater Vancouver area. Patients with preferences on locations and wait times can use this data to choose the hospital they seek care from. Yet another example is the international border crossing facilities located between the U.S. and Canada. As an example, in the pacific northwest, there are 4 border crossing facilities. Almost realtime data on wait times at each one of these facilities is available. Travelers have preferences for location and the amenities available at each facility and make their decisions based on the wait time and the characteristics of each facility.

Due to the complexity of the system, we study the queue-length process in the MQSCC using both fluid and diffusion approximations. Although these approximation techniques, as the norm for studying such stochastic systems are well known for one or two server settings with limited information and heterogeneous servers and/or customers, the general MQSCC is less understood. Our paper develops the following approximations for a general MQSCC. First, under the fluid approximation, we show that the fluid limit process converges to a unique equilibrium which can be characterized as a solution to a nonlinear complementary problem (NCP). Second, using the diffusion approximation, we show that under the heavy traffic regime, the scaled queue-length process converges to a reflected multi-dimensional Ornstein-Uhlenbeck (RMOU) process, which possesses a unique stationary distribution with closed-form density function (truncated multivariate Gaussian) under certain conditions. Third, we show that the stationary distribution of the scaled queue-length process also converges to the stationary distribution of the RMOU. This result is referred to as the \textit{interchange of limits} in the literature (See [3, 8]). These results justify the
validity of using the stationary distribution of the RMOU as an approximation to the stationary distribution of the stochastic process of the MQSCC. Finally, we extend the analysis to MQSCCs with reneging customers.

In proving the above results, we make several important technical contributions to the related research domain. For example, in the equilibrium analysis, the classical theory on NCP only guarantees the uniqueness of the equilibrium. By exploiting the special properties of the MQSCC, we prove the existence of the equilibrium by a constructive approach. When deriving the fluid limit, we identify a new sufficient condition which guarantees the pathwise uniqueness of the solution to a multi-dimensional stochastic differential equation with reflection (SDER). This SDER has been well studied in the literature and a well known sufficient condition for the pathwise uniqueness is Lipschitz continuity of the coefficients (e.g. [10, 53]); whereas our condition requires the drift coefficients to be absolute continuous and have negative definite Jacobean matrix almost everywhere, but the Jacobean can have unbounded entries so the drift coefficients does not have to be (even locally) Lipschitz continuous. Relaxing the Lipschitz assumption allows us to meaningfully model the service system of our interest; while the negative definite property naturally follows from the customer choice model of the MQSCC. We then prove that the stochastic process in the MQSCC, under appropriate scalings, converges to the fluid limit process or the diffusion limit process, respectively. To show this convergence, instead of using the existing proof techniques which heavily rely on the linearity [48] or Lipschitz continuity assumptions [37, 38], our proof leverages the mean-reversion property of the arrival rate function in the MQSCC. Our proof for the interchange of limits uses the framework introduced by Gamarnik and Zeevi [14]. However, adopting this framework to our model with non-Lipschitz drift coefficient is not trivial, and again calls for the mean-reversion property of the arrival rate. We believe, our approach and the proof technique will be useful for analyzing general state-dependent queueing networks with nonlinear and non-Lipschitz net flow rates, which refers to the difference between the arrival and departure rates. We will provide the details in later technical sections.

The results from our analysis have important practical implications. First, we have developed heavy-traffic approximations for the transient states of a large-scale system where customers are heterogeneous in server preference and waiting cost and are able to observe queue-length. Each customer then chooses an SP from a set of heterogeneous candidates. Second, due to the interchange of limits result, we can approximate the stationary distribution of the stochastic process using that of the RMOU. The closed-form characterization of the stationary distribution of RMOU allows us to calculate service-level based measures, which are useful for both performance evaluation as well as optimal control of the service system. We can also evaluate other performance measures such as the value of real time information, individual customer’s interests, and the social welfare for large systems.

The rest of the paper is organized as follows. Section 2 presents a literature review. In Section 3, a customer choice model is formulated. Based on that, we prove some properties of the arrival rate function in the MQSCC that will be used in the main analysis. Section 4 presents some notations and preliminary results that will be used in the subsequent sections. In section 5, we formally define a sequence of MQSCCs, based on which we derive the fluid and diffusion approximations in Section 6 and Section 7, respectively. In Section 8, the customer reneging behavior is considered. Finally, Section 9 concludes with a summary.

2. Literature Review

There are two streams of research closely related to our work. The first stream of papers focus on modeling and analyzing the effect of arriving customers’ queue-joining behavior in various queueing systems. These models are classified in Figure 1. As shown in Figure 1, first, there are two general classes of works in this area classified according to “information level” (IL) with O for observable and U for unobservable queues. Each class is
categorized into six types of models according to “number of queues” (NQ) with M for multiple queues and S for single queue, “customer class” (CC) with H for homogeneous and T for heterogeneous customers, and “server type” (ST) with I for identical and D for different servers. Thus, each type of model can be denoted by the notation with four letters separated by backslash (to distinguish from the forward slash used for Kendall notation). For example, our model can be denoted as $O\backslash M\backslash T\backslash D$ meaning a system with observable multiple queues, heterogeneous customers, and different servers. Customers are different in delay sensitivity and service value, but have the same service speed for the same server, while servers are different in service value and speed. Note that for each node in Figure 1, the left branch is the special case of the right branch. In reviewing the literature, it will be clear that the model we treat here is a more general version of the observable queue setting with customer choice, the one which has been less studied in the literature. In the literature review on the models in the above classification, we mainly focus on those papers that are directly related to our model. A more exhaustive reference can be found in a monograph by Hassin et al. [22] and a more recent survey by Hassin (2014).

Some of the early models of the $O\backslash S\backslash H\backslash I$ type are by Naor [46] and Leeman [30] who investigated homogeneous customers’ decisions on whether to join a queue for service. When the queue is observable, they show that in equilibrium, a pure threshold strategy (i.e., joining the queue when the queue length is below a threshold) maximizes consumer surplus. However this equilibrium solution is sub-optimal with respect to the social welfare. The socially optimal solution is reached by introducing an admission cost (toll) in addition to the waiting cost as shown in [50]. Hassin [19] found that in a last-come-first-serve queue with customer abandonment, the differences between Pareto optimal and social optimal equilibria caused due to possible customers’ negative externalities does not arise. Larsen [27] generalized Noar’s model to the one with heterogeneous customers who differ in their valuation of service. In contrast, Edelson and Hilderbrand [11] and Frutos and Gallego [13] studied the heterogeneous customer model where two classes of customers differ in their marginal waiting cost. The above models belong to $O\backslash S\backslash T\backslash I$ type. When there are multiple parallel observable queues, homogeneous customers, and identical servers (i.e., the $O\backslash M\backslash H\backslash I$ type model), the system generally does not have an equilibrium as indicated in [22], except for some special models (e.g. [20]). For this reason, the $O\backslash M\backslash H\backslash I$ type models are studied under a weaker notion of equilibrium such as the “$\epsilon$-equilibrium”. An example of $O\backslash M\backslash H\backslash D$ type model was considered in Li and Lee [32]. They consider a setting with two queues with heterogeneous servers and homogeneous customers where balkings is not allowed but jockeying is permitted. The most general case is the $O\backslash M\backslash T\backslash D$ type model, which is the far right branch in observable queue class in Figure 1. The MQSCC studied in this paper falls into this category as we assume customers have different wait cost and server preferences.

The first study on the simplest unobservable queue case or $U\backslash S\backslash H\backslash I$ type was done by Edelson and Hilderbrand [11] and Chen and Frank [4]. Two extensions followed the basic unobservable queue model. Littlechild [33] considered an M/M/1 queue with customers of heterogeneous service values which falls under the $U\backslash S\backslash T\backslash I$ type. Later, Mendelson [41] extended the model to a more general GI/G/s setting. Luski [34] generalized the model in [11] to a two-queue system which belongs to the $U\backslash M\backslash H\backslash I$ type and studied the equilibrium pricing strategies. Recently, Hua et al. [23] studied two-tier service systems with either identical or multi-class customers which are examples of $U\backslash M\backslash T\backslash I$ type or $U\backslash M\backslash T\backslash D$ type but they focused on the two queue case only. Thus most models in the unobservable queue class have been studied in the literature and are relatively well understood.

Another related area is on queues differentiated by priority levels and admission prices. Adiri and Yechiali [1] considered a system consisting of multiple parallel queues that correspond to different priority levels and admission prices, in which a customer can join a faster queue by paying a higher admission price. Maglaras et al. [36] studied a more general model in which different queues have
different waiting times and admission prices, and customers have heterogeneous valuations for being served in different queues. Other similar models related to pricing mechanisms in queueing games have been extensively studied in \[15, 21, 28, 42\]. For a comprehensive survey in this area, we again refer to the book \[22\]. However, the pricing mechanisms discussed in these models cannot be implemented in the service system studied in this paper where priority choices cannot be monetized. These previous studies with priority and pricing concerns can be classified as in Figure 1. Based on the discussion above, it should be clear that our model fills an important gap in the literature.

The second stream of related research is the one on fluid and diffusion approximations for service systems with multiple queues. In these models, the system state is usually represented by a vector, with each component representing the length of a queue. There is a rich literature that models this type of systems as multi-dimensional diffusion processes. The closest model to the MQSCC is the state-dependent queueing network studied in \[16, 29, 31, 37, 38, 56\], with some important differences. Compared to a general state-dependent queueing network model, for our model we relax the Lipschitz continuity assumption on the state-dependent arrival rate function, but make two additional assumptions. The two additional assumptions are: (i) the service rate of each SP is a constant, and all customers must leave the system after service completion instead of being routed to other queues; (ii) the arrival rate function satisfies properties in Proposition 1 in the next section as a consequence of the customer choice model. Using the special properties (i) and (ii) of MQSCC, we can derive some important properties of the heavy-traffic limit process for MQSCC: (1) the fluid limit process exhibits the mean-reversion property and converges to a unique equilibrium point; (2) the diffusion limit process is an RMOU process, whose steady-state distribution admits a closed-form characterization. These two characteristics are not valid in a general state-dependent queueing network model.

As noted earlier, our model relaxes the Lipschitz continuity, which is not a realistic assumption for many arrival rate functions in practical applications of the MQSCC. We elaborate on this in
subsequent sections. Without the Lipschitz assumption, the standard framework for proving the convergence to a fluid or a diffusion limit (see, for example, Theorem 4.6 and 7.2 of [38]) cannot be adopted or modified to treat our model. In this paper, we develop different proof techniques which take advantage of the special properties of the MQSCC.

There are several queueing models in which the fluid limit process also converges to a unique equilibrium. A well-known example is a queueing network with constant arrival rates and constant or state-dependent routing matrix [16, 17, 49]. If the arrival rate is smaller than the service rate in each queue, the drift function for each queue stays negative, so the fluid limit process has to be drawn to the unique equilibrium, i.e., the state of all empty queues. The behavior of this fluid limit process differs from that in our model, which converges to a non-zero equilibrium due to the mean-reverting drift functions. Consequently, the diffusion limit process in those queueing network models is a multi-dimensional Brownian motion (RMBM) with a reflection barrier at 0, rather than an RMOU process. Reed and Ward [48] considered a queueing network where the arrival rate is larger than the service rate in each queue and customers renege. In that model, the drift function is also mean-reverting, so the fluid limit process converges to a non-zero equilibrium and its diffusion limit process is a non-reflected multi-dimensional O-U process, both exhibiting similar properties as the MQSCC. Other similar models include a service system with differentiated service levels in [35], or with heterogeneous customer types in [18]. However, the drift function of the stochastic process is linear in those models, but is non-linear in the MQSCC. Therefore, the methods introduced in ([18, 35, 48]), which deals with linear drift functions, cannot be adapted to our model.

3. Customer Choice Model We consider a stochastic service system with \( J \) parallel heterogeneous SPs, indexed by \( j = 1, 2, \ldots, J \). We assume that the service times at server \( j \) are independent and identically distributed (i.i.d.) random variables with a finite mean \( 1/\mu_j \), and use the vector notation \( \mu := \{\mu_j\}_{j=1}^{J} \). When a customer arrives at the system, he decides whether to join any one of the \( J \) queues or balk. After a customer joins a queue, abandonments and switching between queues are not allowed (We consider an extension of the model with reneging customers in Section 8). The service discipline is First-Come-First-Served (FCFS) at each SP. A customer leaves the system permanently after one service completion.

We assume that SPs are heterogeneous so customers have heterogeneous service utilities at different SPs. When a customer arrives at the system at time \( t \), the information set available to him includes the type of each SP, the mean service rate of each SP, and the real time queue-length vector \( X(t) := (X_j(t))_{j=1}^{J} \). As mentioned earlier, such a case is known as the full information scenario in the literature. We use \( \xi \) to index different customer types. Since customer’s parameters have continuous distributions, there are uncountably many customer types. We use \( U_{\xi,j} \) to denote the expected utility perceived by customer type \( \xi \) for joining the \( j \)-th queue. We assume that \( U_{\xi,j} \) takes the following functional form.

\[
U_{\xi,j} = \begin{cases} 
 u_{\xi,j} - c_\xi \tau_j, & \text{if } j \neq 0 \text{ (joining)} \\
 0, & \text{if } j = 0 \text{ (balking)} 
\end{cases} 
\]  

\( u_{\xi,j} \) represents the reward for customer \( \xi \) being served by SP \( j \), which depends on the attributes of both customer \( \xi \) and SP \( j \); \( c_\xi \) is a non-negative coefficient which captures an individual customer’s waiting cost per time unit and depends on customer \( \xi \); \( \tau_j \) denotes the expected waiting time for queue \( j \) given by\(^1\)

\[
\tau_j = \frac{X_j(t)}{\mu_j}. 
\]  

\(^1\) We assume that the customer will ignore the remaining service time of the customer currently being served at time \( t \), which is a reasonable assumption when the queue is long and the service time is negligible in comparison to the waiting time.
The domains of $c_ξ$ and $u_ξ := \{u_ξ,j\}$ are therefore $\mathbb{R}_+ := [0, +\infty)$ and $\mathbb{R}_+^J$, respectively. The values of $u_ξ$ and $c_ξ$ are randomly drawn from the customer population subject to certain probability distributions. The probability distributions are assumed to satisfy some regularity conditions as follows.

1. $(u, c)$ has a continuous cumulative distribution function (cdf) everywhere over $\mathbb{R}_+^{J+1}$.
2. $(u, c)$ has a continuous joint probability density function (pdf) $f(u, c)$ almost everywhere except at $u \in \mathcal{K}_J$ or $c \in \mathcal{K}_I$, where

$$\mathcal{K}_J := \{u \in \mathbb{R}_+^J \mid u_i = u_j \text{ for some } i \neq j \text{ or } u_i = 0 \text{ for some } i\}.$$ (3)

For example, $\mathcal{K}_1 = \{0\}$, $\mathcal{K}_2 = \{(0, x), (x, 0), (x, x) \mid x \in \mathbb{R}_+\}$.

3. The conditional pdf $f_{c|u}$ is positive for all $u \in \mathbb{R}_+^J$ and $c \in \mathbb{R}_+$.

**Remarks on Assumptions:** The above assumptions are reasonable and have been made by balancing the requirements that facilitate our subsequent analysis while not restricting the applicability of the model. Specifically, the first assumption ensures that the parameters $u$ and $c$ would have no point mass. Therefore, we do not have to be concerned about the case when a customer finds two queues with equal expected utility, as this happens with probability zero. More importantly, the first assumption guarantees continuity of the arrival rate function as we will show in Appendix A. The continuity of the arrival rate function is a key assumption for our subsequent analysis.

The second assumption ensures that the arrival rate function has finite derivatives almost everywhere (see Appendix A). However, we have to allow the pdf $f(u, c)$ to be unbounded near points with $u \in \mathcal{K}_J$ or $c \in \mathcal{K}_I$ for the following reasons. The set $\mathcal{K}_J$ contains parameters of customers who are indifferent between server $i$ and server $j$ (thus $u_i = u_j$), as well as customers who have no interest in server $i$ at all (thus $u_i = 0$); $c = 0$ represents the case when a customer is insensitive to waiting. In order to model these types of customer behaviors, we allow the pdf to be unbounded near those points as a proxy of point mass (point mass cannot be allowed due to the first assumption). Unbounded pdf also accommodates several well-known distributions, for example, a Weibull distribution with shape parameter smaller than 1 would have unbounded density near 0.

The third assumption results in the strict diagonal dominance property of the Jacobean matrix of the arrival rate function (See Appendix A). Without this property, the fluid process may have non-unique system equilibria (a formal definition will be provided in later sections).

With the utility function defined in (1), the decision problem for a customer indexed by $ξ$ can be formulated as

$$\max_{0 \leq j \leq J}\{0, U_ξ,1, U_ξ,2, \ldots, U_ξ,J\}$$ (4)

Let $η(ξ, τ)$ denote the optimal choice of queue index for a customer $ξ$ when the waiting times are $τ = \{τ_i\}_{j=1, \ldots, J}$. $η(ξ, τ)$ is thus a random integer whose value depends on the draw of the customer, and follows the distribution

$$\Pr(η(ξ, τ) = 0) = \Pr(0 > U_ξ,k, \ k = 1, \ldots, J)$$
$$\Pr(η(ξ, τ) = j) = \Pr(U_ξ,j > U_ξ,k, \ k = 0, 1, \ldots, J, \ k \neq j),$$ (5)

where $U_ξ,0$ denotes the utility of balking which is assumed to be zero.

Our customer choice model can cover the well-known conditional logit model of McFadden et al. [39]. See the following example.

**Example 3.1** Suppose the service reward for customer $ξ$ is assumed to have the following functional form

$$u_j(ξ) = v_j(ξ) + ε_j(ξ), \text{ for } j = 1, \ldots, J,$$ (6)
where \( v_j(\xi) \) represents the deterministic reward which can be estimated based on attributes of the customer \( \xi \) and \( SP_j \), and \( \epsilon_{\xi,j} \) represents the random effect and is assumed to have a standard type-one extreme value distribution. The expected utility for customer \( \xi \) to choose queue \( j \) is given by
\[
U_{\xi,0} = 0, \quad U_{\xi,j} = v_j(\xi) - c(\xi)\tau_j + \epsilon_j(\xi) \quad \text{for} \quad j = 1, 2, \ldots, J.
\]
(7)

Then according to the conditional logit model McFadden et al. [39], the choice probability has a closed-form expression,
\[
\begin{align*}
\Pr(\eta(\xi, \tau) = 0) &= \frac{1}{1 + \sum_{k=1}^{j} \exp(v_k(\xi) - c(\xi)\tau_k)}, \\
\Pr(\eta(\xi, \tau) = j) &= \frac{1}{1 + \sum_{k=1}^{J} \exp(v_k(\xi) - c(\xi)\tau_k)}, \quad \text{for} \quad j = 1, \ldots, J.
\end{align*}
\]
(8)

We use \( p_j(\tau) \) to denote the probability that a randomly arriving customer chooses to join queue \( j \), and \( p_0(\tau) \) to denote the probability that this customer balks given a waiting-time vector \( \tau \). By assuming that the customer type is independent of the arrival sequence, we have \( p_j(\tau) = \Pr(\eta(\xi, \tau) = j) \) for \( j = 0, 1, \ldots, J \). Let \( \Lambda(\tau) := (p_j(\tau))_{j=1,\ldots,J} \) denote the vector of those choice probabilities, and \( R(\tau) := (\partial p_j(\tau)/\partial \tau_i)_{j=1,\ldots,J} \) denote the corresponding Jacobian matrix. The properties of \( \Lambda(\tau) \) and \( R(\tau) \) are summarized in the following proposition, with its proof provided in Appendix A.

**Proposition 1** The arrival-rate function \( \Lambda(\tau) = (p_j(\tau))_{j=1,\ldots,J} \) is continuous everywhere over \( \mathbb{R}_+^J \), and has continuous Jacobian matrix \( R(\tau) \) over \( \mathbb{R}_+^J \setminus K^J \), where \( K^J \), as defined in (3), is a zero-measured subset of \( \mathbb{R}_+^J \).

Moreover, \( \Lambda(\tau) \) and \( R(\tau) \) have the following properties at all \( \tau \in \mathbb{R}_+^J \):
(a) (Weak Gross Substitutability (WGS))
\[
p_j(\tau) \text{ is non-decreasing in } \tau_k \text{ for } j = 1, \ldots, J \text{ and } k \neq j.
\]
(9)

Thus, \( R(\tau) \) has non-negative off-diagonal entries.

(b) (Negative Diagonals)
\[
p_j(\tau) \text{ is strictly decreasing in } \tau_j \text{ for } j = 1, \ldots, J.
\]
(10)

Thus, \( R(\tau) \) has negative diagonal entries.

(c) (Symmetry) At all \( \tau \),
\[
\frac{\partial p_j(\tau)}{\partial \tau_i} = \frac{\partial p_i(\tau)}{\partial \tau_j}, \quad \text{for } i, j = 1, \ldots, J.
\]
(11)

so \( R(\tau) \) is a symmetric matrix.

(d) (Strict Row Diagonal Dominance) Suppose \( \tau^2 = \tau^1 + t\epsilon \) for any \( t > 0 \), then
\[
p_j(\tau^1) > p_j(\tau^2) \quad \text{for } j = 1, \ldots, J.
\]
(12)

As the result, \( R(\tau) \) has negative row-sums. Together with property (a)(b), we have \( |R_{ii}(\tau)| > \sum_{j \neq i} |R_{ij}(\tau)| \) at all \( \tau \geq 0 \), which implies strict row diagonal dominance.

(e) (Vanishes at Infinity)
\[
\forall \epsilon > 0, \exists M > 0, \text{ such that for all } j \text{ with } \tau_j > M, \quad p_j(\tau) < \epsilon.
\]
(13)
Note that Properties (a)-(d) in Proposition 1 guarantee that the Jacobean $-R(\tau)$ is a non-singular M-matrix (see e.g. Plemmons and Berman [47]) for all $\tau$. Properties (a)(b)(d) are also referred to as the mean-reversion property in this paper for reasons that will be elucidated in the proof of Theorem 3. Property (c) further guarantees that $R(\tau)$ is symmetric. Thus, given Properties (a)(b)(c)(d), $R(\tau)$ must be a symmetric negative definite matrix. Property (e) guarantees that the arrival rate vanishes when the queue length (or equivalently the waiting time) is sufficiently large, and consequently no queue would explode and the MQSCC must be stable.

**Remark 1** The proof of Proposition 1 suggests that the arrival rate function may have unbounded derivatives when $\tau$ approaches points in $\mathcal{K}^J$. Intuitively, we have allowed customers with an infinitely large probability density to be almost indifferent between two SPs (here we consider balking as one special SP with waiting time $\tau_0 = 0$ and service reward $u_0 = 0$), as this is a likely scenario in reality. As a result, when two SPs require the same waiting time, say, $\tau_i = \tau_j$, any small change to $\tau_i$ or $\tau_j$ would cause a large amount of customers to change their choice. As a result, the arrival rate function $\Lambda(\tau)$ could have infinite derivatives at points in $\mathcal{K}^J$, and be unbounded near those points. This prevents us from adopting the existing methodology, e.g., [37, 38], to establish the fluid and diffusion limits and interchange of limits in the subsequent sections.

The main results of this paper (i.e., existence and uniqueness of the equilibrium $\tau^*$, convergence of the fluid limit process to the equilibrium, convergence to the diffusion limit, and a closed-form characterization of the steady-state distribution of the diffusion limit, interchange of limits) are all built on Properties (a)-(e) of the arrival rate function $\Lambda(\tau)$ given in Proposition 1. These properties are not only sufficient, but also necessary in the sense that without some of these properties, some of these results may not hold for certain parameters. We provide examples to illustrate the necessity of those properties in Appendix B. Therefore, we cannot expect the same results to hold in a general state-dependent queueing model.

The assumptions on parameters $u$ and $c$ we made in this section serve for the purpose of deriving the properties of the arrival rate function $\Lambda(\tau)$. The subsequent analysis is based on $\Lambda(\tau)$ satisfying the properties in Proposition 1.

### 4. Notations and Preliminaries

Having set up some preliminaries on the customer choice model, we next discuss some notations to be used throughout the paper. All vectors are in boldface to differentiate from the scalars. For a sequence of random vectors $X^n$, we use $X^n \rightarrow X$ a.s., $X^n \overset{p}{\rightarrow} X$, and $X^n \Rightarrow X$ to denote almost surely point-wise convergence, convergence in probability, and convergence in distribution (weak convergence), respectively. Let $J := \{1, 2, \ldots, J\}$ denote the index set of the SPs. For a vector $a \in \mathbb{R}^J$, we use $\|a\|$ to denote the infinite norm, so $\|a\| := \max_{j \in J} |a_j|$. For two vectors $a, b \in \mathbb{R}^J$, we use $\langle a, b \rangle := \sum_{i=1}^J a_i b_i$ to represent the inner product, and use $a \circ b := (a_i b_i)_{i \in J}$ to represent the Hadamard product. For a given nonnegative vector $\mu \in \mathbb{R}_+^J$, we define the $\mu$-norm as $\|a\|^\mu := \|a \circ \mu\|$. Note that the $\mu$-norm is topologically equivalent to the $\infty$-norm. Let Diag$(a)$ denote a diagonal matrix with its diagonal entries being $a$. We use $B(t)$ to denote a $J$-dimensional standard Wiener process starting at 0.

Let $D([0, +\infty), \mathbb{R}^J)$ denote the space of right-continuous functions with left limits (RCLL) in $\mathbb{R}^J$ with time domain $[0, +\infty)$, endowed with the usual Skorokhod topology ([24]). For any $T > 0$, we define the uniform norm $\|\cdot\|_T$ on space $D([0, +\infty), \mathbb{R}^J)$ as

$$\|y\|_T = \sup\{\|y(s)\|, \ s \in [0, T]\}. \tag{14}$$

We denote $\|y\| = \sup\{\|y(s)\|, \ s \in [0, +\infty)\}$ with a slight abuse of notations. We say that $y^n \rightarrow y$ uniformly on all compact sets (u.o.c.), if $\|y^n - y\|_T \rightarrow 0$ a.s. for all $T > 0$. When $y$ is continuous, the almost surely convergence in the topology induced by the uniform norm is equivalent to almost
surely convergence in the Skorokhod topology [5]. Therefore, it suffices to discuss convergence with respect to the uniform topology on compact sets for continuous processes.

We next introduce the notations of reflection mapping, which is similar to the oblique reflection mapping defined in Chapter 7 of [5] (in our model the reflection has to be vertical to the surface). In this paper, we consider a rectangular domain $\Omega := \prod_{j \in \mathcal{J}} [a_j, b_j]$, with $-\infty \leq a < b \leq +\infty$. We let $(\Phi^{\Omega}, \Psi^{\Omega}, \Upsilon^{\Omega})$ denote the basic reflection mapping with respect to domain $\Omega$ such that $(\Phi^{\Omega}, \Psi^{\Omega}, \Upsilon^{\Omega})(z) : D([0, \infty), \mathbb{R}^J) \to D([0, \infty), \Omega \otimes \mathbb{R}^J_+)$, where $\Upsilon^{\Omega} : \mathbb{R}^J \to \Omega$. If we let $\mathbf{x} := (x_j)_{j \in \mathcal{J}}$, $l := (l_j)$, and $\mathbf{u} := (u_j)$ denote the image of $(\Phi^{\Omega}, \Psi^{\Omega}, \Upsilon^{\Omega})$, such that

$$
\begin{align*}
l_j(t) &= \sup_{0 \leq s \leq t} (a_j(s) + u_j(s) - z_j(s))^+ \\
u_j(t) &= \sup_{0 \leq s \leq t} (z_j(s) + l_j(t) - b_j(s))^+ \\
x_j(t) &= z_j(t) + l_j(t) - u_j(t).
\end{align*}
$$

(15)

It is well known that the $l_j$ and $u_j$ defined as above are the minimal non-decreasing processes which enforce $x_j(t) \in [a_j, b_j]$, and satisfy the following complementary-slackness condition:

$$
\begin{align*}
\mathbf{x} &= \mathbf{z} + \mathbf{l} - \mathbf{u} \\
l_j(0) &= 0, \quad \int_0^\infty (x_j(t) - a_j)^+dl_j(t) = 0 \\
u_j(0) &= 0, \quad \int_0^\infty (b_j - x_j(t))^+du_j(t) = 0
\end{align*}
$$

(16)

Note that in the above definition, we allow $a_j = -\infty$ ($b_j = +\infty$), then the corresponding non-decreasing process $l_j \equiv 0$ ($u_j \equiv 0$).

To analyze the queue-length process in an MQSCC, we introduce a generalized reflection mapping. Given a Lipschitz continuous mapping $\Gamma : \mathbb{R}^J \to \mathbb{R}^J$, we define

$$
(\Phi^\Gamma(z), \Psi^\Gamma(z), \Upsilon^\Gamma(z), \Theta^\Gamma(z)) \equiv (\mathbf{x}, \mathbf{l}, \mathbf{u}, \mathbf{y}),
$$

(17)

such that $(\mathbf{x}, \mathbf{l}, \mathbf{u}, \mathbf{y})$ is the unique solution to the following integral equations

$$
\begin{align*}
y(t) &= z(t) + \int_0^t \Gamma(x(s))ds, \quad \forall \ t \geq 0, \\
(\mathbf{x}, \mathbf{l}, \mathbf{u}) &= (\Phi^\Gamma, \Psi^\Gamma, \Upsilon^\Gamma)(y)
\end{align*}
$$

(18)

The following lemma, similar to Lemma 1 and Proposition 2 in Reed and Ward [48], gives a sufficient condition under which the mapping $(\Phi^\Gamma(\cdot), \Psi^\Gamma(\cdot), \Upsilon^\Gamma(\cdot), \Theta^\Gamma(\cdot))$ is well defined and Lipschitz continuous.

**Lemma 1** Suppose $z(0) \in \Omega$. If $\Gamma : \mathbb{R}^J \to \mathbb{R}^J$ is Lipschitz continuous, then there exists a unique $(\mathbf{x}, \mathbf{l}, \mathbf{u}, \mathbf{y}) \in D([0, +\infty), \Omega \otimes \mathbb{R}^J_+ \otimes \mathbb{R}^J)$, which satisfies the integral equations given in (18). As a result, we may define a mapping $(\Phi^\Gamma, \Psi^\Gamma, \Upsilon^\Gamma, \Theta^\Gamma)(z) : D([0, \infty), \mathbb{R}^J) \to D([0, \infty), \Omega \otimes \mathbb{R}^J_+ \otimes \mathbb{R}^J)$, such that $(\Phi^\Gamma, \Psi^\Gamma, \Upsilon^\Gamma, \Theta^\Gamma)(z) \equiv (\mathbf{x}, \mathbf{l}, \mathbf{u}, \mathbf{y})$. Furthermore, the mapping $(\Phi^\Gamma, \Psi^\Gamma, \Upsilon^\Gamma, \Theta^\Gamma)$ is also Lipschitz continuous.

We will use Lemma 1 to derive the diffusion limit in Section 7. Lemma 1 cannot be directly applied to the non-Lipschitz drift function. Instead, we will first use an affine function to approximate the drift function and then apply Lemma 1 to the affine (thus Lipschitz) function.

**5. Model Setup** We now provide a formal description of our model. We consider a sequence of MQSCCs indexed by $n = 1, 2, \ldots$. In all those MQSCCs, customers arrive according to a time-homogeneous Poisson process. Upon arrival, a customer chooses the $j$-th SP with probability $p_j(\tau)$. We assume that the arrival rate function $A(\tau) = (p_j(\tau))$ satisfies all the properties given in Proposition 1, and is invariant with respect to the system index $n$. We assume that the service time is i.i.d for each SP and has a coefficient of variation $c_{s,j}$ for the $j$-th SP.
In the $n$-th MQSCC, without loss of generality, we assume that customers arrive at the system (including customers who choose to balk) at a mean rate $n$. The service times at the $j$-th SP has mean $\frac{1}{n\mu_j}$, with $\mu_j^n \to \mu_j$. Let $X^n(t)$ and $\tau^n(t)$ denote the queue-length and waiting-time vectors in the $n$-th MQSCC at time $t$, respectively. Consequently, the total number of customers who have joined queue $j$ by time $t$ in the $n$-th MQSCC is given by $N \left( \int_0^t np_j(\tau(s))ds \right)$, where $np_j(\tau(s))$ gives the arrival rate of queue $j$ at time $s$ in the $n$-th MQSCC, $N(x)$ denotes a Poisson random variable with mean $x$. Finally, we use $S^n_j(t)$ to represent the cumulative number of service completions at the $j$-th SP in the $n$-th MQSCC, provided that the SP has been kept busy in $[0,t]$. $S^n_j(t)$, as a renewal process, can be formulated as

$$S^n_j(t) := \{ k | \sum_{i=1}^k b_j(k) \leq n\mu_j^n t \},$$

where $(b_j(k))$ is a sequence of i.i.d. service time random variables with mean 1, and its distribution does not depend on the MQSCC index $n$. Finally, we use $W^n_j(t)$ to denote the cumulative busy time up to time $t$. Therefore, the $j$-th SP has actually served $S^n_j(W^n_j(t))$ customers up to time $t$ in the $n$-th MQSCC.

6. Fluid Approximation Since the underlying stochastic system is difficult to analyze, we first consider a fluid approximation of the MQSCC. We prove that the scaled queue-length process in the sequence of MQSCCs converges to a fluid limit process. Moreover, we show that this fluid limit process converges to an equilibrium state which can be characterized as a solution to a Nonlinear-Complementarity-Problem (NCP).

In the $n$-th MQSCC, we define the scaled queue-length

$$x^n(t) := \frac{1}{n}X^n(t).$$

We next show that the process $x^n$ converges to a fluid limit process.

**Theorem 1** Define $\Omega := [0, +\infty)^J$ and $\Gamma_j(x) := p_j(x \circ \mu^{-1})$ for all $j \in J$. Suppose $x^n(0) \to x(0)$ a.s. when $n \to \infty$ with $x(0) \geq 0$, then for all $T > 0$,

$$\|x^n - x\|_T \to 0, \text{ a.s.}$$

where $x$ is the unique solution to the following differential equation with reflection,

$$x(t) = \Phi^\Omega \left( x(0) + \int_0^t (\Gamma(x(s)) - \mu)ds \right).$$

Mandelbaum et al. [38], in their Theorem 4.6, have proved that the queue-length process in a general state-dependent queueing network converges to the unique fluid limit process when the arrival and service rate functions are locally Lipschitz continuous. Since the arrival rate function $p_j(\cdot)$ in our model may be not locally Lipschitz continuous, our Theorem 1 is not implied by their result. In fact, if the drift coefficients $\Gamma(\cdot) = (p_j(\cdot \circ \mu^{-1}))$ in differential equation (22) are non-Lipschitz, the solution to (22) may not exist, or may not be unique. This suggests that the scaled queue-length may not converge to a unique fluid limit for general non-Lipschitz arrival rates if one makes no further assumptions. To see this, consider a one-dimensional non-stochastic differential equation,

$$x(t) = \int_0^t \sqrt{x(s)}ds.$$
The above equation has a non-Lipschitz drift $\sqrt{x(s)}$, and has two solutions, $x(t) \equiv 0$ and $x(t) = \frac{1}{3}t^2$. We can also provide an example of a differential equation with non-Lipschitz drift, to which a finite solution does not exist for the entire horizon.

$$x(t) = 1 + \int_0^t x^2(s) \, ds.$$  \hfill (24)

The above equation has a finite solution $x(t) = \frac{1}{t^2}$ only over the time window $[0,1)$.

Fortunately, the special properties of the arrival rate function in the MQSCC can replace the Lipschitz condition and allow us to prove Theorem 1. To that end, we first provide a new condition (stated in Lemma 2 below) which guarantees pathwise uniqueness\(^2\) of a solution to the differential equation (22), with its proof provided in Appendix C.

**Lemma 2** Consider the following SDER

$$\mathbf{x}(t) = \mathbf{x}(0) + \int_0^t \mathbf{b}(s, \mathbf{x}(s)) \, ds + \int_0^t \mathbf{\sigma}(s, \mathbf{x}(s)) \, dB(s) + \mathbf{c}(t),$$  \hfill (25)

where $\mathbf{c}$ is a non-decreasing process that keeps $\mathbf{x} \geq 0$ (See (15)). Suppose for all $s$, $\mathbf{\sigma}(s, \cdot)$ is Lipschitz continuous, that is, $\|\mathbf{\sigma}(s, x) - \mathbf{\sigma}(s, y)\| \leq K\|x-y\|$ for some constant $K > 0$, and $\mathbf{b}(s, \cdot)$ is absolute continuous with negative definite Jacobian matrix a.e. Then the solution to SDER (25) must be pathwise unique, if exists.

Tanaka [53] and Dupuis and Ishii [10] proved that there exists a unique solution to (25) if both $\mathbf{b}(\cdot)$ and $\mathbf{\sigma}(\cdot)$ are Lipschitz continuous. Swart [52] and Yamada and Watanabe [57] discussed pathwise uniqueness under some similar but more general conditions. While our Lemma 2 states that the Lipschitz continuity of the drift coefficient $\mathbf{b}(\cdot)$ can be replaced by absolute continuity with negative definite Jacobian a.e. Our result thus complements the existing results on pathwise uniqueness of the solution to (25).

Back to the proof for Theorem 1, the pathwise uniqueness of the solution to Equation (22) follows from Lemma 2\(^3\) with $\mathbf{b}(\cdot) = \mathbf{\Gamma}(\cdot) - \mathbf{\mu}$ and $\mathbf{\sigma}(\cdot) \equiv 0$. However, Lemma 2 does not imply the existence of a solution to Equation (22). To prove the existence of a solution, we show that the sequence $(\mathbf{x}^n)$ is tight, so any of its weak limit would be a solution to (22). A detailed proof is provided below.

**Proof of Theorem 1** By the queueing dynamics, for each queue $j$,

$$x^n_j(t) = x^n_j(0) + \frac{1}{n} \mathcal{N} \left( \int_0^t np_j(\pi^n(s)) \, ds - \frac{1}{n} S^n_j(W^n_j(t)) \right)$$

$$= x^n_j(0) + \frac{1}{n} \int_0^t \left( p_j(\mathbf{x}^n(s) \circ (\mathbf{\mu}^n)^{-1}) - p_j(\mathbf{x}^n(s) \circ (\mathbf{\mu})^{-1}) \right) ds$$

$+ \int_0^t \left( p_j((\mathbf{x}^n(s) \circ (\mathbf{\mu}^n)^{-1}) - (\mathbf{\mu}^n)^{-1}) ds + \ell^n_j(t) \right)$  \hfill (26)

where $x^n_j(t)$ was defined in (20), $\ell^n_j(t) := \mu^n_j(t-W^n_j(t))$ is the minimal non-decreasing process which ensures $x^n_j(t) \geq 0$, and

$$Z^n_j(t) := \left( \mathcal{N} \left( \int_0^t np_j(\mathbf{x}^n(s) \circ (n\mathbf{\mu}^n)^{-1}) \, ds - \int_0^t np_j(\mathbf{x}^n(s) \circ (n\mathbf{\mu}^n)^{-1}) \, ds \right) \right)$$

$+ (n\mu^n_j W^n_j(t) - S^n_j(W^n_j(t)))$  \hfill (27)

represents a mean-zero centered process.

\(^2\)One may refer to Yamada and Watanabe [57] for a rigorous definition of pathwise uniqueness.

\(^3\)For the purpose of proving Theorem 1, we only need a weaker version of Lemma 2 that deals with a non-stochastic differential equation with reflection. We presented Lemma 2 as a general result on SDER, because of its independent interest.
Using the notations of reflection mapping we defined in Section 4, we can express \( x^n(t) \) as

\[
x^n(t) = \Phi^\Omega(\tilde{z}^n + \tilde{y}^n)(t)
\]

(28)

where

\[
\begin{align*}
\tilde{z}^n(t) &:= \frac{1}{n}Z^n(t) + \int_0^t (\Lambda(x^n(s) \circ (\mu^n)^{-1}) - \Lambda(x^n(s) \circ (\mu)^{-1}))ds \\
\tilde{y}^n(t) &:= x^n(0) + \int_0^t (\Gamma(x^n(s)) - \mu)ds
\end{align*}
\]

(29)

with \( \Gamma(x) := \Lambda(x \circ \mu^{-1}) \).

By the functional strong law of large number (e.g., Theorem 5.10 in Chen and Yao [5]), and \( \mu^n \to \mu \), we have

\[
\frac{1}{n} \|N(\int_0^t p_j(x^n(s))ds) - \int_0^t np_j(x^n(s))ds\|_T \to 0
\]

\[
\frac{1}{n} \|\mu^n W^n(t) - S_j^n(W^n_j(t))\|_T \to 0.
\]

(30)

We thus conclude that

\[
\frac{1}{n} \mid Z^n \mid_T \to 0.
\]

(31)

Also, since \( \Lambda(\cdot) \) is continuous and bounded (by one), by bounded convergence, we have

\[
\left\| \int_0^t (\Lambda(x^n(s) \circ (\mu^n)^{-1}) - \Lambda(x^n(s) \circ (\mu)^{-1}))ds \right\|_T \leq \int_0^T \left\| \Lambda(x^n(s) \circ (\mu^n)^{-1}) - \Lambda(x^n(s) \circ (\mu)^{-1}) \right\| ds \to 0
\]

(32)

Equations (31) and (32) imply that \( \|\tilde{z}^n\|_T \to 0 \). Also, as \( \tilde{z}^n \) has bounded variation, we deduce that the sequence \( (\tilde{z}^n) \) is tight. It is straightforward to deduce tightness of the sequence \( (\tilde{y}^n) \). Since the reflection mapping \( \Phi^\Omega(\cdot) \) is Lipschitz continuous, we deduce that the sequence \( x^n = \Phi^\Omega(\tilde{z}^n + \tilde{y}^n) \) is also tight (See Corollary 1 of [56]). Thus, the sequence \( (x^n, \tilde{z}^n, \tilde{y}^n) \) must have at least one weak limit, say \( (\bar{x}, \tilde{z}, \tilde{y}) \). By continuous mapping theorem, this weak limit must solve the limit form of Equation (28), that is,

\[
x(t) = \Phi^\Omega(\tilde{z} + \tilde{y})(t).
\]

(33)

By our previous discussion, \( \tilde{z}^n \) must have the weak limit: \( \tilde{z}(t) \equiv 0 \). It is straightforward to show that \( \tilde{y}(t) = x(0) + \int_0^t (\Gamma(x(s)) - \mu)ds \). Therefore, the weak limit \( x \) must solve Equation (22) in the theorem. The existence of the weak limit \( x \) thus leads to the existence of a solution to (22). By Lemma 2, Equation (22) must have a pathwisely unique solution. Thus, the sequence \( x^n \) must have a unique weak limit \( x \), that is, \( x^n \to x \). Finally, since \( x \) as a solution to (22) must be continuous, the Skorohod Representation Theorem (e.g., Theorem 5.1 in [5]) implies that there exist versions of \( x^n \) and \( x \) such that \( x^n \to x \) u.o.c. with probability one.

We call \( x \) the fluid limit process of the MQSCC. Because there is a one-to-one correspondence between \( X(t) \) and \( \tau(t) \) via equation (2), we can alternatively represent the fluid limit process using \( \tau := \{\tau(t): t \geq 0\} \). We next define the equilibrium (stationary) state of this fluid limit process.

**Definition 1** \( x^* := (x^*_j) \in \mathbb{R}_+^J \) is an equilibrium queue-length vector if given \( x(0) = x^* \), the differential equation (22) has the solution \( x(t) \equiv x^* \). The associated \( \tau^* := (\tau^*_j) = \left(\frac{\tau^*_1}{\mu_j}\right) \) is referred to as an equilibrium waiting-time vector.

Intuitively, a fluid limit process is at an equilibrium state if and only if the net flow rate equals to zero for each queue. This logic leads to the following characterization of an equilibrium state.
Proposition 2 \( \tau^* \) is an equilibrium waiting-time vector of an MQSCC if and only if \( \tau^* \) is the solution to the following nonlinear complementary problem (NCP):

\[
NCP \quad \begin{aligned}
\mu_j - p_j(\tau) &\geq 0, \quad \text{for } j = 1, \ldots, J. \\
\tau_j &\geq 0, \quad \text{for } j = 1, \ldots, J. \\
\sum_{j=1}^J \tau_j(\mu_j - p_j(\tau)) &= 0.
\end{aligned}
\] (34)

Proof. If \( \tau^* \) is an equilibrium, then the arrival and departure rates must be balanced with each other in each queue. So the departure rate in each queue must be \( p_j(\tau^*) \). For queues with excessive service capacity, we must have \( \mu_j - p_j(\tau^*) > 0 \), and that queue must be empty so \( \tau^*_j = 0 \); for other queues, we have \( \mu_j - p_j(\tau) = 0 \). We thus proved the complementary slackness condition in (34).

The other inequality constraints can be proved straightforwardly.

Suppose \( \tau^* \) is a solution to (34). For queues with \( \tau^*_j > 0 \), by the complementary slackness condition in (34), we have \( \mu_j - p_j(\tau) = 0 \), which implies that the service rate and arrival rate are balanced for those queues; for queues with \( \tau^*_j = 0 \), we know that the arrival rate has not exceeded the service capacity due to the inequality constraint \( \mu_j - p_j(\tau) \geq 0 \). Since those queues are empty, the arrival and departure rates must be balanced. Thus, the drift coefficient in equation (22) must equal to zero at \( \tau^* \), which implies \( \tau(t) \equiv \tau^* \) provided that \( \tau(t) \) is a solution to (22) with \( \tau(0) = \tau^* \).

The special properties of the arrival rate function (a)-(e) allow us to establish the existence and uniqueness of an equilibrium waiting-time vector, which otherwise cannot be expected in a general state-dependent queueing network, i.e., there can be either no equilibrium or multiple equilibria. We will provide examples of these cases after the following theorem.

Theorem 2 There exists a unique equilibrium waiting-time vector \( \tau^* \) for each MQSCC.

Proof. We first prove that \( -\Lambda(\cdot) := -(p_j(\cdot))_{j=1, \ldots, J} \) satisfies the so-called P-property (Moré and Rheinboldt [43]):

\[
P\text{-Property: } \forall \tau^1, \tau^2 \in \mathbb{R}^J_+, \quad \tau^1 \neq \tau^2, \quad \min_{j=1}^J (\tau^1_j - \tau^2_j)(p_j(\tau^1) - p_j(\tau^2)) < 0. \quad (35)
\]

Without loss of generality, we assume that \( \tau^1_j - \tau^2_j = \max_j(\tau^1_j - \tau^2_j) > 0 \) for some \( j^* \), and define

\[
\Delta \tau := \tau^1_j - \tau^2_j. \quad (36)
\]

Then to prove (35), it suffices to prove that \( p_{j^*}(\tau^1) < p_{j^*}(\tau^2) \). By the definition of \( \Delta \tau \), we have \( \tau^1 \leq \tau^2 + \Delta \tau e \), but \( \tau^1_{j^*} = \tau^2_{j^*} + \Delta \tau \). Therefore, property (a) in Proposition 1 implies that

\[
p_{j^*}(\tau^1) \leq p_{j^*}(\tau^2 + \Delta \tau e). \quad (37)
\]

If we define a univariate function \( f(x) := p_{j^*}(\tau^2 + xe) \) and apply the mean value theorem to \( f(\cdot)^4 \), we get

\[
f(\Delta \tau) - f(0) = \Delta \tau f'(\zeta). \quad (38)
\]

for some \( \zeta \in [0, \Delta \tau] \). That implies

\[
p_{j^*}(\tau^2 + \Delta \tau e) - p_{j^*}(\tau^2) = \Delta \tau \sum_{i \neq j^*} R_{j^*i}(\tau^2 + \zeta e) + \Delta \tau \sum_{i \neq j^*} R_{j^*i}(\tau^2 + \zeta e)
\]

\[
< 0
\] (39)

\(^4\)The mean value theorem holds even if at some point \( x \), the derivative \( f'(x) \) may equal to \( +\infty \) or \(-\infty \) (when \( \tau^2 + xe \in K^j \)), as long as \( f'(x) \) has no jumps.
for some $\zeta \in [0, \tau]$, where $R_{ij}(\tau^2 + \zeta e)$ represents the entry at the $j^{th}$ row and $i^{th}$ column of the Jacobean matrix evaluated at $\tau^2 + \zeta e$, and the last inequality follows from Properties (b) and (d) in Proposition 1. Inequality (37) and (39) together imply that $p_j(\tau^1) < p_j(\tau^2)$, which leads to the P-property.

By Theorem 2.3 of More [44] or the comments after Theorem 1.6 of Megiddo and Kojima [40], the P-property of $-\Lambda(\tau)$ ensures that the solution to the NCP (34) is unique. However, the P-property of $-\Lambda(\tau)$ alone, in general is insufficient for the existence of a solution to the NCP. The most well known condition for existence of a solution to NCP is that the Jacobian of $-\Lambda(\tau)$ is positive bounded (every principle minor of the Jacobian of $-\Lambda(\tau)$ is bounded between $[\delta, \delta^{-1}]$ everywhere) (Cottle [7]) or that $-\Lambda(\tau)$ is a uniform P-function $\min(p_i(\tau^1) - p_j(\tau^2)) \leq -c\|\tau^1 - \tau^2\|^2$ for some $c > 0$ (Karamardian [26], More [45]). Unfortunately, neither condition is satisfied by $-\Lambda(\tau)$, as its Jacobian can be arbitrarily close to a singular matrix when $\|\tau\| \to \infty$.

The next step of the proof involves proposing a new set of sufficient conditions for the existence of a solution to an NCP of the form of (34). We claim that if $\Lambda(\tau)$ satisfies properties given (a), (b), and (e) in Proposition 1, then the NCP above must possess a solution. Note that property (d) is only needed to prove the P-property of $-\Lambda(\tau)$ and the uniqueness of the solution, but not needed for the proof of existence.

We use a constructive approach to prove the existence of the equilibrium. We prove that the equilibrium state can be achieved by iterative adjustment of the waiting times $\tau$. This adjustment process is referred to as a tatonnement process in the economics literature [2, 54]. We start with $\tau = 0$. In each iteration, we check sequentially if $\mu_j - p_j(\tau) < 0$ for each $j = 1, 2, \ldots, J$. Suppose for some $j$, $\mu_j - p_j(\tau) < 0$, then we increase the value of $\tau_j$ and keep the other components of $\tau$ unchanged until $\mu_j - p_j(\tau) = 0$. Such a $\tau$ always exists as a result of properties (b) and (e) in Proposition 1(property (b) implies that $\mu_j - p_j(\tau)$ increases continuously in $\tau_j$; property (e) ensures that $\mu_j - p_j(\tau) \to 0$ when $\tau_j \to \infty$). We repeat the above procedure sequentially for $j = 1, 2, \ldots, J$ until at some $j$, $\mu_j - p_j(\tau) \geq 0$ for $k > j$. Note that after $\tau_j$ being increased, the value of $\mu_j - p_j(\tau)$ can only decrease and turn negative again for some $\ell < j$ due to property (a), the weak gross substitutability. Therefore, we have to run the above algorithm for another iteration, that is, checking if $\mu_j - p_j(\tau) < 0$ for some $j$ and increase $\tau_j$ to make the equality to hold.

According to the above discussion, either at the very beginning $\mu_j - p_j(\tau) > 0$, or $\mu_j - p_j(\tau) < 0$ throughout the entire algorithm. We use $\tau^N$ to denote the updated value of $\tau$ in the N-th iteration. If in some iteration $N$, $\mu_j - p_j(\tau^N) \geq 0$ for all $j$, then $\tau^N$ is a solution to the NCP because $\mu_j - p_j(\tau) > 0$ only if the value of $\tau_j$ has never been updated (so $\tau_j = 0$); otherwise, we obtain a sequence of waiting-time vectors $\{\tau^N|N=1,2,\ldots\}$. We next show that $\tau^N \to \tau^* < \infty$ and $\tau^*$ is the unique solution to the NCP (34).

Without loss of generality, we assume that the value of $\tau_j$ has been updated (so $\tau_j > 0$) at iteration $N_1, N_2, \ldots, N_j, \ldots$. After each time $\tau_j$ was updated, the waiting-time vector $\tau = (\tau_1^{N_1}, \tau_2^{N_2}, \ldots, \tau_{j+1}^{N_{j+1}}, \ldots, \tau_j^{N_j-1})$ must solve the equation $\mu_j - p_j(\tau) = 0$. Therefore, the following equation must hold for each $l = 1, 2, \ldots$,

$$\mu_j - p_j(\tau_1^{N_1}, \ldots, \tau_j^{N_j}, \tau_{j+1}^{N_{j+1}}, \ldots, \tau_j^{N_j-1}) = 0.$$  (40)

Since the value of $\tau_j^N$ can only increase after each iteration, the monotone convergence theorem implies that $\tau_j \to \tau_j^*$. By property (e) in Proposition 1, $\tau_j^*$ must be a finite number, otherwise we have $\mu_j - p_j(\tau^N) \to \mu_j - 0 > 0$, which contradicts the complementarity slackness condition. By letting $l \to \infty$ and taking the limit on both sides of equation (40), we get $\mu_j - p_j(\tau^*) = 0$. By repeatedly applying this argument for $j = 1, 2, \ldots, J$, we prove that $(\mu - \Lambda(\tau^*), \tau^*)$ is a solution to the NCP (34). \qed
Remark 2  Since our proof for the existence of an NCP solution is constructive, the tatonnement algorithm introduced in the proof can be used to calculate the equilibrium queue-length vector (or equilibrium waiting-time vector).

The next theorem shows that given any initial state $\tau(0) \in \mathbb{R}_+^I$, the fluid limit process must converge to the unique equilibrium state.

Theorem 3  For any given $\tau(0) \geq 0$, the fluid limit process converges to the unique equilibrium state, i.e.,

$$\tau(t) \to \tau^*, \quad \text{when } t \to \infty. \quad (41)$$

Proof.  We define $\bar{\Delta} \tau(t) = \max_j \tau_j(t) - \tau_j^*(t)$ and $\underline{\Delta} \tau(t) = \min_j \tau_j(t) - \tau_j^*(t)$. We first prove that for any $\delta > 0$, if $\bar{\Delta} \tau(t) > \delta$, then $\bar{\Delta} \tau'(t) \leq -h(\delta)$, where $h(\delta)$ is a positive constant which depends on the value of $\delta$.

Suppose $\tau_j^*(t) - \tau_j^* = \bar{\Delta} \tau(t) \geq \delta$. Since $\tau_j^*(t) > 0$, the complementarity slackness condition implies that $\mu_j = p_j(\tau^*)$. Thus,

$$\frac{\tau_j^*(t)}{\mu_j^*} = \frac{X_j^*(t)}{\mu_j^*} = \frac{p_j(\tau(t))}{\mu_j^*} - 1 = \frac{p_j^*(\tau(t))}{\mu_j^*} - 1. \quad (42)$$

With the above equality, to show that $\tau_j^*(t) \leq -h(\delta)$, it suffices to show that

$$\frac{p_j^*(\tau(t)) - p_j^*(\tau^*)}{p_j^*(\tau^*)} \leq -h(\delta). \quad (43)$$

We prove the above inequality using a similar argument as in the proof of P-property of Theorem 2. By substituting $\tau^1 = \tau(t)$ and $\tau^2 = \tau^*$ into inequality (37) and (39), we get

$$p_j^*(\tau(t)) - p_j^*(\tau^*) \leq p_j^*(\tau^* + \bar{\Delta} \tau(t)e) - p_j^*(\tau^*)$$

$$\leq p_j^*(\tau^* + \delta e) - p_j^*(\tau^*)$$

$$= \delta R_{j^1, j^1}(\tau^* + \zeta e) + \delta \sum_{i \neq j^1} R_{j^1, i}(\tau^* + \zeta e) \quad (44)$$

for some $\zeta \in [0, \delta]$. In Equation (44), the first inequality follows from inequality (37) (which uses property (a)), and the second inequality follows from $\bar{\Delta} \tau(t) \geq \delta$ and property (d). We then define

$$h(\delta) := \frac{-\delta}{p_j^*(\tau^*)} \left( \max \{z \in [0, \delta] \mid R_{j^1, j^1}(\tau^* + ze) + \sum_{i \neq j^1} R_{j^1, i}(\tau^* + ze) \} \right). \quad (45)$$

Using properties (b)(d) in Proposition 1, we deduce that $R_{j^1, j^1}(\tau^* + ze) + \sum_{i \neq j^1} R_{j^1, i}(\tau^* + ze) < 0$ for all $z \in [0, \delta]$. Therefore, $h(\delta)$ is a positive constant that is independent of $\tau(t)$. With $h(\delta)$ defined as in (45), inequality (44) directly leads to (43). Therefore, $\tau_j^*(t) \leq -h(\delta)$ whenever $\bar{\Delta} \tau(t) \geq \delta$. An analogous argument can be used to prove that $\bar{\Delta} \tau'(t) \geq h(\delta)$ whenever $\bar{\Delta} \tau(t) \leq -\delta$. Therefore, whenever the maximum deviation of $\tau(t)$ from $\tau^*$ is greater than $\delta$, the maximum deviation is reduced at a rate of at least $h(\delta)$. We refer to this property as mean-reversion, which builds on properties (a)(b)(d) of the arrival rate function. The mean-reversion property guarantees that the maximum deviation must converge to zero, and thus the conclusion of Theorem 3 follows. \qed
7. Diffusion Approximation  According to the fluid model, the queue length vector of the MQSCC always converges to a unique equilibrium vector. Such a result provides the policy maker with the ability to forecast the average waiting time at each of the SPs. However, fluid models, as well known in the literature, only give the first-order characterization of the system dynamics and ignore the underlying stochasticity. To derive probability based risk measures or service levels, we develop a diffusion approximation to the dynamics of the MQSCC. In contrast to the fluid model, a diffusion process can capture the randomness of the queue length process. We show that the diffusion process converges to an RMOU process when the throughput rates go to infinity.

We continue to examine the sequence of MQSCCs defined in Section 5. In the $n$-th MQSCC, we define the virtual equilibrium $\tau^{n,*}$ as the solution to the following NCP

\[
\begin{align*}
\text{NCP} & : \quad \sum_{j=1}^{J} n \mu_j n - n p_j(\tau^{n,*}) \geq 0, \quad \text{for } j = 1, \ldots, J. \\
& \quad \tau^{n,*}_{j} \geq 0, \quad \text{for } j = 1, \ldots, J. \\
& \quad \sum_{j=1}^{J} \tau^{n,*}_{j}(n \mu_j n - np_j(\tau^{n,*})) = 0. \quad (46)
\end{align*}
\]

The virtual equilibrium can be interpreted as a state at which the mean arrival rate and service rate are balanced in each queue in the $n$-th MQSCC. Since we have assumed that $\mu_j n \to \mu_j$, the continuity of $p_j(\tau)$ implies that the limit of $\tau^{n,*}$ must solve the NCP (34) for the fluid model. Since the solution to (34) is unique according to Theorem 2, we deduce that $\tau^{n,*} \to \tau^*$. We use $\rho^*_j := \frac{\mu_j n}{\mu_j}$ to denote the traffic intensity at the equilibrium waiting-times. Correspondingly, we denote the traffic intensity of queue $j$ in the fluid model by $\rho_j := \lim_{n \to \infty} \rho^*_j$. We consider four exclusive cases of the limiting behaviors of the sequences $(\tau^{n,*}_j)$ and $(\rho^*_j)$. Note that these four cases are not exhaustive, but they can cover scenarios which have been most often considered in the literature (e.g., Ward and Glynn [55]).

**Largely Under-demand Queues**

$\mathcal{J}^- := \{ j | \rho^*_j \to \rho_j < 1 \}

**Balanced or Slightly Under-demand Queues**

$\mathcal{J}^- := \{ j | \tau^{n,*}_j = 0, \quad \rho^*_j \leq 1 \text{ for all } n, \quad \rho^*_j \to 1, \quad \sqrt{n} (\mu_j n - p_j(\tau^{n,*})) \to 0 \}

**Slightly Over-demand Queues**

$\mathcal{J}^+ := \{ j | \tau^{n,*}_j > 0 \text{ for all } n, \quad \tau^{n,*}_j \to \tau^*_j = 0, \quad \sqrt{n} (\mu_j n \tau^{n,*}_j - \mu_j \tau^*_j) \to 0 \}

**Largely Over-demand Queues**

$\mathcal{J}^{++} := \{ j | \tau^{n,*}_j \to \tau^*_j > 0, \quad \sqrt{n} (\mu_j n \tau^{n,*}_j - \mu_j \tau^*_j) \to 0 \}

Note that $\rho^*_j$ is no greater than one in all queues by the NCP condition. $\tau^{n,*}_j > 0$ implies that $\rho^*_j = 1$ by complementarity slackness.

We next investigate the diffusion approximation for the scaled queue-length process

\[
Q^*_j(t) := \sqrt{n}(x^*_j(t) - x^*_j). \quad (48)
\]

where $x^*$ represents the queue-lengths under fluid scaling that has been defined in Equation (20), and $x^*_j = \mu_j \tau^*_j$ gives the length of queue $j$ at the virtual equilibrium. For largely under-demand queues where $\rho_j < 1$, it is known that there is no diffusion for those queues, i.e., $Q^*_j \not\to 0$ (see e.g. Choudhury et al. [6]). Therefore we can assume that $\mathcal{J}^- = \emptyset$ without loss of generality, as those queues have constant length of zero under diffusion scaling. We can focus on characterizing the asymptotic behavior of the scaled queue-length process for other queues.

To prove existence of the diffusion limit, we need the arrival rate function to have a finite Jacobean matrix at the equilibrium $\tau^*$. Let $\mathbf{R}^*$ denote the Jacobean matrix at $\tau^*$. According to Proposition 1, $\mathbf{R}^*$ exists as long as $\tau^* \not\in \mathcal{K}^J$. The next lemma guarantees that $\tau^* \not\in \mathcal{K}^J$ with probability one if the service rates $\mathbf{\mu}$, as input data, has a continuous distribution (i.e., has no point mass).
Lemma 3 Suppose \( J^- = \emptyset \), and that the distribution of service rate vector \( \mathbf{\mu} = (\mu_j) \) has no point mass. Then with probability one, a finite \( \mathbb{R}^r \) exists.

Proof. According to Proposition 1, \( \mathbb{R}^r \) exists as long as \( \tau^* \notin K^J \). Thus, it suffices to prove that \( \tau^* \notin K^J \) with probability zero. Since \( J^- = \emptyset \), we have \( \mu_j = p_j(\tau^*) \) for all \( j \in J \). Thus, \( \Lambda(\tau^*) = \mathbf{\mu} \). In order to have \( \tau^* \in K^J, \mathbf{\mu} \) must lie in the range of \( \Lambda(K^J) \). Since \( K^J \) is a zero-measured set and \( \Lambda(\cdot) \) is a continuous mapping, \( \Lambda(K^J) \) must also have measure zero. If \( \mathbf{\mu} \) has no point mass, then \( \mathbf{\mu} \in \Lambda(K^J) \) with probability zero. Consequently, \( \tau^* \in K^J \) with probability zero.

Since \( \mathbf{\mu} \), as the reciprocal of the mean service time, has a continuous distribution in most practical applications, we can safely assume that \( \tau^* \notin K^J \) and therefore a finite \( \mathbb{R}^r \) exists. We will keep this assumption throughout our subsequent analysis.

Theorem 4 Let \( Y \) denote a \( J \)-dimensional diffusion process defined on the domain

\[
\Omega = \otimes [0, +\infty)^J = \otimes (-\infty, +\infty)^{J^+},
\]

which satisfies the following SDER,

\[
Y(t) = \int_0^t (\mathbf{R}^r \text{Diag}(\mathbf{\mu}^{-1}))(Y(s) - \mathbf{\vartheta}) - \mathbf{\theta}) \, ds + \mathbf{\Sigma}B(t) + \mathbf{L}(t) \tag{50}
\]

where \( \mathbf{\Sigma} \) is a \( J \)-by-\( J \) diagonal matrix with \( \sqrt{(1 + c_{s,j})\mu_j} \) as its \( j \)-th diagonal entry, \( \mathbf{B}(t) \) is a \( J \)-dimensional standard Brownian motion with covariance matrix \( I \) (identity matrix), and \( \mathbf{L}(t) \) is a \( J \)-dimensional minimal non-decreasing process which makes \( Y_j(t) \geq 0 \) for all \( j \in J^- \cup J^+, \mathbf{\vartheta} \) and \( \mathbf{\theta} \) are both \( J \)-dimensional vectors with

\[
\vartheta_j = \begin{cases} 
\lim_{n \to \infty} \sqrt{n}(\mu_j - p_j(\tau^{n,*})) & \text{if } j \in J^- \\
0 & \text{otherwise},
\end{cases} \quad \theta_j = \begin{cases} 
\lim_{n \to \infty} \sqrt{n}(\mu_j \tau^{n,*} - \mu_j) & \text{if } j \in J^+ \cup J^{++} \\
0 & \text{otherwise.}
\end{cases}
\]

Suppose \( Q^0(0) \Rightarrow Y(0) \) and \( \mathbb{E}[\|Y(0)\|] < \infty \). Then for all \( T > 0 \),

\[
\|Q^n - Y\|_\tau \to 0 \ a.s. \tag{52}
\]

Before proving the above result, we make a few remarks. First, according to Theorem 4, the diffusion process has a reflection barrier at 0 only for \( j \in J^- \cup J^+ \), but has no reflection barrier for \( j \in J^{++} \). Intuitively, for \( j \in J^- \cup J^+ \), we have \( Q^n_0(t) = \sqrt{n}x^n_j(t) \). Thus, \( Q^n_0(0) = 0 \) (so \( x^n_0(t) = 0 \)) means that queue \( j \) is empty, at which time the server has to stop working and prevents \( Q^n_0(t) \) from decreasing further. Therefore, if \( j \in J^- \cup J^+ \), 0 is a reflecting barrier for \( Q^n_0(t) \). For \( j \in J^{++} \), since \( x^n_j = n\tau^n_j > 0 \), an empty queue \( x^n_j(t) = 0 \) corresponds to \( Q^n_0(t) = \sqrt{n}(0 - x^n_j) \to -\infty \) when \( n \to \infty \). That means, if \( j \in J^{++} \), the reflection barrier for \( Q^n_0(t) \) is at \( -\infty \), which is equivalent to the case of no reflection barrier.

Second, we provide some interpretations of the two vectors \( \mathbf{\vartheta} \) and \( \mathbf{\theta} \) in Equation (51). For \( j \in J^- \), \( \vartheta_j = 0 \), while \( -\theta_j \) represents the negative drift that brings down \( Q^n_0(t) \) towards zero, due to the fact that the center of the RMOU is actually negative along the \( j \)-th coordinate. For \( j \in J^+ \cup J^{++} \), \( \theta_j = 0 \), and \( \theta_j \) can be considered as the center of the RMOU for queues along the \( j \)-th coordinate. Figure 2 depicts the behavior of \( Y_j \) and illustrates the role of \( \mathbf{\vartheta} \) and \( \mathbf{\theta} \) in the cases when \( j \) is in \( J^- \), \( J^+ \), and \( J^{++} \), respectively.

Finally, we want to elaborate on the relationship between our result and Theorem 7.2 in Mandelbaum et al. [38]. They developed diffusion approximation for \( \sqrt{n}(x^n(t) - x(t)) \), which is the
deviation of the scaled queue lengths from the fluid limit amplified by $\sqrt{n}$. The same result, nevertheless, cannot be expected in our model. This is because the drift coefficients $R(\tau(t))$ in the SDER (50) may have infinite values when the fluid limit $\tau(t) = x^n(t) \circ (\mu)^{-1}$ trespasses points in $K'$. Should that happen, the sequence of $Q^n$ may be not tight and the diffusion limit is not well defined.

Although in our model we cannot develop a diffusion approximation for $\sqrt{n}(x^n(t) - x(t))$, we can do so for $\sqrt{n}(x^n(t) - x^*)$, i.e., the deviation of the scaled queue lengths from the equilibrium. To do that, we assume that the fluid limit starts with the steady state ($x(0) = x^*$, or equivalently, $Q^n(0)$ converges to a finite random variable). Then by the definition of equilibrium, we know the fluid limit must keep invariant as $x(t) \equiv x^*$. Therefore, we actually developed a diffusion approximation for the deviation of the scaled queue length from its fluid limit. Moreover, in our model, the drift coefficient in the diffusion limit is the net flow rate at the equilibrium, which allows an affine approximation using the Jacobean at the equilibrium $R^*$. So we can derive the diffusion limit as an RMOU process, which has a stationary distribution due to negative definiteness of $R^*$. Such a result, however, cannot be expected in a general state-dependent queueing network, because the fluid limit there may not has an equilibrium, and the drift function would not exhibit similar properties (i.e., can be approximated by an affine function with negative definite coefficient matrix).

We want to emphasize that the framework introduced in Theorem 7.2 of [38] cannot be adapted to derive our Theorem 4, even by assuming $x(0) = x^*$ in their proof. This is because their proof framework heavily relies on the bounded derivative (or global Lipschitz) condition for the state-dependent net flow rates. Without the Lipschitz condition, several of their intermediate results cannot hold in general, including their Lemma 14.12 (compact containment), Lemma 14.13 (C-tightness), and Lemma 14.14 (characterization of the limit process); while those results are all needed for their proof of Theorem 7.2. In particular, their Lemma 14.12 states that for all $T > 0$, \{Q^n(t) | t \in [0, T]\}, as defined in (48), will be contained in a compact set with probability approaching one when $n \to \infty$. This conclusion, nevertheless, is not valid if the net flow rates (thus the drift coefficients) are non-Lipschitz. To see this, recall the example we gave in the differential equation (24), in which the drift coefficient is non-Lipschitz and its solution becomes infinitely large for $t \in [1, +\infty)$. Although that differential equation is non-stochastic, adding a stochastic term will not change the boundedness of the solution. Therefore, a non-Lipschitz net flow rate, if without additional constraints, may lead to a queue-length process that violates the compact containment condition.

In Lemma 5 of this paper, we will prove compact containment of $Q^n$ for non-Lipschitz, but mean-reverting net flow rates. Below we provide some intuition behind the proof of Lemma 5. Without the Lipschitz assumption, a small deviation of $Q^n$ may result in a large net flow rate that pushes $Q^n$ away from the equilibrium (zero), which causes compact containment to fail. However, the mean-reversion properties (properties (a)(b)(d)) ensure that any deviation of $Q^n$ can only result in a net flow rate that pulls $Q^n$ back towards the equilibrium (even though the drift can be quite large). Thus, the mean-reversion property can replace the Lipschitz condition and guarantee compact containment of $Q^n$. After proving compact containment, the rest of the proof for Theorem 4 can be completed using the generalized reflection mapping technique. This technique leverages the special characteristics of our model that the associated fluid limit must stay at the equilibrium $x^*$, and therefore differs from Lemma 14.13 and Lemma 14.14 of [38], both of which require the Lipschitz assumption.

Now we present a formal proof of Theorem 4.

**Proof of Theorem 4**  
The proof takes three major steps.
1. Step 1.
(a) When queue $j$ is slightly under-demand, $Y_j$ tends to move toward the virtual equilibrium $\theta_j = 0$ at a constant downward drift rate $\theta_j$. Meanwhile, 0 is a reflection barrier for $Y_j$.

(b) When queue $j$ is slightly over-demand (or balanced), $Y_j$ oscillates around the virtual equilibrium $\theta_j$ and is subject to a reflection barrier at 0.

(c) When queue $j$ is largely over-demand, $Y_j$ oscillates around the virtual equilibrium $\theta_j$ in an unbounded domain.

Figure 2. Typical sample paths of $Y_j$ in the cases of $j \in J^-, J^+, J^{++}$.

Because $R^*\text{Diag} (\mu^{-1}) (Y(s) - \vartheta) - \theta$ and the constant matrix $\Sigma$ are both Lipschitz continuous functions of $Y$, Theorem 4.1 of Tanaka [53] can be applied to prove that there is always a unique solution to the integral equation (50). Therefore the diffusion limit is well defined.

2. Step 2.

Instead of directly proving the convergence of the possibly unbounded process $Q^n$, we study a bounded modification of $Q^n$. We define

$$\Omega(\kappa) := [0, +\kappa]^{J^- \cup J^+} \otimes [-\kappa, +\kappa]^{J^{++}},$$

and $Q^{\kappa,n} = \Phi^{\Omega(\kappa)}(Q^n)$ as the process created from $Q^n$ by imposing reflection barriers on the finite boundary of $\Omega(\kappa)$. We prove in the following lemma that on any compact set $[0, T]$, this (bounded) multi-dimensional birth-and-death process $Q^{\kappa,n}$ weakly converges to $Y^{\kappa} := \Phi^{\Omega(\kappa)}(Y)$, which is a diffusion process created from the unbounded RMOU $Y$ by imposing reflection conditions on the boundary of $\Omega(\kappa)$.

Lemma 4 Given the conditions specified in Theorem 4, for any finite $T > 0$,

$$\{Q^{\kappa,n}(t) | 0 \leq t \leq T\} \Rightarrow \{Y^{\kappa}(t) | 0 \leq t \leq T\}.$$  

The proof of Lemma 4 is provided in Appendix D.

3. Step 3

Finally, we show that given any $T > 0$, when $\kappa \to \infty$, with probability approaching one, it takes a time larger than $T$ for $Q^{\kappa,n}$ to hit the non-zero boundary. Therefore, $Q^{\kappa,n} \equiv Q^n$ with probability approaching one over any compact horizon $[0, T]$ when $\kappa \to \infty$. We show that a similar conclusion holds for the reflected diffusion process $Y^{\kappa}$. These findings lead to the following lemma.

\footnote{In our case, the reflecting takes effect only in a subset of dimensions. However, it does not make essential difference to the proof of Theorem 4.1 of Tanaka [53].}
Lemma 5 \hspace{1em} (Compact Containment) For any $T > 0$, $\epsilon > 0$, when $\kappa \to \infty$, we have
\[
\limsup_{n \to \infty} \Pr(|Q^n_T| > \kappa) = \limsup_{n \to \infty} \Pr(|Q^{\kappa,n}_n - Q^n|_T \neq 0) \to 0
\] (55)

The proof of Lemma 5 is provided in Appendix E.

With the above results, we can prove the conclusion of Theorem 4. For all bounded, continuous real-valued function $f$ with domain $D([0,T), \mathbb{R}^J)$, when $\kappa \to \infty$, we have
\[
\limsup_{n \to \infty} \left| \mathbb{E}f(Q^n) - \mathbb{E}f(Y) \right| \
\leq \limsup_{n \to \infty} \left| \mathbb{E}f(Q^{\kappa,n} - Q^{\kappa,n}) + \mathbb{E}f(Q^{\kappa,n} - Q^\kappa) \right| \
\leq \limsup_{n \to \infty} 2\bar{f} \Pr(|Q^n - Q^{\kappa,n}|_T \neq 0) \
\to 0
\] (56)

where $\bar{f}$ represents an upper bound for $|f|$, $\limsup_{n \to \infty} \mathbb{E}f(Q^{\kappa,n} - Q^\kappa) = 0$ follows from Lemma 4, $\limsup_{n \to \infty} 2\bar{f} \Pr(|Q^n - Q^{\kappa,n}|_T \neq 0) \to 0$ follows from Lemma 5, and $\mathbb{E}f(Y^\infty - Ef(Y)) \to 0$ follows from bounded convergence and the continuous mapping theorem. Equation (56) implies that $Q^\kappa \Rightarrow Y$ on $[0,T]$. We then deduce pointwise convergence by the Skorohod’s representation theorem.

Perhaps the most useful characterization of a stochastic process is its stationary distribution. The diffusion limit process $Y$ is an RMOU and falls into the category of multi-dimensional reflected diffusion processes, the stationary distribution of which has been studied in [9, 25]. Based on the results of [25], we can derive a closed-form characterization of the stationary distribution of $Y$ under additional assumption that all SPs have the same coefficient of variation, i.e., $c_{s,j} = c_{s,1}$ for all $j \in \mathcal{J}$.

Proposition 3 \hspace{1em} The RMOU process $Y$ has a unique stationary distribution. Furthermore, if $c_{s,j} = c_{s,1}$ for all $j \in \mathcal{J}$, then the stationary distribution of $Y$ defined in Theorem 4 has a truncated multivariate Gaussian distribution. Its density function is given by
\[
\pi_Y(z) = \begin{cases} 
\frac{\pi(z)}{\int_\Omega \pi(z)dz} & \text{if } z \in \Omega, \\
0 & \text{otherwise.}
\end{cases}
\] (57)

where $\pi(z)$ is the density function of a multivariate Gaussian distribution with mean $\vartheta + \text{Diag}(\mu)(R^*)^{-1}\theta$ and covariance matrix $-\frac{1}{2}(1 + c_{s,1})\text{Diag}(\mu)(R^*)^{-1}\text{Diag}(\mu)$.

Proof. According to Example 3.10, Claim 1 of Kang and Ramanan [25], if the diffusion limit is a solution to an SDER with affine drift coefficient $C x$, and if $C^* := [A - \mathbf{Q}^{-1}Q]^{-1}C$ (see definitions in [25]) is symmetric, then $p(x) = e^{x^T C^* x}$, after normalization, gives the stationary distribution of the diffusion limit. We next check whether with the parameters in our setting, $C$, is symmetric and $p(x)$ is proportional to $\pi(z)$ as defined in the proposition. For that, we note that $Q = 0$, because the row vectors of the matrix $\mathbf{Q}$ represent the component of the reflection vector that is tangential to the boundary (see the comments after Theorem 3 in [25]), which is zero in our model where the reflection is always normal. By comparing the SDER in [25] to Equation (50), we have $A = \Sigma^2 = (1 + c_{s,1})\text{Diag}(\mu)$, $x = z - \vartheta - \text{Diag}(\mu)R^* \theta$ and $C = R^* \text{Diag}(\mu)^{-1}$. Thus, $C^* := A^{-1}C = (1 + c_{s,1})^{-1}\text{Diag}(\mu)R^* \text{Diag}(\mu)^{-1}$ is symmetric and negative definite as $R^*$ is symmetric and negative definite. We thus conclude that
\[
p(x) = \exp(x^T C^* x)
= \exp((z - \vartheta - \text{Diag}(\mu)R^*)^{-1}\vartheta)(1 + c_{s,1}^{-1})\text{Diag}(\mu)R^* \text{Diag}(\mu)^{-1}(z - \vartheta - \text{Diag}(\mu)R^* \theta)
= \exp(-\frac{\theta}{2}(z - \vartheta - \text{Diag}(\mu)R^* \theta)^T(-\frac{1}{2}(1 + c_{s,1})\text{Diag}(\mu)(R^*)^{-1}\text{Diag}(\mu))^{-1}
(z - \vartheta - \text{Diag}(\mu)R^* \theta)
\] (58)
is proportional to the density of the stationary distribution of the diffusion limit, \( \pi_Y(z) \). By looking into the above expression, we find that \( p(x) \) is proportional to the density of a multivariate Gaussian random variable with mean \( \theta + (\text{Diag}(\mu)R^*)^{-1}\theta \) and covariance matrix \(-\frac{1}{2}(1 + c_{s,1})\text{Diag}(\mu)(R^*)^{-1}\text{Diag}(\mu) \), which is denoted by \( \pi(z) \). Therefore, \( \pi_Y(z) \) is proportional to \( \pi(z) \). Normalizing \( \pi(z) \) thus leads to an exact expression for \( \pi_Y(z) \) in (57).

Note that the assumption \( c_{s,j} \equiv c_{s,1} \) is indispensable in deriving the closed-form distribution of \( Y \). Otherwise we lose symmetry of \( \pi \) and cannot apply the result of [25].

Remark 3 The multivariate Gaussian steady-state distribution provides the MQSCC manager with some practical insights. Since the covariance matrix of such a distribution is proportional to the inverse of the Jacobean \((R^*)^{-1}\), the spread of the distribution is decreasing in the scale of \( R^* \). Thus if one wishes to reduce the variability of the queue-length process of the MQSCC, one may consider increasing the scale of \( R^* \), which depends on customers’ delay sensitivity. Roughly, if customers are more sensitive to the non-zero waiting times (so a larger \( c_{s,j} \)), then \( R^* \) will have a larger scale which leads to a lower spread of the multivariate Gaussian distribution. Thus the diffusion limit process will be more concentrated at its center. Such a reduction in queue length variability will load the multi-queue service system in a more balanced way which reduces the idle times of all servers and increases the system throughput. Therefore, the MQSCC manager has an incentive to exert effort to persuade customers to actively use the queue-length information. As a result, the customer’s delay sensitivity can be increased so that the pooling effect of the MQSCC can be enhanced and system efficiency can be improved.

Proposition 3 presented above states that the stationary distribution of the limiting process \( Y \), denoted by \( \pi \), is truncated multivariate Gaussian and has a closed-form density function. In the \( n \)-th MQSCC, if we let \( \hat{b}^n_{j}(t) \), \( j = 1, 2, \ldots, J \) denote the remaining service time for the customer currently being served by the \( j \)-th SP, and define \( b^n(t) := (\hat{b}^n_{j}(t)) \). Then \( \Xi^n(t) := (Q^n(t), b^n(t)) \) is a Markov process and can be proved to have a stationary distribution. Let \( \pi^n \) denote the projection of the stationary distribution onto \( Q^n \). Then we can prove that \( \pi^n \) weakly converges to \( \pi \) when \( n \) approaches infinity. This result is also termed as interchange of limits and illustrated in Figure 3. The interchange of limits was proved when \( \Xi^n \) is the Markov process in a generalized Jackson network by Gamarnik and Zeevi [14]. We adopt their machinery and show that the interchange of limits holds in the MQSCC model. However, [14] considered a state-independent queueing network with constant mean arrival and service rates, while our model has non-Lipschitz arrival rates. Therefore, the adoption of their methods is not trivial and must exploit the special property of the MQSCC. Specifically, we use the mean-reversion property of the arrival rate to prove an important intermediate result, Proposition 4, which states that the sequence of Lyapunov functions is uniformly upper bounded.

Theorem 5 The sequence of stationary distributions, \( \pi^n \), weakly converges to \( \pi \).

Proof. The main idea of the proof is to define a Lyapunov function and to show that \( \pi^n \) is uniformly tight, which yields the existence of a limiting distribution \( \hat{\pi} \). The interchange of limits can then be proved by arguing that any such \( \hat{\pi} \) must coincide with the unique stationary distribution of \( Y \), \( \pi \).

Before we get into the details of the proof, we first introduce some definitions and notations. Note that the sequence of the Markov processes \( \Xi^n(t) \) have the common state space \( \Omega := \mathbb{R}^{2n} \) for all \( n \). A function \( V: \Omega \to \mathbb{R}^+ \) is said to be a Lyapunov function with drift size parameter \(-\gamma < 0\) and drift time parameter \( t_0 > 0 \) and exception parameter \( \kappa \) for a Markov process \( \Xi \) if

\[
\sup_{\Xi(0) \in \Omega; V(\Xi(0)) > \kappa} \{ E_{\Xi(0)}[V(\Xi(t_0)) - V(\Xi(0))] \} \leq -\gamma. \quad (59)
\]
Figure 3. The interchange-of-limit result implies that the steady-state distribution of $Y^n$, $\pi$, can be approximated by $\pi_n$, the projection of the steady-state distribution of $\Xi^n$ onto the subspace of $Q^n$.

For each $n$, define

$$
L_1(u,t,n) := \sup_{\Xi^n(0) \in \Omega} \mathbb{E}[\exp(u(\Xi^n(t)) - V(\Xi^n(0)))|\Xi^n(0)]
$$

$$
L_2(u,t,n) := \sup_{\Xi^n(0) \in \Omega} \mathbb{E}[(V(\Xi^n(t)) - V(\Xi^n(0)))^2 \exp(u(V(\Xi^n(t)) - V(\Xi^n(0)))^+)|\Xi^n(0)]
$$

for any $u > 0$, $t \geq 0$.

The key to proving Theorem 5 is to utilize the following proposition, which is in analogue to Proposition 3 in [14] but deals with the MQSCC case. Note that we have used different notations from those used in [14]: our $Q^n(t)$ corresponds to the notation $\frac{1}{\sqrt{n}}Q^n(nt)^\mu$ in their paper. Because we have used a different scale, the bound we derived with respect to the $\|\cdot\|_0$ norm is exactly the bound derived in their paper the interval $[0, nt_0]$.

**Proposition 4** Let $V(\Xi^n(t)) := \|Q^n(t)\|^{\mu^{-1}}$. Then for sufficiently large $n$, $V(\cdot)$ is a Lyapunov function with drift size parameter $-1$, drift time parameter $t_0$, and exception parameter $\kappa$ for some $\kappa, t_0 > 0$. In addition, there exists $u_0$ such that

$$
\limsup_{n \to \infty} L_1(u_0,t_0,n) < \infty
$$

$$
\limsup_{n \to \infty} L_2(u_0,t_0,n) < \infty
$$

(61)

Proposition 4 is proved in Appendix F.

With Proposition 4, the rest of the proof for Theorem 5 follows the same routine as in [14]. Specifically, Proposition 4 implies that $V(\cdot)$ is a Lyapunov function with parameter $-1$, $t_0$, and $\kappa$. Moreover, the second inequality in (61) implies that there exists $u_0$, such that $u_0 L_2(u_0,t_0,n) < 1$ for all sufficiently large $n$. Thus, both conditions of Theorem 6 in [14] are satisfied for all sufficiently large $n$. We then invoke their Theorem 6 and deduce that $1 - u_0/2 > 0$ and the following inequality holds for all sufficiently large $n$,

$$
\Pr_{\pi^n}(\|Q^n(0)\|_T > s) \leq (1 - u_0/2)^{-1} L_1(u_0,t_0,n) \exp(-u_0(s - \kappa)).
$$

(62)

By the above inequality and the inequality in (61), we have

$$
\Pr_{\pi^n}(\|Q^n(0)\|_T > s) \leq H_1 \exp(-h_1 s),
$$

(63)

for properly selected positive constants $H_1$ and $h_1$. Inequality (63) implies the uniform tightness of the sequence of distributions $(\pi^n)$. The rest of the proof follows exactly as in Theorem 8 of [14].

In the next section, we discuss the results for the case where customers can renege after joining a queue.
8. MQSCC with Reneging Customers Our main results on MQSCC can be extended to the case when customers may renege (or abandon) before they get served. Note that in most past studies on the queues with customer choice, the reneging feature was not considered due to the reason that the joining decision has been made based on the expected service utility. However, we added this feature due to the motivation of our model. If the model is for modeling the healthcare service, the death or unexpected medical condition change may lead to customer abandonment. Since the analysis is quite similar to the one above, we will elaborate the results where the technical differences are significant.

We assume that customers renege after an exponentially distributed period with mean of $\frac{1}{d}$. When the system is Markovian (the inter-arrival times, reneging times, and service times are all exponentially distributed), the following expression given in [58] can be used to compute the expected waiting time

$$\tau_j = \frac{1}{d} \log(1 + \frac{X_j d}{\mu_j}). \quad (64)$$

We assume that all customers use (64) to compute their expected waiting time, and choose a queue which maximizes their payoff $U_{\xi,j}$ as given in (1), which leads to state-dependent arrival rate function $\Lambda(\tau)$. Because our proof for Proposition 1 does not rely on the functional form of $\tau_j$ with respect to $X_j$, the proof can be adapted to showing that all the properties of the arrival rate function $\Lambda(\cdot)$ hold in the presence of reneging customers.

**Corollary 1** All the properties listed in Proposition 1 hold in an MQSCC with customer reneging.

**Remark 4** If customer reneging is allowed, the MQSCC is always stable. So condition (e) of Proposition 1 is no longer necessary.

We next prove that the fluid process in an MQSCC with customer reneging converges to the equilibrium state, which is the unique solution to an NCP with a slightly different formulation compared to the non-reneging case.

**Theorem 6** The equilibrium waiting times of an MQSCC with reneging is the unique solution to the following Nonlinear Complementary Problem (NCP).

$$\begin{align*}
NCP \\
Z_j := \mu_j \exp(\tau_j d) - p_j(\tau) &\geq 0, \quad \text{for } j = 1, \ldots, J. \\
\tau_j &\geq 0, \quad \text{for } j = 1, \ldots, J. \\
\tau_j Z_j &= 0, \quad \text{for } j = 1, \ldots, J.
\end{align*} \quad (65)$$

Moreover, if we use $\tau(t)$ to denote the waiting-time vector in a fluid model, then for any given $\tau(0) \geq 0$, $\tau(t) \to \tau^*$. 

**Proof.** By defining $\tilde{p}_j := \mu_j \exp(\tau_j d) - p_j(\tau)$, the above NCP can be rewritten into a similar form of (34) by replacing the arrival function $\Lambda(\cdot)$ with $\tilde{\Lambda}(\cdot) := (\tilde{p}_j(\cdot))_{j=1,\ldots,J}$. Note that the Jacobian for $\tilde{\Lambda}(\tau)$ has the form $\tilde{R} = \sigma(\tau) + R$, where $R$ is the Jacobian of $p(\tau)$ and is a symmetric negative definite matrix by Corollary 1, and $\sigma(\tau)$ is a diagonal matrix with the $j$-th entry $\sigma_{jj}(\tau) = \mu_j d \exp(\tau_j d) > 0$. Because of the extra term $\sigma(\tau)$, we are now able to prove that $\tilde{p}_j$ satisfies the uniform P-property, i.e.,

**Uniform P-Property:** $\forall \tau^1, \tau^2 \in \mathbb{R}^J_+, \tau^1 \neq \tau^2$, $\min_{j=1}^J (\tau_j^1 - \tau_j^2)(\tilde{p}_j(\tau^1) - \tilde{p}_j(\tau^2)) < c\|\tau^1 - \tau^2\|^2, \quad (66)$

with $c > d \max_j \mu_j > 0$. Thus, the classical theorem by Cottle [7] implies the existence of a unique solution to the NCP (65).
To prove \( \tau(t) \to \tau^* \), we define \( \Delta \tau(t) = \max_j (\tau_j(t) - \tau_j^*) \), and \( \Delta \tau(t) = \min_j (\tau_j(t) - \tau_j^*) \). We want to prove that \( \Delta \tau'(t) \leq \kappa(\delta) \) for some constant \( \kappa(\delta) > 0 \) whenever \( \Delta \tau(t) \geq \delta \). Without loss of generality, assume that \( \tau_j(t) - \tau_j^* = \Delta \tau(t) \), then \( \tau_j(t) > 0 \) and (64) imply that

\[
\tau'_j(t) = \frac{p_j(\tau) - \mu_j}{X_j(t)d + \mu_j} \leq \frac{p_j - \mu_j}{p_j(\tau^*)}.
\]

(67)

where the inequality follows from the NCP constraint \( Z_j = \mu_j(\exp(\tau_jd)) - p_j(\tau^*) = \mu_j + X_j^*d - p_j(\tau^*) \geq 0 \).

The rest of the proof resembles the proof of Theorem 3, i.e., we prove facts (1) and (2) and show that \( \Delta \tau'(t) \leq -\kappa(\delta) \). We then use the similar argument to show that \( \Delta \tau(t) \geq \kappa(\delta) \) whenever \( \Delta \tau(t) \leq -\delta \) and prove \( \tau(t) \to \tau^* \).

The proof of the convergence to the diffusion limit is a simple extension of Theorem 4 by including an extra term \(-dI\) in the drift matrix as a result of reneging. We summarized the result below. The notations follow from the definitions in the previous sections.

As before we partition the index set of queues into four subsets \( \mathcal{J}^- \), \( \mathcal{J}^- \cup \mathcal{J}^+ \), and \( \mathcal{J}^+ \) according to (47), and redefine \( \theta \) and \( \vartheta \) as following

\[
\theta_j = \begin{cases} 
\lim_{n \to \infty} \sqrt{n}(\mu^0_j - p_j(\tau^{n*} - \tau^*)) & \text{if } j \in \mathcal{J}^- \\
0 & \text{otherwise},
\end{cases}
\]

\[
\vartheta_j = \begin{cases} 
\lim_{n \to \infty} \sqrt{n}(X_j^{n*} - X_j^*) & \text{if } j \in \mathcal{J}^\cup \mathcal{J}^+ \\
0 & \text{otherwise}.
\end{cases}
\]

(68)

where \( X_j^{n*} = \mu^0_j \frac{1}{\sqrt{n}}(\exp(d\tau_j^{n*}) - 1) \) represents the queue length corresponding to expected waiting time of \( \tau_j^{n*} \). We next derive a diffusion approximation for the MQSCC.

**Theorem 7** Suppose \( \mathcal{J}^- = \emptyset \), and \( Q^n(0) \Rightarrow Y(0) \), with \( \mathbb{E}\|Y(0)\| < \infty \). Suppose \( \{Y(t) | t \geq 0\} \) is a \( J \)-dimensional diffusion process defined on the domain

\[
\Omega = [0, +\infty)^{\mathcal{J}^\cup \mathcal{J}^+} \otimes (-\infty, +\infty)^{\mathcal{J}^+},
\]

and satisfies the following stochastic-differential-equation (SDE),

\[
Y(t) = \int_0^t \left( (R^\ast \text{Diag}((1 + \tau^* d) \circ \mu)^{-1}) - dI \right) Y(s - \theta) \, ds + \int_0^t \Sigma^R dB(s) + L(t)
\]

(70)

where \( I \) is an \( J \)-by-\( J \) identity matrix, \( \Sigma^R \) is a \( J \)-by-\( J \) diagonal matrix with \( \sqrt{\left( \sigma_{x,j}^2 + \exp(\tau_j^* d) \right) \mu_j} \) as its \( j \)-th diagonal entry, \( B(t) \) is a \( J \)-dimensional Brownian motion, and \( L(t) \) is a \( J \)-dimensional minimal non-decreasing process which makes \( Y_j(t) \geq 0 \) for all \( j \in \mathcal{J}^- \cup \mathcal{J}^+ \). Suppose \( Q^n(0) \Rightarrow Y(0) \) and \( \mathbb{E}\|Y(0)\| < \infty \). Then for all \( T > 0 \),

\[
\|Q^n - Y\|_T \to 0 \ a.s.\]

(71)

**Proof.** The proof is mostly similar to that of Theorem 4, but differs in two places: (1) the derivative of \( Q^n_j(t) \) includes an extra term \(-dX_j(t)\), which counts the aggregate reneging rate at time \( t \); (2) the waiting time is no longer linear in \( X_j(t) \) but has to be computed using equation (64). We will prove Lemma 4 by highlighting the parts due to the above differences. The rest of the proof, including Lemma 5, follows the same routine as in the proof for Theorem 4 and we will not repeat it here.

We next prove Lemma 4, that is, by restricting the process to stay inside the bounded domain \( \Omega(\kappa) \), the bounded process \( \{Q^n_j(t) | 0 \leq t \leq T\} \) weakly converges to \( \{Y^\ast(t) | 0 \leq t \leq T\} \). We first express \( Q^n_j(t) \) in a similar way to (103) as follows:
\[ Q_{\kappa,n}^+(t) = Q_{\kappa,n}^+(0) + n^{-1/2}N \left( n \int_0^t p_j(\tau^n(s))ds \right) - n^{-1/2}N \left( f_0^t dX_{\kappa,n}^+(s)ds \right) - n^{-1/2}S_{\kappa,n}^+(t) \]
\[ + n^{-1/2}L_{\kappa,n}^+(t) - n^{-1/2}U_{\kappa,n}^+(t) \]
\[ = Q_{\kappa,n}^+(0) + \int_0^t \left( n^{-1/2} \left( p_j(\tau^n(s)) - \mu_j^n - n^{-1}dX_{\kappa,n}^+(s) \right) - \left( \sum_{i} R_{ji}^* \left( \frac{1}{1 + \tau_i d} \right) \right) \right) ds \]
\[ + n^{-1/2}Z_{\kappa,n}^+(t) + \int_0^t \left( \sum_{i} R_{ji}^* \left( \frac{1}{1 + \tau_i d} \right) \right) \left( Q_{\kappa,n}^+(s) - \vartheta_i - \theta_j \right) ds + \frac{1}{\sqrt{n}} L_{\kappa,n}^+(t) - \frac{1}{\sqrt{n}} U_{\kappa,n}^+(t), \]

where \( X_{\kappa,n}^+(s) \) denotes the queue-length process restricted to the domain \( \Omega(\kappa) \). Note that the centered process \( Z_{\kappa,n}^+ := \left( Z_{\kappa,n}^i \right) \) has included an extra term for the reneging customers, which has the expression

\[ Z_{\kappa,n}^+(t) := \left( N\left( \int_0^t np_j(\tau^n(s))ds \right) - \int_0^t np_j(\tau^n(s))ds \right) \]
\[ + \left( n\mu_j^n t - S_{\kappa,n}^+(t) \right) - \left( N\left( \int_0^t dX_{\kappa,n}^+(s)ds \right) - \int_0^t dX_{\kappa,n}^+(s)ds \right) \]

We next analyze the terms labeled as (A.1)-(A.3) in (72).

1. Our assumption of the initial value implies that (A.1) \( \Rightarrow Y(0) \).

2. Since \( \tau_j \) has to be computed using (64), the expression for \( \Delta \tau_{\kappa,n}^+ \) will be

\[ \Delta \tau_{\kappa,n}^+ := \tau_{\kappa,n}^+(s) - \tau^* \]
\[ = \frac{1}{d} \log \left( 1 + \left( n^{1/2}Q_{\kappa,n}^+(s) + \tau^* \right) \circ (n\mu^n)^{-1} \right) - \frac{1}{d} \log \left( 1 + \left( X_{\kappa,n}^+ \circ (n\mu^n)^{-1} \right) \right). \]

It is not difficult to show that \( n^{2} \| \Delta \tau_{\kappa,n}^+ \|_1 \) is uniformly bounded and thus it suffices to expand the Taylor series of \( n^{1/2} \left( p_j(\tau_{\kappa,n}^+ + \Delta \tau_{\kappa,n}^+(s)) - p_j(\tau_{\kappa,n}^+) \right) \) till its first-order term. Thus, using a similar argument to the one we used in deriving equation (108), we have

\[ n^{1/2} \left( p_j(\tau_{\kappa,n}^+ + \Delta \tau_{\kappa,n}^+(s)) - p_j(\tau_{\kappa,n}^+) \right) \rightarrow \sum_{i} \frac{Q_{\kappa,n}^+(s) - \vartheta_i}{\left( 1 + \tau_j d \right) \mu_i^n} R_{ji}^* \]

Thus, by our definition of \( \theta_j \) and \( \vartheta_j \), we have

\[ n^{1/2} \left( p_j(\tau_{\kappa,n}^+(s)) - \mu_j^n - n^{-1}dX_{\kappa,n}^+(s) \right) \]
\[ = n^{1/2} \left( p_j(\tau_{\kappa,n}^+ + \Delta \tau_{\kappa,n}^+(s)) - p_j(\tau_{\kappa,n}^+) + n^{1/2} \left( p_j(\tau_{\kappa,n}^+) - \mu_j^n - n^{-1}dX_{\kappa,n}^+ \right) \right) \]
\[ + n^{1/2} \left( n^{-1}dX_{\kappa,n}^- - n^{-1}dX^+ \right) - n^{1/2} \left( n^{-1}dX_{\kappa,n}^+(s) - n^{-1}dX^+ \right) \]
\[ \rightarrow \sum_{j} \frac{Q_{\kappa,n}^+(s) - \vartheta_j}{\left( 1 + \tau_j d \right) \mu_j^n} R_{ji}^* - \theta_j + d\vartheta_j - dQ_{\kappa,n}^+(s) \]

The above convergence leads to that \( (A.2) \rightarrow 0 \) uniformly over any compact set.

3. \( n^{-1/2}Z_{\kappa,n}^+(t) \) is the sum of three centered processes. We have shown in the proof of Theorem 4 that the sum of the first two terms converges to \( \Sigma B(t) \) with \( \Sigma \) a diagonal matrix and \( \Sigma_{jj} = \sqrt{\left( 1 + c_{\kappa,j} \right) \mu_j} \), respectively. Since \( \frac{1}{n} \int_0^t dX_{\kappa,n}^+(s)ds \rightarrow \frac{1}{n} dX_j^+t \) uniformly on any compact set \( t \in [0,T] \), and \( \frac{1}{n} \int_0^t dX_{\kappa,n}^-(s)ds \) is a non-decreasing process in \( t \), we may invoke the random time-change theorem and FCLT to prove that

\[ n^{-1/2} \left( N\left( \int_0^t dX_{\kappa,n}^+(s)ds \right) - \int_0^t dX_{\kappa,n}^+(s)ds \right) \Rightarrow B_j^0(t). \]
where $B_D(t)$ is a Brownian motion whose covariance matrix is a diagonal matrix and the $j$-th entry of its diagonal is given by $(\exp(\tau_j^* d) - 1)\mu_j$. Since $n^{-1/2}Z^{\kappa,n}(t)$ is the sum of three independent Brownian processes, we deduce that

$$n^{-1/2}Z^{\kappa,n}(t) \Rightarrow \Sigma^R B(t).$$  \hspace{1cm} (78)

The rest of the proof follows the proof for Lemma 4.

9. Conclusions and Future Research  
Our paper is the first to apply the heavy traffic approximations to the general MQSCC problem and derive properties (and existence) of the equilibrium queue length. As mentioned, we not only make the theoretical contributions but also address some of the managerial issues from practical systems. For one, our results can help practitioners implement system evaluation metrics for controlling this kind of stochastic systems. We have made several significant assumptions that enable us to do the analysis. We have ignored the issue of endogenous abandonments. In some situations, waiting customers may abandon the current queue and join a different queue. Usually, when a customer abandons the current queue, he has to lose his priority in that queue and has to wait at the end of the new queue. This usually happens when there is sudden change to a customer’s waiting cost or utility because of accidental event occurring or additional information becoming available. The switching behavior is equivalent to the event that a customer reneges in one queue and a new customer joins another queue. Our conjecture is that it will not change the behavior of the MQSCC and thus will not affect the results significantly. Relaxing some of our assumptions, such as Poisson arrival and exponential reneging times, can be an interesting but more challenging topic for future research. With the performance measures developed in this paper, investigating the discrepancy between the social welfare optimization and customer’s self-interest maximization in a MQSCC is another direction of future research.

Acknowledgments.  
The authors are grateful to Professor Peter Glynn, Professor Kavita Ramanan, and Professor Assaf Zeevi for their valuable suggestions on this problem. The authors are thankful for the anonymous referee team to provide many valuable suggestions. The research of Yichuan Ding is partially supported by NSERC PGPIN 436156-13.

References


Appendix A: Proof of Proposition 1 We first prove that the arrival rate function $\Lambda(\tau)$ has finite derivatives everywhere except at points in $K_J$. 

We first discuss the partial derivative $\frac{\partial p_j(\tau)}{\partial \tau_i}$ with $j \neq i$. If this partial derivative exists, it must equal to the following limit

$$\lim_{t \to 0} \frac{1}{t} (p_j(\tau + t\epsilon) - p_j(\tau)).$$

A customer $\xi$ chooses queue $i$ at $\tau$, but chooses queue $j$ at $\tau + t\epsilon$, if and only if this customer has a parameter set $(u_i, c) \in S^1 \cap S^2(t)$, where

$$S^1 := \{(u, c) | u_j - c\tau_j > \max\{0, u_k - c\tau_k, \ \ \ \ k \neq i, j\}\}$$

$$S^2(t) := \{(u, c) | c(\tau_i - \tau_j) \leq u_i - u_j < c(\tau_i - \tau_j + t)\}$$

Intuitively, $\xi \in S^1$ if queue $i$ and queue $j$ are the top two choices of customer $\xi$, $\xi \in S^2$ if the expected utility of queue $i$ and queue $j$ are so close that a small change of $\tau_i$ would change the choice of customer $\xi$. The probability for $\xi \in S^1 \cap S^2(t)$ is thus exactly the difference $p_j(\tau + t\epsilon) - p_j(\tau)$.

If $\tau_i \neq \tau_j$, the limit (79) can be calculated as

$$\lim_{t \to 0} \frac{1}{t} (p_j(\tau + t\epsilon) - p_j(\tau))$$

$$= \lim_{t \to 0} \frac{1}{t} \Pr((u, c) \in S(t))$$

$$= \lim_{t \to 0} \frac{1}{t} \int_{(u, c) \in S^2(t)} I((u, c) \in S_1) f(u, c) \, du \, dc$$

$$= \lim_{t \to 0} \frac{1}{t} \int_{u_i - u_j}^{u_i - u_j \tau_i / \tau_j} f_c(u) I((u, c) \in S_1) \, dc \, f(u) \, du$$

$$= \int \lim_{t \to 0} \frac{1}{t} \int_{u_i - u_j}^{u_i - u_j \tau_i / \tau_j} f_c(u) I((u, c) \in S_1) \, dc \, f(u) \, du$$

Equality (81) is due to dominated convergence. To see that, note that the term inside $[\cdot]$ has the following limit

$$\lim_{t \to 0} \frac{1}{t} \int_{u_i - u_j}^{u_i - u_j \tau_i / \tau_j} f_c(u) I((u, c) \in S_1) \, dc = f_c(u_i - u_j) I((u, u_i - u_j) \in S_1).$$

Thus, for sufficiently small $t$, the term inside $[\cdot]$ is upper bounded by $2f_c(u_i - u_j) I((u, u_i - u_j) \in S_1)$, which has finite expectation with respect to $u$ (upper bounded by the marginal pdf $f_c(\cdot)$ at $u_i - u_j \tau_i / \tau_j$).
Therefore, if \( \tau_i \neq \tau_j \), the partial derivative \( \frac{\partial p_j(\tau)}{\partial \tau_i} \), as the limit of \( \frac{1}{t}(p_j(\tau + te_i) - p_j(\tau)) \), exists and has the following expression,

\[
\frac{\partial p_j(\tau)}{\partial \tau_i} = \int I((u, \frac{u_i - u_j}{\tau_i - \tau_j}) \in S_i) f(u, \frac{u_i - u_j}{\tau_i - \tau_j}) du. \tag{84}
\]

Since the RHS of above equation is a continuous function of \( \tau_i \) and \( \tau_j \) when \( \tau_i \neq \tau_j \), the partial derivative \( \frac{\partial p_j(\tau)}{\partial \tau_i} \) must also be continuous in \( \tau_i \) and \( \tau_j \) at points where \( \tau_i \neq \tau_j \).

If \( \tau_i = \tau_j \), then we have

\[
\lim_{t \to 0} \frac{1}{t} \left( p_j(\tau + te_i) - p_j(\tau) \right) = \lim_{t \to 0} \frac{1}{t} \Pr \left( (u, c) \in S(t) \right) = \lim_{t \to 0} \int \int \left[ \frac{1}{t} \int_{u_j}^{u_j + ct} f_{u_j|u_{-j},c}(u_j) I((u, c) \in S_i) du_j \right] f_{u_{-j},c}(u_{-j}, c) dcdu_{-i} \tag{85}
\]

\[
\geq \int \int \liminf_{t \to 0} \left[ \frac{1}{t} \int_{0}^{ct} f_{u_j|u_{-j},c}(u_j + x) I((u, c) \in S_i) dx \right] f_{u_{-j},c}(u_{-j}, c) dcdu_{-i} \tag{86}
\]

\[
= \int \int cf_{u_j|u_{-j},c}(u_j) I((u, c) \in S_i) f_{u_{-j},c}(u_{-j}, c) dcdu_{-i} = \int \int cf_{u,c}(u_j, u_{-j}, c) I((u, c) \in S_i) dcdu_{-i} \tag{87}
\]

where \( u_{-i} \) represents the vector obtained by removing the \( i \)th entry from \( u \), inequality (86) follows from the Fatou’s Lemma. Since the RHS of Equation (87) can be regarded as the marginal pdf of at \( u_i = u_j \) (with an additional constraint that \( (u, c) \in S_i \)), it can be \( +\infty \). Consequently, the partial derivative \( \frac{\partial p_j(\tau)}{\partial \tau_i} \) may not exist at points with \( \tau_i = \tau_j \), and its value can be unbounded in a neighborhood of such points.

The above argument proves that if \( j \neq i \), then \( \frac{\partial p_j(\tau)}{\partial \tau_i} \) exists and is continuous except at points \( \tau \), with \( u_j = u_i \), near which \( \frac{\partial p_j(\tau)}{\partial \tau_i} \) may go to infinity. We next discuss the \( j = i \) case. Because \( \sum_{i=0}^{J} p_j(\tau) \equiv 1 \), we know that

\[
\frac{\partial p_j(\tau)}{\partial \tau_i} = - \sum_{j \neq i, j=0,1,...,J} \frac{\partial p_j(\tau)}{\partial \tau_i} \tag{88}
\]

Note that the summation at the RHS consists of \( \frac{\partial p_j(\tau)}{\partial \tau_i} \) for all \( j \neq i \) (including \( j = 0 \)). \( p_0(\tau) \) represents the proportion of customers who choose to balk, or equivalently, to join a queue indexed by \( 0 \) with expected waiting time \( \tau_0 = 0 \) and service utility \( u_0 = 0 \). Thus, \( \frac{\partial p_0(\tau)}{\partial \tau_i} \) exists and is continuous except at \( \tau \)s with \( \tau_i = 0 \). We have proved that \( \frac{\partial p_j(\tau)}{\partial \tau_i} \) (\( j \neq 0, i \)) exists and is continuous except at \( \tau \)s with \( \tau_i = \tau_j \). Therefore, \( \frac{\partial p_0(\tau)}{\partial \tau_i} \) exists and is continuous except those \( \tau \) with \( \tau_i = 0 \) or \( \tau_i = \tau_j \) for some \( j \neq 0, i \), which is exactly the set \( K^J \) defined in (3).

So far, we have proved that the arrival rate function \( p_j(\tau) \) has finite derivative everywhere except at points in \( K^J \). Next we show that even at points in \( K^J \), \( p_j(\tau) \) is continuous, though it may not have finite derivatives. Specifically,

\[
\lim_{t \to 1} p_j(\tau + te_i) - p_j(\tau) = \lim_{t \to 0} \Pr \left( (u, c) \in S(t) \right) = \lim_{t \to 0} \int \int e^{ct} f_{u_j|u_{-j},c}(u_j + x) I((u, c) \in S_i) du_j f_{u_{-j},c}(u_{-j}, c) dcdu_{-i} = 0. \tag{89}
\]
where equality (89) follows from Equation (85), and equality (90) follows from that
\[
\lim_{t \to 0} \int_{0}^{t} f_{u_{ij} \mid u_{i-1} = c(u_{j} + x)} I((u_{i}, c) \in S_{1}) \, du_{i} = 0.
\]
We may repeatedly apply the above logic for each coordinate \( i \neq j \) and establish continuity of \( p_{j}(\tau) \) at points in \( K^{J} \).

We next prove properties (a)-(e) in Proposition 1.

(a) Suppose \( \tau_{k}^{i} > \tau_{k}^{j} \), and \( \tau_{k}^{2} = \tau_{k}^{1} \) for \( j \neq k \). For a customer indexed by \( \xi \), if his choice is queue \( j \neq k \), then
\[
\tau_{k}^{i} - c\tau_{k}^{j} < u_{k} - c\tau_{k}^{1} \leq u_{j} - c\tau_{j}^{1} = \tau_{j}^{1} - c\tau_{j}^{2},
\]
where the first inequality is due to \( \tau_{k}^{i} > \tau_{k}^{j} \), the second inequality follows from the fact that the customer’s optimal choice is queue \( j \) instead of queue \( k \), and the last equality follows from \( \tau_{j}^{1} = \tau_{j}^{2} \). Therefore, if a customer’s initial choice is queue \( j \), then his choice remains the same when the waiting-time vector is changed from \( \tau_{1} \) to \( \tau_{2} \). We thus deduce that \( p_{j}(\tau) \) is non-decreasing in \( \tau_{k} \).

(b) Note that \( p_{j}(\tau) \) must be non-increasing with \( \tau_{j} \), as a result of (a) and \( \sum_{k=0}^{J} p_{k} = 1 \). So it suffices to prove \( p_{j}(\tau) \) is strictly decreasing when \( \tau_{1} \) has been replaced by \( \tau_{2} \), where \( \tau_{j}^{1} > \tau_{j}^{2} \) but \( \tau_{k}^{2} = \tau_{k}^{1} \) for \( k \neq j \). A customer \( \xi \) will choose to join queue \( j \) given expected waiting-times vector \( \tau_{1} \), but not join queue \( j \) when the waiting-time vector is changed to \( \tau_{2} \), if and only if
\[
(u_{\xi}, c_{\xi}) \in \{ (u, c) \mid \begin{align*}
&u_{j} - c\tau_{j}^{1} > \max\{0, u_{k} - c\tau_{k}^{1}, \ k \neq j\}, \ k \neq j \ \\
&u_{j} - c\tau_{j}^{2} < \max\{0, u_{k} - c\tau_{k}^{2}, \ k \neq j\}
\end{align*}\}
\]
Because the parameter \( c \) has positive conditional pdf \( f_{c \mid u} \) over \( \mathbb{R}_{+} \), the above set must have a positive probability mass. Therefore, a positive proportion of customers must switch to queues other than \( j \) when the waiting time of queue \( j \) has been increased from \( \tau_{j}^{1} \) to \( \tau_{j}^{2} \). Therefore, \( p_{j}(\tau) \) is strictly decreasing in \( \tau_{j} \).

(c) Equation (81) implies that
\[
\frac{\partial p_{j}(\tau)}{\partial \tau_{i}} = \lim_{t \to 0} \frac{1}{t} (p_{j}(\tau + te_{i}) - p_{j}(\tau))
= \lim_{t \to 0} \frac{1}{t} \Pr \left( \begin{array}{c}
(u_{j} - c\tau_{j}^{1} > \max\{0, u_{k} - c\tau_{k}^{1}, \ k \neq i, j\}) \\
(u_{j} - c\tau_{j}^{2} < \max\{0, u_{k} - c\tau_{k}^{2}, \ k \neq j\})
\end{array} \right)
\]
Similarly,
\[
\frac{\partial p_{i}(\tau)}{\partial \tau_{j}} = \lim_{t \to 0} \frac{1}{t} (p_{i}(\tau) - p_{i}(\tau - te_{j}))
= \lim_{t \to 0} \frac{1}{t} \Pr \left( \begin{array}{c}
(u_{i} - c\tau_{i} > \max\{0, u_{k} - c\tau_{k}^{1}, \ k \neq i, j\}) \\
(u_{i} - c\tau_{i} < \max\{0, u_{k} - c\tau_{k}^{2}, \ k \neq j\})
\end{array} \right)
\]
Notice that the set at the RHS of Equation (93) and (94) are identical. The intuition is that it is the same group of customers who will switch to queue \( j \), when either \( \tau_{j} \) has been decreased by \( t \), or \( \tau_{i} \) has been increased by \( t \). Property (c) then follows immediately.

(d) Given \( \tau_{2} = \tau_{1} + te \), the linear form of \( U_{\xi} \) implies that if \( U_{\xi,j} \geq U_{\xi,k} \) for all \( k \neq j \) (including \( k = 0 \)) at \( \tau_{2} \), then the same inequalities must hold at \( \tau_{1} \). Therefore, we deduce that \( p_{j}(\tau_{2}) \geq p_{j}(\tau_{1}) \) for all \( j \neq 0 \). To prove the strict inequality in (12), we notice that customer \( \xi \) joins some queue at \( \tau \), but balks at \( \tau_{2} \) if
\[
(u, c) \in \left\{ u, c \mid \begin{align*}
0 < &\max\{u_{k} - c\tau_{k}^{1}, \ k = 1, \ldots, J\} \\
0 > &\max\{u_{k} - c(\tau_{k}^{1} + t), \ k = 1, \ldots, J\}
\end{align*}\}
\]
Because the parameter \( c \) has positive conditional pdf \( f_{c \mid u} \) over \( \mathbb{R}_{+} \), the above set must have a positive probability mass, so the strict inequality (12) is proved.
(e) A customer will joint queue $j$ only if $u_j - c\tau > 0$. Therefore, when $\tau_j \to \infty$,

$$p_j(\tau) \leq \Pr(u_j - c\tau > 0) = \int_0^{u_j/\tau_j} \int f_{\tau|u}(c) dc \, f_u(du) \to 0.$$  \hspace{1cm} (96)

where the equality follows from bounded convergence. Equation (96) leads to property (e).

\section*{Appendix B: Examples when properties of arrival rate function are violated}

Here we provide some examples in which the MQSCC do not have the desired characteristics (existence and uniqueness of equilibrium, convergence of the fluid limit process to the equilibrium, convergence to the diffusion limit, and a closed-form characterization of the steady-state distribution of the diffusion limit) without some of the properties (a)-(e).

The first example shows that the MQSCC may not have an equilibrium without property (e).

\textbf{Example B.1} Consider an MQSCC with arrival rate function $\Lambda(\tau) = (0.4 + 0.1 \exp(-\tau_1), 0.4 + 0.1 \exp(-\tau_2)$, and $\mu = (0.3, 0.3)^T$. One can verify that this arrival rate function and its Jacobean matrix $R(\tau) = \begin{pmatrix} -0.1 \exp(-\tau_1) & 0 \\ 0 & -0.1 \exp(-\tau_2) \end{pmatrix}$ satisfy all properties in Proposition 1 except for property (e). Because the arrival rates for both queues are lower bounded by $0.4$, which is strictly larger than the service rate, both queues have to grow to infinitely large and the MQSCC has no equilibrium.

The second example shows that property (d) is necessary to guarantee the uniqueness of the equilibrium.

\textbf{Example B.2} Consider an MQSCC has $R(\tau) \equiv \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}$, $\mu = (0.5, 0.5)^T$, $\Lambda(\tau) = (0.5, 0.5)^T + R\tau$. Then any vector in the form of $(z, z)^T$ with $z \geq 0$ can be an equilibrium queue-length vector for this MQSCC.

The third example shows that without properties (a) and (b), then even if the MQSCC has a unique equilibrium, the fluid limit process may not converge to that equilibrium.

\textbf{Example B.3} Consider an MQSCC has arrival rate function $\Lambda(\tau) = \left(\frac{\tau_1^2}{2}, \frac{2-\tau_1}{2}\right)$, $\mu = (0.5, 0.5)^T$. The Jacobean matrix of $\Lambda(\tau)$ is $R(\tau) = \begin{pmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{pmatrix}$, which violates properties (a) and (b). This fluid limit process has a unique equilibrium state $\tau^* = (1, 1)^T$. However, if we start from $\tau = (0, 1)^T$, the fluid process $\tau(t)$ will move along the circle $(\tau_1 - 1)^2 + (\tau_2 - 1)^2 = 1$ and never converges to the equilibrium.

Finally, property (c) is needed to ensure that the row diagonal dominance (property (d)) of the Jacobean matrix also implies column diagonal dominance, and thus negative definiteness, and that the steady-state distribution of the diffusion process has a closed form (see Proposition 3). Property (c) is not a necessary condition for uniqueness and existence of the equilibrium as it has not been used in our proof of Theorem 2.
Appendix C: Proof of Lemma 2  Suppose \( x \) and \( y \) are both solutions to SDER (25). Then by the first equation in the proof of Theorem 4.1 in (Tanaka [53], page 175), we have
\[
\| x(t) - y(t) \|^2 \\
\leq \| \int_0^t (\sigma(s, x(s)) - \sigma(s, y(s))) dB(s) \|^2 + 2 \int_0^t \langle x(s) - y(s), b(s, x(s)) - b(s, y(s)) \rangle ds + \text{the remainder.}
\] (97)
where the remainder has zero expectation. We thus have
\[
E \| x(t) - y(t) \|^2 \\
\leq E \int_0^t \| \sigma(s, x(s)) - \sigma(s, y(s)) \|^2 ds + 2 \int_0^t \langle x(s) - y(s), b(s, x(s)) - b(s, y(s)) \rangle ds
\] (98)
where the inequality follows from Lipschitz continuity of \( \sigma(s, \cdot) \). By absolute continuity of \( b(s, \cdot) \), we have
\[
b(s, x(s)) - b(s, y(s)) = \int_0^1 R(y(s) + \xi(x(s) - y(s))) (x(s) - y(s)) d\xi,
\] (99)
with the Jacobean matrix \( R(y(s) + \xi(x(s) - y(s))) \) negative definite for almost all \( \xi \in [0, 1] \). Consequently,
\[
\langle x(s) - y(s), b(s, x(s)) - b(s, y(s)) \rangle = \langle x(s) - y(s), \int_0^1 R(y(s) + \xi(x(s) - y(s)))(x(s) - y(s)) d\xi \rangle
\] (100)
\[
= \int_0^1 \langle x(s) - y(s), R(y(s) + \xi(x(s) - y(s)))(x(s) - y(s)) d\xi \rangle
\] (101)
which, together with Equation (98), leads to
\[
E \| x(t) - y(t) \|^2 \leq K^2 \int_0^t E \| x(s) - y(s) \|^2 ds.
\] (101)
Then by the Gronwall’s inequality (e.g., [12], page 498), we have \( \| x(t) - y(t) \| = 0 \).

Appendix D: Proof of Lemma 4  The main idea of the proof is to express \( Q^{n,n} \) as the image of a generalized reflection mapping and then to apply the continuous mapping theorem to establish the convergence. In view of (26), we can express \( Q^n \) as
\[
Q^n(t) = n^{-1/2}(X^n(t) - \tau^* \mu^*)
\] (102)
By imposing the reflection barrier on the finite non-zero boundary of \( \Omega(\kappa) \), we derive the expression of \( Q^{n,n} \) as
\[
Q^{n,n}_j(t) = \Phi^{Q^n}(Q^n)
\] (103)
\[
= Q^{n,n}_j(0) + n^{-1/2} \mathcal{N} \left( n \int_0^t p_j(\tau^{n,n}(s)) ds \right) - n^{-1/2} S^n_j(t) + n^{-1/2} L^n_j(t)
\] (104)
By imposing the reflection barrier on the finite non-zero boundary of \( \Omega(\kappa) \), we derive the expression of \( Q^{n,n} \) as
\[
Q^{n,n}_j(t) = \Phi^{Q^n}(Q^n)
\] (105)
\[
= Q^{n,n}_j(0) + n^{-1/2} \mathcal{N} \left( n \int_0^t p_j(\tau^{n,n}(s)) ds \right) - n^{-1/2} S^n_j(t) + n^{-1/2} L^n_j(t)
\] (106)
By imposing the reflection barrier on the finite non-zero boundary of \( \Omega(\kappa) \), we derive the expression of \( Q^{n,n} \) as
\[
Q^{n,n}_j(t) = \Phi^{Q^n}(Q^n)
\] (107)
\[
= Q^{n,n}_j(0) + n^{-1/2} \mathcal{N} \left( n \int_0^t p_j(\tau^{n,n}(s)) ds \right) - n^{-1/2} S^n_j(t) + n^{-1/2} L^n_j(t)
\] (108)
By imposing the reflection barrier on the finite non-zero boundary of \( \Omega(\kappa) \), we derive the expression of \( Q^{n,n} \) as
\[
Q^{n,n}_j(t) = \Phi^{Q^n}(Q^n)
\] (109)
\[
= Q^{n,n}_j(0) + n^{-1/2} \mathcal{N} \left( n \int_0^t p_j(\tau^{n,n}(s)) ds \right) - n^{-1/2} S^n_j(t) + n^{-1/2} L^n_j(t)
\] (110)
where \( \tau^\kappa,s(n) = (n^{1/2}Q^\kappa,s(n) + n\tau^* \circ \mu^*) \circ (n\mu^n)^{-1} \) denotes the waiting-time vector at time \( s \) in the \( n \)-th MQSCC, \( L^n := (L^n_s) \) and \( U^n := (U^n_s) \) are the minimal non-decreasing processes which ensure that the corresponding \( Q^\kappa,n \) stays in \( \Omega(\kappa) \), \( Z^\kappa,n := (Z^\kappa,n_j) \) is the centered process associated with \( Q^\kappa,n \), such that

\[
Z^\kappa,n_j(t) := \left( \mathcal{N} \left( \int_0^t p_j(\tau^\kappa,n(s)) \, ds \right) - \int_0^t np_j(\tau^\kappa,n(s)) \, ds \right) + \left( n\mu^n_t - S^\kappa(n) \right).
\]

Note that in the expression of (103), we have used \(-n^{-1/2}S^\kappa_j(t)\) instead of \(-n^{-1/2}S^\kappa_j(W^\kappa,n_j(t))\) because the loss of service capacity has already been absorbed into \( L^n_j(t) \).

We next analyze the terms labeled as (A.1)-(A.3) in (103).

1. By our assumption of the initial value, (A.1) \( \Rightarrow Y^\kappa(0) \), where \( Y^\kappa(0) = \Phi^{\Omega(\kappa)}(Y(0)) \).

2. To estimate (A.2), we first estimate the term \( n^{1/2}(p_j(\tau^\kappa,n(s)) - \mu^*_j) \) at any time \( s \). Define

\[
\Delta \tau^\kappa,n(s) := (n^{1/2}Q^\kappa,n(s) + n\tau^* \circ \mu) \circ (n\mu^n)^{-1} - \tau^\kappa,n^\ast.
\]

We have

\[
n^{1/2}(p_j(\tau^\kappa,n(s)) - \mu^*_j) = n^{1/2}(p_j((n^{1/2}Q^\kappa,n(s) + n\tau^* \circ \mu) \circ (n\mu^n)^{-1}) - \mu^*_j) \\
= n^{1/2}(p_j(\tau^\kappa,n^\ast + \Delta \tau^\kappa,n(s)) - p_j(\tau^\kappa,n^\ast)) + n^{1/2}(p_j(\tau^\kappa,n^\ast) - \mu^*_j) \\
\rightarrow n^{1/2}(p_j(\tau^\kappa,n^\ast + \Delta \tau^\kappa,n(s)) - p_j(\tau^\kappa,n^\ast)) - \theta_j,
\]

and

\[
n^{1/2}\Delta \tau^\kappa,n(s) \rightarrow n^{1/2}(Q^\kappa,n(s) \circ (n\mu^n)^{-1} + (\tau^* \circ \mu - \tau^\kappa,n^\ast \circ \mu^n) \circ (\mu^n)^{-1}) \\
\rightarrow (Q^\kappa,n(s) - \theta) \circ \mu^{-1}.
\]

Given that \( \sup_n \|Q^\kappa,n(s)\| \leq \kappa \), Equation (107) implies that \( n^{1/2}\|\Delta \tau^\kappa,n\| \) is uniformly upper bounded for any \( t > 0 \). Thus, to estimate \( n^{1/2}(p_j(\tau^\kappa,n^\ast + \Delta \tau^\kappa,n(s)) - p_j(\tau^\kappa,n^\ast)) \), it suffices to expand the Taylor series until the first-order term, because the higher-order terms uniformly converge to zero on the compact set \([0, t] \) when \( n \to \infty \). Following this logic, we get

\[
\int_0^t n^{1/2}(p_j(\tau^\kappa,n(s)) - \mu^*_j) \, ds \\
= \int_0^t n^{1/2}(p_j(\tau^\kappa,n^\ast + \Delta \tau^\kappa,n(s)) - p_j(\tau^\kappa,n^\ast)) + n^{1/2}(p_j(\tau^\kappa,n^\ast) - \mu^*_j) \, ds \\
= \int_0^t \nabla p_j(\tau^\kappa,n^\ast) \cdot (n^{1/2}\Delta \tau^\kappa,n(s)) \, ds + o(n^{1/2}\|\Delta \tau^\kappa,n\|_2) + \int_0^t n^{1/2}(p_j(\tau^\kappa,n^\ast) - \mu^*_j) \, ds \\
\rightarrow \int_0^t \nabla p_j(\tau^\kappa,n^\ast) \cdot (Q^\kappa,n(s) - \theta) \circ \mu^{-1} - \theta_j \, ds \\
= \int_0^t \left( \sum_i \frac{Q_i^\kappa,n(s) - \theta_i}{\mu_i} R_i^\kappa,n \right) \, ds
\]

where the convergence follows from continuity of \( \nabla p_j(\cdot) \) at \( \tau^* \), Equation (107), and the bounded convergence theorem. (108) implies that (A.2) \( \rightarrow 0 \) uniformly on any compact set \([0, t] \).

3. \( Z^\kappa,n(t) \) is the sum of two centered processes. The first centered process represents the cumulative difference between the Poisson arrival and its average up to time \( t \). Define the cumulative arrival process \( A^\kappa,n(t) := \int_0^t p_j(\tau^\kappa,n(s)) \, ds \). (108) implies that \( A^\kappa,n(t) \to \mu^*_j t \to \mu_j t \) uniformly on any compact set \([0, T] \). Moreover, for each \( j \), \( A^\kappa,n \) is a non-decreasing process with \( A^\kappa,n(0) = 0 \). By invoking the functional central limit theorem (FCLT) (see Theorem 5.11 in [5]) and the random time-change theorem (Theorem 5.3 in [5]), we have

\[
n^{1/2} \left( \mathcal{N}(nA^\kappa,n_j(t)) - nA^\kappa,n_j(t) \right) \Rightarrow B_j(\mu_j t),
\]

where \( B(t) := \{B_j(t)\} \) denotes a \( J \)-dimensional standard Brownian motion with covariance matrix \( I \) (identity matrix).
For the second process in the expression (104) for $Z^{k,n}(t)$, we first argue that for every $j \in \mathcal{J}$, $W^{k,n}_j(t) \rightarrow t$. If $j \in \mathcal{J}^+\cup\mathcal{J}^-$, it is straightforward to verify that $W^{k,n}_j(t) \rightarrow t$; if $j \in \mathcal{J}^-$, by looking into the fluid limit process $x^{k,n} = \frac{1}{n}X^{k,n}$ and using a similar argument in the proof of $\frac{1}{n}L^n(t) \rightarrow 0$ for Theorem 1, we deduce that

$$\frac{1}{n}L^{k,n}_j(t) = \mu_j(t - W^{k,n}_j(t)) \rightarrow 0,$$

which indicates that $W^{k,n}_j(t) \rightarrow t$.

Given that, using the random time-change theorem and FCLT, we deduce that

$$n^{-1/2}(n\mu_j W_j^{k,n}(t) - S_j^{k,n}(W_j^{k,n}(t))) \Rightarrow B_j^S(\mu_j t),$$

(110)

where $B^S = (B^S_j)$ is a $J$-dimensional standard Brownian motion whose covariance matrix is diagonal and its $j$-th diagonal entry is $c^2_{s,j}$. By summing up (109) and (110), we get

$$n^{-1/2}Z^{k,n}(t) \Rightarrow \Sigma B(t),$$

(111)

where $\Sigma$ is a diagonal matrix and has $\sqrt{(1+c^2_{s,j})\mu_j}$ as its $j$-th diagonal entry, and $B$ still represents a $J$-dimensional standard Brownian motion with covariance matrix $I$.

Given the analysis above, we deduce that

$$(A.1) + (A.2) + (A.3) \Rightarrow Y_j^k(0) + 0 + \sqrt{(1+c^2_{s,j})\mu_j}B_j(t)$$

(112)

We can express $Q_j^{k,n}(t)$ using the generalized reflection mapping as we defined before Lemma 1,

$$Q_j^{k,n}(t) = (A.1) + (A.2) + (A.3) + \int_0^t \Gamma_j^{k,n}(Q_j^{k,n}(s))ds + n^{-1/2}L_j^n(t) - n^{-1/2}U_j^n(t)$$

(113)

where $\Gamma_j^{k,n}(z) := \sum_{i \in \mathcal{J}} R^{k,n}_i(z_i - \theta_j) - \theta_j, \bar{L} = \frac{1}{\sqrt{n}}L^n$ and $\bar{U} = \frac{1}{\sqrt{n}}U^n$ are the minimal non-decreasing processes which ensure $Q_j^{k,n}$ to lie in $\Omega(\kappa)$, and converge to zero with respect to the $||\cdot||_T$ norm when $n \rightarrow \infty$. Since $\Gamma^* := (\Gamma^*_j)$ is affine and thus has to be Lipschitz continuous, we may invoke Lemma 1 and deduce that the mapping $\Phi^{(\kappa)}_{\Gamma^*}$ is continuous. Therefore, Equation (112) and the continuous mapping theorem (Theorem 5.2 in [5]) leads to

$$Q_j^{k,n}(t) = \Phi^{(\kappa)}_{\Gamma^*}((A.1) + (A.2) + (A.3)) \Rightarrow \Phi^{(\kappa)}_{\Gamma^*}(Y^k(0) + \Sigma B(t)) = \Phi^{(\kappa)}_{\Gamma^*}(Y(0) + \Sigma B(t)).$$

(114)

Let $\tilde{\gamma}^k := \Theta^{(\kappa)}_{\Gamma^*}(Y^k(0) + \Sigma B(t))$. By the definition of $\Theta^{(\kappa)}_{\Gamma^*}$, $\tilde{\gamma}^k$ solves the following integration equation,

$$\tilde{\gamma}^k(t) := Y(0) + \Sigma B(t) + \int_0^t \Gamma^*(\Phi^{(\kappa)}_{\Gamma^*}(\tilde{\gamma}^k)(s))ds$$

(115)

By comparing the integral equation (115) to the one in the statement of Equation 4 (Equation (50)), it is straightforward to show that $\tilde{\gamma}^k(t)$ solves (115) if and only if $Y^k = \tilde{\gamma}^k + L$ ($L$ is the minimal non-decreasing process which enforces $\tilde{\gamma}^k \geq 0$ when $j \in \mathcal{J}^- \cup \mathcal{J}^+$) solves Equation (50). Therefore, we have $Y^k := \Phi^{(\kappa)}(Y) = \Phi^{(\kappa)}(\tilde{\gamma}^k) = \Phi^{(\kappa)}(\Theta^{(\kappa)}_{\Gamma^*}(Y^k(0) + \Sigma B(t))) = \Phi^{(\kappa)}_{\Gamma^*}(Y(0) + \Sigma B(t))$. Thus, (114) actually implies that $Q_j^{k,n} \Rightarrow Y^k$. □
Appendix E: Proof of Lemma 5  Similar to $\Delta \tau^{n^\circ}(s)$ in Equation (105), we define $n^{1/2}\Delta \tau^n(s)$ for a given $Q^n(s)$ as

$$n^{1/2}\Delta \tau^n(s) = n^{1/2}(n^{1/2}Q^n(s) + n\tau^\circ \circ \mu)(n\mu^\circ)^{-1} - \tau^n$$

$$= Q^n(s)(\mu^n)^{-1} + (\tau^\circ \circ \mu - \tau^n)(\mu^n)^{-1}$$

(116)

We have shown in Equation (107) that the second term at the RHS of (116) converges to $-\vartheta \circ \mu^{-1}$, so the second term must be bounded for all $n$. Also, the sequence $\{\mu^n\}$ is bounded as it converges to $\mu$. Thus, there exists $\epsilon > 0$, such that for sufficiently large $n$,

$$n^{1/2}\Delta \tau^n_j(s) + \frac{\delta_j}{\mu_j} - \epsilon \leq \frac{Q^n_j(s)}{\mu_j} \leq n^{1/2}\Delta \tau^n_j(s) + \frac{\delta_j}{\mu_j} + \epsilon,$$

(117)

which implies that $n^{1/2}\Delta \tau^n(s)$ is bounded if and only if $Q^n(s)$ is bounded. We let $\Delta \tau^n(t)$ and $\Delta \tau^n(t)$ denote the maximal and minimal entry in the vector $\Delta \tau^n(t)$, respectively. To prove Lemma 5, it suffices to prove that for any fixed $T > 0$, when $\kappa \to \infty$,

$$\limsup_n \Pr(\{n^{1/2}\Delta \tau^n(t) \mid t \in [0, T]\} > \kappa) \to 0$$

$$\liminf_n \Pr(\{n^{1/2}\Delta \tau^n(t) \mid t \in [0, T]\} < -\kappa) \to 0$$

(118)

To prove (118), we first derive an expression for $Q^n$ in analogue to the expression for $Q^{n^\circ}$ (103) by ignoring the reflection barrier at $\pm \kappa$,

$$Q^n_j(t) = Q^n_j(0) + \int_0^t \Gamma^n_j(\tau^n + ^\Delta \tau^n(s))ds + n^{-1/2}Z^n_j(t) + n^{-1/2}L^n(t),$$

(119)

where $\Delta \tau^n(s)$ is defined as in (116) for a given $Q^n(s)$, $\Gamma^n_j(\tau) := n^{1/2}(p_j(\tau) - \mu^n_j)$ represents the deterministic drift that can be non-Lipschitz, and $Z^n_j(t)$ represents a mean-zero stochastic process which was defined in Equation (27).

We next consider the scenario when $n^{1/2}\Delta \tau^n(s) = n^{1/2}(\tau^n_j(s) - \tau^n_j^\circ) > \delta$ in some interval $[a_1, b_1]$ and for some fixed $j^\circ \in \{j = 1, \ldots, J\}$. That means, $\tau^n$ has the largest positive deviation from the equilibrium $\tau^n_j^\circ$ along its $j^\circ$-th dimension over $[a_1, b_1]$. We can then prove by the mean-reversion property of $\Gamma^n$, that over $[a_1, b_1]$, the drift term would be upper bounded by a negative constant (See (123)), and consequently the deviation $\Delta \tau^n(s)$ would decrease by at least an amount proportional to $b_1 - a_1$ (See (127)).

Formally, we have

$$\Gamma^n_j(\tau^n + ^\Delta \tau^n(s)) = n^{1/2}(p_j(\tau^n + ^\Delta \tau^n(s)) - \mu^n_j)$$

$$= n^{1/2}(p_j(\tau^n + ^\Delta \tau^n(s)) - p_j(\tau^n)) + n^{1/2}(p_j(\tau^n) - \mu^n_j)$$

(120)

We next provide an upper bound for the RHS of Equation (120). In inequality (43) (which builds on the mean-reversion property (a)(b)(d) in Proposition 1), by replacing $\tau(t)$ with $\tau + ^\Delta \tau^n(s)$, and by noting that $\Delta \tau^n(s) \geq n^{1/2}/\delta$, we get

$$p_j(\tau^n + ^\Delta \tau^n(s)) - p_j(\tau^n) \leq -n^{1/2}/\delta,$$

(121)

where $h^n(\cdot)$ follows a similar functional form of $h(\cdot)$ as given in Equation (45), that is,

$$h^n(\delta) := \frac{\delta}{p_j(\tau^n)} \left( \max\{z \in [0, n^{-1/2}/\delta] \mid R_j^\circ, (\tau^n + z\epsilon) + \sum_{i \neq j^\circ} R_j^\circ, (\tau^n + z\epsilon) \} \right) > 0$$

(122)

Inequality (121) allows us to upper bound the RHS of (120) as

$$\Gamma^n_j(\tau^n + ^\Delta \tau^n(s)) \leq -h^n(\delta) + n^{1/2}(p_j(\tau^n) - \mu^n_j)$$

$$\leq \frac{\delta}{p_j(\tau^n)} \left( R_j^\circ, (\tau^n) + \sum_{i \neq j^\circ, j} R_j^\circ, (\tau^n) - \theta_j \right)$$

(123)
That means, for sufficiently large $n$,

$$
\Gamma^n_j(\tau^{n,*} + \Delta \tau^n(s)) < \frac{\delta}{p_j(\tau^{n,*})} \left( R_{j^*}(\tau^{n,*}) + \sum_{i \neq j^*} R_{j^*}(\tau^{n,*}) \right) - \theta_{j^*} := -\Delta_n < 0
$$

(124)

where $R_{j^*}(\tau^{n,*}) + \sum_{i \neq j^*} R_{j^*}(\tau^{n,*}) < 0$ by properties (a)(b)(d) in Proposition 1. By looking into the sequence $\{\Delta_n\}$, we deduce that it converges to some positive constant, $\Delta > 0$. Inequality (119) and (124) imply that

$$
Q^n_{j^*}(b_1) - Q^n_{j^*}(a_1) \leq -\Delta_n(b_1 - a_1) + n^{-1/2}(Z^n_{j^*}(b_1) - Z^n_{j^*}(a_1)) + n^{-1/2}(L_{j^*}(b_1) - L_{j^*}(a_1))
$$

(125)

If $j^* \in J^- \cup J^+$, then $\tau^n_j(s) - \tau^n_{j^*} > 0$ implies that $Q^n_{j^*}(s) > 0$ over $[a_1, b_1)$. Consequently, $L_{j^*}(b_1) - L_{j^*}(a_1) = 0$. Thus in either case, $L_{j^*}(b_1) - L_{j^*}(a_1) = 0$ and inequality (125) implies that

$$
Q^n_{j^*}(b_1) - Q^n_{j^*}(a_1) \leq -\Delta_n(b_1 - a_1) + n^{-1/2}(Z^n_{j^*}(b_1) - Z^n_{j^*}(a_1)).
$$

(126)

which leads to

$$
n^{1/2}(\Delta \tau^n(b_1) - \Delta \tau^n(a_1)) = n^{1/2}(\tau^n_{j^*}(b_1) - \tau^n_{j^*}(a_1))
$$

$$
= \frac{1}{\mu_j} (Q^n_{j^*}(b_1) - Q^n_{j^*}(a_1)).
$$

(127)

$$
\leq \frac{1}{\mu_j} (-\Delta_n(b_1 - a_1) + n^{-1/2}(Z^n_{j^*}(b_1) - Z^n_{j^*}(a_1)))
$$

That means, the largest deviation $\Delta \tau^n$ keeps decreasing. For any interval $[a, b] \subseteq [0, T]$ over which $n^{1/2}\Delta \tau^n \geq \delta$, we can partition $[a, b]$ into countably many intervals $\cup^\infty_{i=1}[a_i, b_i]$ such that $\Delta \tau^n(s) = \tau^n_{j^*}(s) - \tau^n_{j^*}$ for the same index $j^* \in \{1, 2, \ldots, J\}$ and for all $s \in [a_i, b_i)$. Using this notation, we derive the following inequality

$$
n^{1/2}(\Delta \tau^n(b) - \Delta \tau^n(a)) = \sum_{i=1}^{\infty} n^{1/2}(\Delta \tau^n(b_i) - \Delta \tau^n(a_i)) \leq \sum_{i=1}^{\infty} \frac{1}{\mu_j} \left( -\Delta_n(b_i - a_i) + n^{-1/2}(Z^n_{j^*}(b_i) - Z^n_{j^*}(a_i)) \right)
$$

(128)

Now let $\delta = \frac{\xi}{2}$. If $\Delta \tau^n(\cdot)$ has ever exceeded $\frac{\xi}{2}$ over $[0, t]$, then we let $a = \sup\{s \in [0, t] : \Delta \tau^n(s) \leq \frac{\xi}{2}\}$ and $b = t$. The selection of $a$ and $b$ guarantees that $\Delta \tau^n(a) = \frac{\xi}{2}$ and $\Delta \tau^n(s) \geq \frac{\xi}{2}$ for all $s \in [a, b]$. Thus, Equation (128) implies that

$$
n^{1/2}(\Delta \tau^n(t) - \frac{\xi}{2}) = n^{1/2}(\Delta \tau^n(b) - \Delta \tau^n(a)) \leq \frac{1}{\min_j \mu_j} (n^{-1/2}\|Z^n\|_T).
$$

(129)

If $\Delta \tau^n(\cdot)$ is always upper bounded by $\frac{\xi}{2}$ over $[0, t]$, then the above inequality holds trivially. We thus have

$$
n^{1/2} \sup\{\Delta \tau^n(t) \mid t \in [0, T]\} \leq \frac{\xi}{2} + \frac{1}{\min_j \mu_j} n^{-1/2} \sup\{\|Z^n(t)\| \mid t \in [0, T]\}
$$

(130)

\[6\] To derive (129), we have only used a weaker upper bound (128) for $\Delta \tau^n(b) - \Delta \tau^n(a)$ by ignoring the negative drift $-\Delta_n(b - a)$. The original upper bound (128) including $-\Delta_n(b - a)$, however, is needed in the later proof for Proposition 4.
When \( \kappa \to \infty \), we deduce that

\[
\limsup_n \Pr(\sup\{n^{1/2}\Delta t^n(t) \mid t \in [0,T]\} > \kappa) \\
\leq \limsup_n \Pr(\sup\{n^{1/2}\Delta t^n(t) \mid t \in [0,T]\} > \kappa \mid n^{1/2}\Delta t^n(0) \leq \frac{\delta}{2}) \Pr(n^{1/2}\Delta t^n(0) \leq \frac{\delta}{2}) \\
+ \limsup_n \Pr(n^{1/2}\Delta t^n(0) > \frac{\delta}{2}) \\
\to \limsup_n \Pr(\sup\{n^{1/2}\Delta t^n(t) \mid t \in [0,T]\} > \kappa \mid n^{1/2}\Delta t^n(0) \leq \frac{\delta}{2}) \cdot 1 + 0 \quad (131)
\]

for some positive constants \( c_i \), \( i = 1, 2, 3, 4 \). In Equation (131), the convergence result follows from \( \limsup_n \Pr(n^{1/2}\Delta t^n(0) > \frac{\delta}{2}) \to 0 \) as \( Q^n(0) \) (so \( n^{1/2}\Delta t^n(0) \)) is assumed to have finite expectation; the second inequality follows from (130), and the last inequality follows from the upper bound (145) for the tail probability of \( n^{-1/2}\|Z^n\|_T \) (See Lemma 6 in Appendix G). Note that the second term of RHS in Equation (131) is dominated by \( \exp(-\frac{\delta}{2}\kappa\sqrt{n}) \) when \( n \) is large, so the RHS has to converge to zero when \( \kappa \to \infty \), which leads to the first convergence equation in (118).

The second convergence in (118) can be proved using an analogous argument and is omitted here.

**Appendix F: Proof of Proposition 4** Equation (117) implies that when \( n \) is sufficiently large, the difference between \( V(\Xi^n(t)) = \|Q^n\|^{\mu_1} \) and \( \|n^{1/2}\Delta t^n(t)\| \) is almost a constant (i.e., within \( \pm \epsilon \)). So proving Equation (61) is equivalent to proving the same bounded condition for \( \|n^{1/2}\Delta t^n(t)\| \), that is, for some \( \epsilon_0 > 0, t_0 \geq 0, \)

\[
\limsup_{n \to \infty} \sup_{\Xi^n(t) \in \Omega} \mathbb{E}[\exp(u_0(\|n^{1/2}\Delta t^n(t_0)\| - \|n^{1/2}\Delta t^n(0)\|)^{\pm}) \mid \Xi^n(0) \] < \infty \\
\limsup_{n \to \infty} \sup_{\Xi^n(t) \in \Omega} \mathbb{E}[\exp(u(\|n^{1/2}\Delta t^n(t_0)\| - \|n^{1/2}\Delta t^n(0)\|)^{\pm}) \mid \Xi^n(0) \] < \infty \quad (132)
\]

To prove (132), we first consider the case when \( \|n^{1/2}\Delta t^n(s)\| > \frac{\kappa}{2} \) for all \( s \in [0,T] \). By Equation (128) (which builds on the mean-reversion properties (a)(b)(d) of the arrival rate) and by plugging into \( a = 0 \) and \( b = t_0 \), we have

\[
n^{1/2}\|\Delta t^n(t)\| - n^{1/2}\|\Delta t^n(0)\| \leq \frac{1}{\min_j \mu_j} \left( -\Delta_n t_0 + n^{-1/2}\|Z^n(t_0)\| \right) \quad (133)
\]

where \( \Delta_n \) was defined in (124), which converges to a positive constant \( \Delta > 0 \). By choosing

\[
t_0 = \frac{\min_j \mu_j}{\Delta} \left( n^{1/2}\|\Delta t^n(0)\| - \frac{\kappa}{2} \right)^+, \quad (134)
\]

for sufficiently large \( n \), Equation (133) implies that

\[
n^{1/2}\|\Delta t^n(t)\| \leq \frac{\kappa}{2} + \frac{1}{\min_j \mu_j}n^{-1/2}\|Z^n(t_0)\|. \quad (135)
\]

In the other case when \( \|n^{1/2}\Delta t^n(s)\| \leq \frac{\kappa}{2} \) for some \( s \in [0,T] \), we can also deduce (135) using a similar argument as we establish inequality (130) in the proof for Lemma 5.

In view of (135), we deduce that there exists \( \epsilon_0 > 0 \) such that

\[
\limsup_{n \to \infty} \sup_{\Xi^n(t) \in \Omega} \mathbb{E}[\exp(u_0(\|n^{1/2}\Delta t^n(t_0)\| - \|n^{1/2}\Delta t^n(0)\|)^{\pm}) \mid \Xi^n(0) \] \leq \limsup_{n \to \infty} \sup_{\Xi^n(t) \in \Omega} \mathbb{E}[\exp(u_0(\|n^{1/2}\Delta t^n(t_0)\|) \mid \Xi^n(0) \] \leq \limsup_{n \to \infty} \sup_{\Xi^n(t) \in \Omega} \mathbb{E}[\exp(u_0(\frac{\kappa}{2} + \frac{1}{\min_j \mu_j}n^{-1/2}\|Z^n(t_0)\|)) \mid \Xi^n(0) \] < +\infty \quad (136)
\]

by the ergodicity assumption and the finiteness of expectation of \( \|Z^n(t)\|_T \).
where the last inequality follows from (140) in Lemma (6) (See Appendix G). Similarly, there exists \( u_0 > 0 \), such that

\[
\limsup_{n \to \infty} \sup_{\Xi^n(0) \in \Omega} \mathbb{E} \left[ (\|n^{1/2} \Delta \tau^n(t_0)\| - \|n^{1/2} \Delta \tau^n(0)\|)^2 \exp(u_0(\|n^{1/2} \Delta \tau^n(t_0)\| - \|n^{1/2} \Delta \tau^n(0)\|)) \right] \\
\leq \limsup_{n \to \infty} \sup_{\Xi^n(0) \in \Omega} \mathbb{E} \left[ \max(\|n^{1/2} \Delta \tau^n(0)\|, \frac{\gamma}{2} + \frac{1}{\min_j \rho_j} n^{-1/2}\|Z^n(t_0)\|)^2 \exp(u_0(\frac{\gamma}{2} + \frac{1}{\min_j \rho_j} n^{-1/2}\|Z^n(t_0)\|))\Xi^n(0) \right] \\
< +\infty,
\]

where the last inequality follows from (141) in Lemma (6). We have thus proved (132), and thus (61) in Proposition 4.

It remains to show that \( V(\cdot) \) is a Lyapunov function with drift size parameter \(-1\), drift term parameter \( t_0 \), and exception parameter \( \kappa \) for \( \Xi \), or equivalently, to prove condition (59) for \( \gamma = 1 \). Because \( V(\Xi^n(t)) \) and \( n^{-1/2}\|\Delta \tau^n(t_0)\| \) only differs by almost a constant, proving (59) is equivalent to proving the same condition for \( \|n^{1/2} \Delta \tau^n(t)\| \) for some positive constant \( \gamma \). To that end, we choose \( t_0 \) as (134) and get

\[
\sup_{n^{1/2} \Delta \tau^n(0) > \kappa} \mathbb{E} \left[ \|n^{1/2} \Delta \tau^n(t_0)\| \mid \|n^{1/2} \Delta \tau^n(0)\| \right] \\
\leq \sup_{n^{1/2} \Delta \tau^n(0) > \kappa} \mathbb{E} \left[ \frac{\gamma}{2} + \frac{1}{\min_j \rho_j} n^{-1/2}\|Z^n(t_0)\| \mid \|n^{1/2} \Delta \tau^n(0)\| \right] - \kappa
\]

for some constant \( c' > 0 \). In (138), the first inequality follows from inequality (135) and that \( \|n^{1/2} \Delta \tau^n(0)\| > \kappa \), and the second inequality follows from (139) in Lemma 6 that \( n^{-1/2}\|Z^n(t_0)\| \) is uniformly upper bounded. By choosing a sufficiently large \( \kappa \), we can have \( c' - \frac{\kappa}{2} < -1 \), which proves that \( V(\cdot) \) is a Lyapunov function with drift size parameter \(-1\).

\[\blacksquare\]

**Appendix G: Lemma 6 and its Proof** The following Lemma was used in both Lemma 5 and Proposition 4.

**Lemma 6** There exists a constant \( u_0 > 0 \), such that the following inequalities hold for all fixed \( t_0 \geq 0 \),

\[
\limsup_{n \to \infty} \sup_{\Xi^n(0) - 0 > \kappa} n^{-1/2} \mathbb{E}[\|Z^n\|_{t_0} | \Xi^n(0)] < \infty,
\]

(139)

\[
\limsup_{n \to \infty} \sup_{\Xi^n(0) \in \Omega} \mathbb{E}[\exp(n^{-1/2} u_0 \|Z^n\|_{t_0}) | \Xi^n(0)] < \infty,
\]

(140)

\[
\limsup_{n \to \infty} \sup_{\Xi^n(0) \in \Omega} \mathbb{E}[\|Z^n\|_{t_0}^2 \exp(n^{-1/2} u_0 \|Z^n\|_{t_0}) | \Xi^n(0)] < \infty,
\]

(141)

where \( \Xi^n(0) \) gives the initial state of the Markovian process, and \( Z^n(t) \) is a \( J \)-dimensional centered process defined in (27).

**Proof.** Using the argument provided at the beginning of the proof for Lemma A.1 in Gamarnik and Zeevi [14], inequality (140) implies (139) and (141). To prove (140), define \( A^n_j(t) := \int_0^t p_j(X^n(s) o (n\mu^n)^{-1}) \text{d}s \). Let \( S_j^n(t) \) denote the cumulative number of customers that have completed service at the \( j \)-th SP up to time \( t \),
By change of the time variables, we can derive the following bound for $n^{-1/2} \|Z_n^a\|_{t_0}$,
\[
\begin{align*}
    n^{-1/2} \|Z_j^n\|_{t_0} &
    \leq \|n^{-1/2}(\mathcal{N}(nt) - nt)\|_{A_j^n(t_0)} + \|n^{-1/2}(n\mu_j^n t - S_j^n(t))\|_{W_j^n(t_0)} \\
    &= \|n^{-1/2}(\mathcal{N}(t) - t)\|_{nA_j^n(t_0)} + \|n^{-1/2}(t - S_j^n(\frac{t}{n\mu_j^n}))\|_{n\mu_j^n W_j^n(t_0)} \\
    &\leq \|n^{-1/2}(\mathcal{N}(t) - t)\|_{nA_j^n(t_0)} + \|n^{-1/2}(t - S_j^n(\frac{t}{n\mu_j^n}))\|_{n\mu_j^n W_j^n(t_0)} \\
    &\leq n^{-1/2} \|\mathcal{N}(t) - (t + B_j(t))\|_{nA_j^n(t_0)} + n^{-1/2} \|B_j\|_{n\mu_j^n} + n^{-1/2} \|B_j\|_{n\mu_j^n},
\end{align*}
\]
(142)

where the second inequality follows from $A_j^n(t_0) \leq t_0$, $W_j^n(t) \leq t$, and $\mu_j^n < 2\mu_j$ for a sufficiently large $n$.

We next derive the tail bounds for each term at the RHS of (142). Using standard bounds for Brownian motion, we can bound the following two terms with constants $c_1, c_2 > 0$ which depend on $t_0$ but not on $n$,
\[
\begin{align*}
    \Pr(\|B_j\|_{nA_j^n(t_0)} > \frac{c_1}{4} a \sqrt{n}) &= c_1 \exp(-c_2 a^2) \\
    \Pr(\|B_j\|_{n\mu_j^n} > \frac{c_1}{4} a \sqrt{n}) &= c_1 \exp(-c_2 a^2).
\end{align*}
\]
(143)

Using the functional strong approximation theorem (FSAT) (Theorem 5.14 and Remark 5.17 in Chen and Yao [5]), we may upper bound the tail probability of the other two terms in (142) with constants $c_3, c_4 > 0$ as follows:
\[
\begin{align*}
    \Pr(n^{-1/2} \|\mathcal{N}(t) - (t + B_j(t))\|_{nA_j^n(t_0)} \geq \frac{c_1}{4} a) &\leq n^{-c_3} \exp(-c_4 an^{-1/2}) \\
    \Pr(n^{-1/2} \|S_j^n(t) - (t + B_j(t))\|_{n\mu_j^n} \geq \frac{c_1}{4} a) &\leq n^{-c_3} \exp(-c_4 an^{-1/2})
\end{align*}
\]
(144)

(142), (143), and (144) together imply that
\[
\Pr(n^{-1/2} \|Z_j^n\|_{t_0} \geq a) \leq 2c_1 \exp(-c_2 a^2) + 2n^{-c_3} \exp(-c_4 a \sqrt{n}).
\]
(145)

We can then upper bound the expectation $\mathbb{E}[\exp(n^{-1/2}u_0 \|Z^n\|_{t_0})\Xi^n(0)]$ for all sufficiently large $n$ and initial state $\Xi^n(0)$ using the tail probability bounds,
\[
\begin{align*}
    \mathbb{E}[\exp(n^{-1/2}u_0 \|Z^n\|_{t_0})\Xi^n(0)] &\leq 2 + \int_{-2}^{\infty} \Pr(\exp(n^{-1/2}u_0 \|Z^n\|_{t_0}) > a) \, da \\
    &= 2 + \int_{-2}^{\infty} \Pr(\exp(n^{-1/2} \|Z^n\|_{t_0}) > \frac{\log x}{u_0}) \, dx \\
    &\leq 2 + \int_{-2}^{\infty} 2c_1 \exp(-c_2 \frac{\log x}{u_0^2}) \, dx + \int_{-2}^{\infty} 2n^{-c_3} \exp(-c_4 \frac{\log x}{u_0} a^{-1/2}) \, dx \\
    &< 2M,
\end{align*}
\]
(146)

where the second inequality follows from (145) by replacing $a$ with $\frac{\log x}{u_0}$, and the last inequality follows from the fact that both integrals can be uniformly upper bounded by a constant $M > 0$ for sufficiently large $n$. Thus we have proved inequality (140). ■