OPERATIONS RESEARCH
Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 0030-364X | EISSN 1526-5463 | 00 | 0000 | 0001

INFORMS
DOI 10.1287/xxxx.0000.0000

Asymptotic Behavior of Parallel Queueing Systems with Discrete-Choice Driven Arrivals

Yichuan Ding
Desautels Faculty of Management, McGill University, Montreal, Quebec H3A 1G5, Canada, daniel.ding@mcgill.ca

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

We study a parallel-queue system where each queue is served by a dedicated server at different locations or facilities. Upon arrival, customers observe real-time queue lengths at each facility and choose to join one or balk, whichever that maximizes their expected utility, defined as the service value minus the waiting cost. Our model acknowledges heterogeneity in customer preferences for services at different facilities and their varying waiting time tolerances. We derive fluid and diffusion limit processes to approximate the asymptotic behaviour of the queueing system, exploiting the distinctive features of the arrival rate functions dictated by the discrete choice model and sidestepping the traditional reliance on the Lipschitz-continuity assumption. We prove the uniqueness of the fluid limit process and its converges to a unique equilibrium. At the equilibrium, our analysis under the conditional logit assumption indicates that the social welfare is maximized as long as all service providers are operating at their capacity. Furthermore, we characterize the diffusion limit for the centered process as a reflected multi-dimensional Ornstein-Uhlenbeck process. Analysis of the diffusion model reveals that publicizing real-time wait times does not change the social welfare. Applications of our theoretical results are illustrated through a case study of vehicle queues at U.S.-Canada border-crossing ports.

Key words: Discrete Choice Model, Nonlinear Complementarity Problem, Fluid and Diffusion Approximation, Reflected Multi-Dimensional Ornstein-Uhlenbeck Process

1. Introduction

The discrete choice model has widely explored in the literature to model consumer behavior. As a typical scenario studied in the literature, a consumer chooses from an assortment of products with different features and prices to maximize her utility. A customer's net utility of choosing a product is the difference between the reward and cost of obtaining the product. In this paper, we apply the discrete-choice model to a parallel queue system, unveiling the intricate dynamics between consumer choice behavior and queueing processes. This application necessitates novel methodologies and yields unique insights into system management.

We study a service system consisting of service providers (SP) operating at different locations. Although the services offered by different SPs are fundamentally similar – implying a customer requires service from only one SP – differences in SP characteristics may influence customer preferences. Consequently, customers might value the service from each SP differently. To model customer

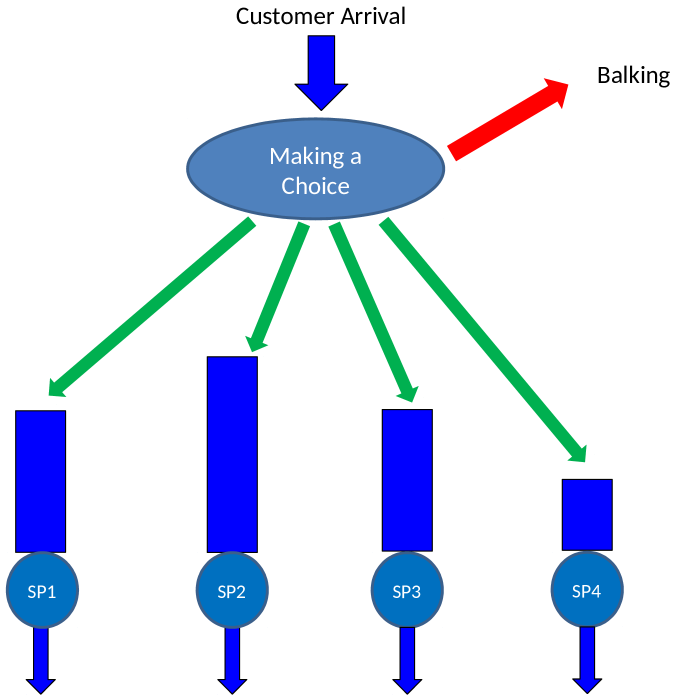


Figure 1 The parallel queue system studied in this paper

decision-making, we assume that all customers are fully informed about both the service utility and the expected waiting time at each SP, allowing them to assess the expected net utility of joining any given queue. Alternatively, a customer may choose to forgo service by not joining any queue, thereby accruing zero utility. If we apply the classical "discrete choice model" to this system, then a customer's decision hinges on selecting the SP that maximizes her expected utility. Figure 1 provides a graphical illustration of such a system.

To broaden the applicability of our model, we study a discrete choice model with random coefficients (Berry et al., 1995; Nevo, 2000),which covers the case with deterministic coefficients as a special case. This model assumes that the coefficients in a customer's utility function, that is, the service value at each SP and the waiting cost per unit time, are randomly distributed. When real-time waiting time estimates are available to customers, under mild regulation assumptions on the parameter distribution, we can show that the mean arrival rate for each queue is an absolutely continuous function of the waiting time estimates at each different SPs. Furthermore, when the customer's choice is formally modelled, the arrival rate function satisfies the following *waiting-aversion* property: as a queue becomes longer, some customers will be discouraged to join that queue and will instead join other queues or balk. Consequently, the mean arrival rate of a queue decreases with its own length due to waiting-aversion; but is non-decreasing with the lengths of other queues. A formal mathematical description of these properties will be provided in Section 3. Such characteristics of the arrival rate function stem from the underlying discrete choice model. The main objective of this paper is to provide asymptotic characterization of discrete-choice (driven) parallel queues, or briefly, DCPQ.

Our study is motivated by several practical instances that fit the DCPQ model that are both widely observed and are areas of research in the Operations Research literature. One example is the kidney transplant waitlist for patients with end-stage renal disease. Kidneys from deceased donors are allocated to patients who have registered on the transplant list according to a given policy. One

allocation policy proposed and tested by Su and Zenios (2006) partitions kidneys 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 parallel queues, each corresponding to a unique type of service (organs). The stylized models analyzed by Su and Zenios (2006) and Ata et al. (2021) are a simplified version of the DCPQ, by assuming that the patient uses the steady-state queue-lengths to calculate the corresponding waiting times.

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`, announces real-time waiting time estimates for major hospitals with emergency rooms in the Metro Vancouver area. Recent studies (Dong et al., 2019; Park et al., 2023) have provided empirical evidence supporting the idea that waiting time announcements influence the emergency department choices of some patients.

The third example is the international border crossing facilities located between the U.S. and Canada. In the Pacific northwest, there are four border crossing facilities. Almost real-time wait time at each one of these facilities is available. Travellers have preferences for location and the amenities available at each facility and make their choice based on the wait time and the characteristics of each facility. Using a novel data from the Canada-US border crossing in the Pacific Northwest, we calibrate our model and illustrate the managerial implications of our theoretical results.

There is rich literature on queueing systems with customer choice. A number of assumptions about the number of queues (usually a single queue) or congestion information (usually non-real-time) or consumer types (usually single class and homogeneous) or server types (usually homogeneous) have to be made in the literature. However, many of these assumptions may not apply to stochastic service systems in practice. The DCPQ model does not impose any of these restrictive assumptions. Thus, not surprisingly, an exact analysis of DCPQ is challenging. For this reason, we study the queue-length process of DCPQ using fluid and diffusion approximations. Even under such approximations, however, few results are known for parallel queues with general state-dependent arrival rates, e.g., the existence of a system equilibrium, and stationary distribution of the queue length process, etc. See Section 2 for a more detailed literature review. However, we show that these results hold owing to the distinctive characteristics of the arrival rate function driven by the discrete choice model, and from there, we derive managerial implications regarding service capacity allocation and information disclosure.

We develop the following approximations for DCPQ. First, under the fluid approximation, we show that the fluid limit process converges to a unique equilibrium which can be characterized as the solution to a nonlinear complementarity problem (NCP). Second, using the diffusion approximation, we show that under diffusion scaling, the centered 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. We also prove that interchange of limit holds, that is, the stationary distribution of the scaled queue-length process converges to the stationary distribution of the RMOU.

By establishing the above results, we make several important contributions to the related research domain.

1. We propose a fairly general model, i.e., the DCPQ, by incorporating a discrete choice model into a parallel queue system. This type of behavior has been empirically identified in emergency department choices (Dong et al., 2019; Park et al., 2023). We approximate the transient and stationary behaviors of the queue-length process in DCPQ via fluid and diffusion approximation. We propose an algorithm to compute the equilibrium state of the fluid limit process, and derive the closed-form stationary distribution for the diffusion limit process. These results facilitate the efficient evaluation of the long-term performance metrics of DCPQ, such as the social welfare.

2. The asymptotic characterizations provide system managers with fresh qualitative insights to the management of DCPQ. Based on the fluid model, we show a perhaps surprising result that under a logit model, when the staffing cost is the same at different locations, the social welfare remains a constant regardless how the service capacity is allocated among SPs with a non-empty queue. In particular, the congestion level or service values at different SPs do not play a role in determining the optimal service rates, as long as no service capacity is wasted.
3. The examination of the diffusion limit's stationary state reveals that sharing waiting time information has almost no impact on social welfare, underpinning a fundamental distinction from findings that have been documented for single-queue systems (Ibrahim, 2018; Guo and Zipkin, 2007; Hu et al., 2018; Wang and Hu, 2020; Hassin and Roet-Green, 2020). This disparity underscores the unique complexities inherent in managing DCPQ.
4. The choice driven properties of the arrival rate function allow us to establish the following technical results in lieu of the Lipschitz continuity assumption: uniqueness of the fluid limit process, convergence of the original stochastic process to the fluid limit and diffusion limit, and interchange of limits. We show that these results may not hold in parallel queues with general, non-Lipschitz arrival rates; but they hold when the arrival rates are non-Lipschitz but have the choice-driven properties. We thus provide a new proof technique for the above results that does not rely on the Lipschitz assumption as the classical methods (e.g., (Mandelbaum et al., 1998a,b)) do. The technical results may be of independent interest to the applied probability society.

2. Literature Review

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 2. As shown in Figure 2, 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 rate at the same server, while servers are different in service value and service rate. Note that for each node in Figure 2, 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. (2006).

Some of the early models of the $O\backslash S\backslash H\backslash I$ type are by Naor (1969) and Leeman (1964) who investigated homogeneous customers' decisions on whether to join a queue for service. When the queue is observable, they showed 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 Stidham Jr (1978). Hassin (1986a) found that in a last-come-first-serve queue with customer abandonment, the differences between Pareto optimal and social optimal equilibria due to possible customers' negative externality does not arise. Larsen (1998) generalized Noar's model to the

one with heterogeneous customers who differ in service value. In contrast, Edelson and Hilderbrand (1975) and Frutos and Gallego (1999) 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 Hassin et al. (2006), except for some special models (e.g. Hassin (2009)). For this reason, the $O\backslash M\backslash H\backslash I$ type models are studied under a weaker notion of equilibrium such as the "ϵ-equilibrium". An example of $O\backslash M\backslash H\backslash D$ type model was considered in Li and Lee (1994). They considered a setting with two queues with heterogeneous servers and homogeneous customers where balking 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 DCPQ studied in this paper falls into this category as we assume customers have different sensitivity with delay and heterogeneous preferences among SPs. Related studies in this category focus on the case where customers receive delayed information about waiting time estimates; see Pender et al. (2020) and Dong et al. (2019). There are several fundamental differences between our work and these two papers. Two important ones are: (1) our paper considers a more general customer choice model which requires different analytical methods and (2) the steady-state characterizations derived for our model may not hold when information is delayed.

The first study on the simplest unobservable queue case or $U\backslash S\backslash H\backslash I$ type was done by Edelson and Hilderbrand (1975) and Chen and Frank (2004). Two extensions followed the basic unobservable queue model. Littlechild (1974) 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 (1985a) extended the model to a more general GI/G/s setting. Luski (1976) generalized the model in Edelson and Hilderbrand (1975) 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. (2014) 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. Other queueing models involving strategic behavior of customers or servers include Adiri and Yechiali (1974); Maglaras et al. (2016); Afèche and Ata (2013); Ward and Armony (2013); Ibrahim et al. (2016); Dong et al. (2015); Gupta and Zhang (2014).

The second stream of related research is the one on fluid and diffusion approximations for service systems with multiple queues. In the models in this stream, 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 DCPQ is the state-dependent queueing network studied in Haddad and Mazumdar (2012); Lee and Puhalskii (2015); Leite and Fragoso (2008); Mandelbaum et al. (1998a,b); Yamada (1995), with some important differences. Compared to a general state-dependent queueing network model, the choice-driven property allows us to derive several characterizations for the fluid and diffusion limit processes (e.g., the fluid limit process converges to a unique equilibrium point, the diffusion limit process is an RMOU process, whose steady-state distribution admits a closed-form characterization). Those characteristics are otherwise not valid in a general state-dependent queueing network. Furthermore, we show that the choice-driven property can substitute for the Lipschitz property in the proofs of the above results in Mandelbaum et al. (1998b). Other papers on state-dependent queues investigated the case when the service speed depends on the workload in the buffer, e.g., (Abouee-Mehrizi and Baron, 2016; Delasay et al., 2016; Dong et al., 2015).

There are several queueing models in which the fluid and diffusion limits exhibit similar ergodic properties. A well-known example is a queueing network with constant arrival rates and constant or state-dependent routing matrix; see Harrison and Reiman (1981) and Reiman (1984). In these

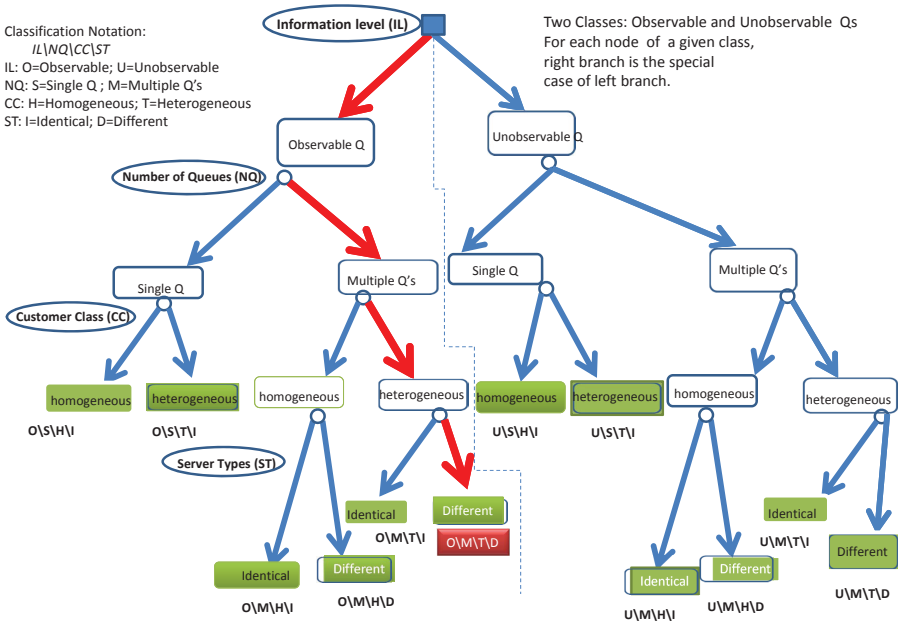


Figure 2 Classification of Queueing Models with Customer Choice.

models, the fluid limit process has a unique equilibrium $\mathbf{0}$, owing to the negative drifts and the non-negative constraint enforced by the reflecting barrier. Consequently, the diffusion limit process in those models is a multi-dimensional Brownian motion with a reflection barrier at $\mathbf{0}$. These characterizations differ from the DCPQ, in which the equilibrium state of the fluid process results from the choice-driven property, and the diffusion limit is thus an RMOU rather than a reflected Brownian motion. Another related model is an overloaded queueing network where customers in each queue renege after an exponentially distributed time. For such a model, Reed and Ward (2004) showed that the fluid limit has a non-zero equilibrium and the diffusion limit process is a non-reflected multi-dimensional O-U process. Other similar models include a service system with differentiated service levels in Maglaras and Zeevi (2004), or with heterogeneous customer types in Harrison and Zeevi (2004). In all these models, the drift is a linear function of the system state; whereas our model allows the drift function to be nonlinear and possibly non-smooth. Therefore, to adapt the existing methods to our model, we need to show that the original process can be approximated by a diffusion process with a linear drift when it is close to the equilibrium.

The third strand of literature explores the impact of waiting time announcements on customer behaviour and social welfare. Hassin (1986b) demonstrates that revealing queue lengths optimizes social welfare primarily under conditions of significant congestion. Chen and Frank (2004) and Hassin and Roet-Green (2020) uncovered a similar insight with respect to maximizing throughput. Both Guo and Zipkin (2007) and Wang and Hu (2020) have evaluated various strategies for information disclosure, discovering that an increase in available information does not invariably improve social welfare. Hu et al. (2018) found that certain degrees of information heterogeneity could potentially elevate social welfare or throughput. Additionally, Armony et al. (2009) delved into the effects of delay announcements within a call center equipped with many servers. While these studies predominantly focus on single-queue systems, the examination of waiting time announcements in

parallel queue systems remains comparatively sparse. Armony and Maglaras (2004) investigated a two-parallel queue system where customer choices are influenced by real-time waiting information. Although their model shares similarities with ours, their analysis is confined to a one-dimensional stochastic process within the Quality and Efficiency Driven (QED) regime. In contrast, our study encompasses a multi-dimensional stochastic process, with queues that may exist in various regimes. Further contributions to this area include Dong et al. (2019) and Pender et al. (2020), who assessed the effects of accuracy and delays in releasing waiting time information on the performance of a parallel-queue system. Singh et al. (2023) investigated the impact of waiting time announcements on two competing parallel queues by analyzing an asymmetric Join-the-Shortest Queue system. Our research contributes to the ongoing discourse by presenting a distinctive conclusion: the degree of information sharing about waiting times does not impact social welfare within a DCPQ setting, characterized by customer heterogeneity in terms of service valuation and waiting time tolerance.

3. The DCPQ Model

3.1. The Discrete Choice Model

We consider a system with J parallel heterogeneous service providers, indexed by $j = 1, 2, \ldots, J$. We assume that the customers' queue-joining behavior follows the classical discrete choice model with random coefficients (e.g. Train (1986); Berry et al. (1995)). Under this assumption, the resulting arrival rate function exhibits discrete-choice driven properties that will be defined later in this section. Formally, for a customer of type ξ, the information available to that customer includes the service value at the j^{th} SP, $u_{\xi,j}$, the customer's waiting cost per unit time, $c_\xi \geq 0$, and the system state. The system state can be described by a J-dimensional vector of waiting time estimates for the customer to join each queue right before time t, that is, $\boldsymbol{\tau}(t-) := (\tau_j(t-))_{j=1,\ldots,J}$, where $\boldsymbol{\tau}(t-)$ denotes the left-limit of $\boldsymbol{\tau}(\cdot)$ at time t. We assume that there are uncountably many different customer types ξ. Note that both the service value $u_{\xi,j}$ and the waiting cost c_ξ vary by customer type ξ, and can be regarded as random variables that follow a fixed probability distribution as customers are drawn from a fixed population. Given waiting time estimates $\boldsymbol{\tau}(t-)$, a customer indexed by ξ can compute her expected utility $U_{\xi,j}$ of joining the j-th queue at time t as follows,

$$U_{\xi,j} = \begin{cases} u_{\xi,j} - c_\xi \tau_j(t-), & \text{if } j \neq 0 \text{ (joining)} \\ 0 & \text{if } j = 0 \text{ (balking)} \end{cases} \tag{1}$$

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

$$\arg\max\{U_{\xi,j} \,|\, j = 0, \ldots, J\}, \tag{2}$$

where the utility of balking is assumed to be zero without loss of generality. For example, suppose an arrived customer sees two queues with waiting time estimates $\tau_1(t) = 1$ and $\tau_2(t) = 2$. If the parameters of a customer are $u_{\xi,1} = 0$, $u_{\xi,2} = 3$, and $c_\xi = 1$, then his utility of joining queue 1 and 2 are $U_{\xi,1} = -1$ and $U_{\xi,2} = 1$, respectively, in which case he will join queue 2. If we change the value of c_ξ from 1 to 2, then his utility will be $U_{\xi,1} = -2$ and $U_{\xi,2} = -1$, in which case he will choose to balk (queue 0), receiving a utility 0.

Since the parameters $\boldsymbol{u}_\xi := (u_{\xi,j})_{j=1,\ldots,J}$ and c_ξ have a fixed joint distribution, we can compute the probability for a randomly drawn arrived customer to choose a queue $j = 0, 1, \ldots, J$, where queue 0 corresponds to balking by slightly abuse of notation. The choice probabilities have the following expressions,

$$\begin{aligned} p_0(\boldsymbol{\tau}(t-)) &= \Pr(0 > u_{\xi,k} - c_\xi \tau_k(t-) \ , \ k = 1, \ldots, J) \\ p_j(\boldsymbol{\tau}(t-)) &= \Pr(u_{\xi,j} - c_\xi \tau_j(t-) > 0 \text{ and } u_{\xi,j} - c_\xi \tau_j(t-) > u_{\xi,k} - c_\xi \tau_k(t-), \ k = 1, \ldots, J, \ k \neq j). \end{aligned} \tag{3}$$

As will be discussed later, we assume that $(\boldsymbol{u_\xi}, c_\xi)$ has a continuous distribution and thus a tie happens with zero probability. For brevity in the remainder of this paper, we will omit $-$ in $\boldsymbol{\tau}(t-)$.

Next,we introduce assumptions on the random coefficients $(\boldsymbol{u}_\xi, c_\xi)$. These assumptions are minimal and are able to accommodate a wide range of applications. Under these assumptions, we prove certain desirable properties of the arrival rate function which in turn facilitate the asymptotic characterization of the DCPQ. Later, we will show that our choice model subsumes several well known models such as the conditional logit model and probit model. For the sake of brevity, we will omit the subscript ξ and denote the random parameters as u_j and c when there is no ambiguity.

Assumption 1 *(Waiting Aversion)* $c > 0$ *a.e.*

Assumption 1 posits that all customers exhibit aversion to waiting, which is a key feature of the choice model studied in this work. Waiting aversion ensures that, whenever a queue becomes longer, fewer customers will join this queues while more will join other queues or balk, which leads to the choice-driven properties of the arrival rate function to be discussed later.

Assumption 2 *(Continuous Distribution of Service Values) The vector* $\boldsymbol{u}$ *follows an absolutely continuous joint cumulative distribution function (cdf).*

Assumption 2 requires $\boldsymbol{u}$ to possess a finite probability density function (pdf) almost everywhere. This assumption is not applied to the distribution of c, which is allowed to be continuous or discrete, or a hybrid of both. The existence of a density function of $\boldsymbol{u}$ is crucial as it ensures that any small variations in queue lengths influence only a small fraction of customers' choices. Without such a density function, the choice probability may be discontinuous in queue lengths, presenting considerable obstacles for our analysis. Specifically, in a two-queue setup without a density function for $\boldsymbol{u}$, we could encounter a situation where $\Pr(u_1 = u_2) > 0$, resulting in a join-the-shortest-queue (JSQ) model where the arrival rate is discontinuous at $X_1 = X_2$. Such discontinuity deviates from the expected behavior in the DCPQ framework and calls for a very different analytical framework; see (Eschenfeldt and Gamarnik, 2018; Cao et al., 2019).

While it is necessary to posit that $(\boldsymbol{u})$ possesses a finite density function almost everywhere to facilitate our analysis, we accommodate the possibility of it approaching to infinity over sets of measure zero. This allows our model to approximate scenarios where a non-negligible proportion of customers might not be served at SP j (where $u_j = 0$) or exhibit indifferent between servers j and k (where $u_j = u_k$). Consequently, although the cdf maintains absolute continuity, certain points may feature an infinite derivative. Many distributions exhibit this characteristic, including the Weibull, Beta, and Gamma distributions within specific parameter ranges.

The general formulation presented encompasses a variety of well studied choice models. By positing that the random utility $U_{\xi,j}$ adopts the parametric form

$$U_{\xi,j} = v_j - c\tau_j + \epsilon_{\xi,j}, \tag{4}$$

where v_j and $c \geq 0$ are constants, and $\epsilon_{\xi,j}$ is drawn from n i.i.d. standard type-1 extreme value distribution, we align with the well-known conditional logit model McFadden et al. (1973). Within this model, the choice probability has closed-form expressions,

$$p_0(\boldsymbol{\tau}) = \frac{1}{1 + \sum_{k=1}^{J} \exp(v_k - c\tau_k)}, \quad p_j(\boldsymbol{\tau}) = \frac{\exp(v_j - c\tau_j)}{1 + \sum_{k=1}^{J} \exp(v_k - c\tau_k)} \text{ for } j = 1, \ldots, J. \tag{5}$$

Similarly, we can get a probit model by assuming $\epsilon_{\xi,j}$ to follow an i.i.d. normal distribution.

3.2. Arrival Process

We next characterize the arrival process under the discrete choice model. Formally, we assume that the service times at server j are i.i.d. random variables with a finite mean $1/\mu_j$. We use the vector notation $\boldsymbol{\mu} := \{\mu_j\}_{j=1,\dots,J}$. Customers arrive at the system according to a time-homogeneous Poisson process with a constant rate 1. 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, abandonment and switching between queues are not allowed (Though an extension of the model with exponential abandonment time is doable and discussed in Section 9). The service discipline is First-Come-First-Served (FCFS) at each queue. A customer leaves the system permanently after service completion.

We describe the system state at time t using a queue-length vector $\boldsymbol{X}(t) := (X_j(t))_{j=1,\dots,J}$, where $X_j(t)$ denotes the number of customers in queue j including the one currently in service. In most practical applications of DCPQ, the remaining service time of the customer at the head of line cannot be observed by either the customer or the system manager. Therefore, we assume that the customers or the system manager will simply use the average service time of a new job to estimate that remaining service time. This approximation is typically accurate because the queue length in many realistic applications of the DCPQ are usually much larger than one. Using this approximation, the waiting time estimator $\tau_j(\xi)$ has the following expression:

$$\tau_j(t) = \frac{X_j(t)}{\mu_j}. \tag{6}$$

In the rest of the paper, we refer to $\tau_j(t)$ as the waiting time estimate or the delay estimate.

Recall that we use $p_j(\boldsymbol{\tau})$ $(j = 0, 1, \dots, J)$ to denote the probability for a randomly arriving customer to choose queue j, which is assumed to be independent of the arrival sequence. Since the aggregate arrival rate is one, the mean arrival rate for queue j is exactly $p_j(\boldsymbol{\tau})$. We denote the state-dependent arrival rates by $\boldsymbol{\Lambda}(\boldsymbol{\tau}) := (p_j(\boldsymbol{\tau}))_{j=1,\dots,J}$, and term this the *arrival rate function*. Let $\boldsymbol{R}(\boldsymbol{\tau}) := (\frac{\partial p_i(\boldsymbol{\tau})}{\partial \boldsymbol{\tau}_j})_{i,j=1,\dots,J}$ denote the Jacobian matrix of $\boldsymbol{\Lambda}(\boldsymbol{\tau})$. We next define two properties for the arrival rate function.

Definition 1 *$\boldsymbol{\Lambda}(\cdot)$ is said to satisfy the stability condition, if for each j, there exists $K > 0$ such that*

$$p_j(\boldsymbol{\tau}) < \mu_j \text{ for all } \boldsymbol{\tau} \in \mathbb{R}_+^J \text{ with } \tau_j \geq K. \tag{7}$$

The stability condition guarantees that whenever a queue is sufficiently long, the state-dependent arrival rate is strictly capped by the service capacity, so the queue length never approaches to infinity. Equation (7) can be considered as the "state-dependent" version of the well-known stability condition "$\lambda < \mu$" in a single queue.

Definition 2 *$\boldsymbol{\Lambda}(\boldsymbol{\tau}) := (p_j(\boldsymbol{\tau}))_{j=1,\dots,J}$ is said to satisfy the choice driven (CD) properties if it is absolutely continuous in $\boldsymbol{\tau}$, and its Jacobian matrix $\boldsymbol{R}(\boldsymbol{\tau})$ is continuous everywhere*[1] *and satisfies the following properties for almost every $\boldsymbol{\tau} := (\tau_j)$:*

1. *(CD-a) Non-Negative Off-Diagonals:*

$$p_j(\boldsymbol{\tau}) \text{ is non-decreasing in } \tau_k \text{ for } j = 1, \dots, J \text{ and } k \neq j. \tag{8}$$

Or equivalently, its Jacobian $\boldsymbol{R}(\boldsymbol{\tau})$ has non-negative off-diagonal entries.

[1] If $\partial p_j(\boldsymbol{\tau})/\partial \tau_i = +\infty\,(-\infty)$ at $\boldsymbol{\tau}$. Then continuity at $\boldsymbol{\tau}$ means $\lim_{n\to\infty} \partial p_j(\boldsymbol{\tau}^n)/\partial \tau_i \to +\infty\,(-\infty)$ for any sequence $\boldsymbol{\tau}^n \to \boldsymbol{\tau}$.

2. *(CD-b) Negative Diagonals:*

$$p_j(\boldsymbol{\tau}) \text{ is strictly decreasing in } \tau_j \text{ for } j=1,\ldots,J. \tag{9}$$

Or equivalently, its Jacobian $\boldsymbol{R}(\boldsymbol{\tau})$ *has negative diagonal entries.*

3. *(CD-c) Strict Row and Column Diagonal Dominance:*

$$p_j(\boldsymbol{\tau}+t\boldsymbol{e}) < p_j(\boldsymbol{\tau}) \quad \text{for } j=1,\ldots,J,\ t>0, \tag{10}$$

where $\boldsymbol{e}$ *denotes an all-one vector. Or equivalently,* $\boldsymbol{R}$ *has negative row sums.*

$$\sum_{k=1}^{J} p_k(\boldsymbol{\tau}+te_j) < \sum_{k=1}^{J} p_k(\boldsymbol{\tau}) \quad \text{for } j=1,\ldots,J,\ t>0, \tag{11}$$

where e_j *denotes a vector with its* j^{th} *entry equal to one and all other entries equal to zero. Or equivalently,* $\boldsymbol{R}(\boldsymbol{\tau})$ *has negative column sums.*

The absolute continuity of $\boldsymbol{\Gamma}(\cdot)$ ensures that its Jacobian, $\boldsymbol{R}(\cdot)$, is finite a.e. Nonetheless, $\boldsymbol{R}$ may be unbounded, leading to non-Lipschitz continuous arrival rate function $\boldsymbol{\Lambda}(\boldsymbol{\tau})$; see an example in in the end of Appendix A. Moreover, the properties (CD-a)-(CD-c) together imply that the Jacobian matrix $\boldsymbol{R}(\cdot)$ is non-symmetric negative definite a.e. and all its eigenvalues have negative real parts (see e.g. Plemmons and Berman (1979)).

We next provide some intuition towards the above properties for the arrival rate function. Note that $\tau_j(t)$ is proportional to the queue length. Thus, a larger $\tau_j(t)$ corresponds to a longer queue. Property (CD-a) stands for weak gross substitutability (WGS) across different SPs – the arrival rate tends to increase when other queues become longer. Property (CD-b) means that the arrival rate of a queue decreases when it becomes longer. To interpret Property (CD-c), i.e., Conditions (10) and (11), consider a scenario when the estimated waiting times in all queues have increased by the same amount, then the difference in the expected waiting times across different queues will keep the same. As a result, a customer's preference order between any two queues will not be altered. However, the increased queue lengths lead more customers to balk, so each queue ends up with a smaller arrival rate. This gives strict row diagonal dominance. Also, when one queue becomes longer, it may push some customers to other queues, but may also push some other customers to balk. So the total arrival rate for all queues has to decrease. This gives the column diagonal dominance.

The next proposition shows that the discrete choice model outlined in Section 3 results the CD properties (i.e., (CD-a), (CD-b) and (CD-c)) of the arrival rate function. All proofs in this paper are provided in the Appendices.

Proposition 1 *The arrival-rate function given by* (3) *satisfies Properties (CD-a), (CD-b), and (CD-c) as well as the stability condition* (7)*.*

In fact, we can prove an even stronger property of the arrival rate function under our choice model – its Jacobian matrix is symmetric everywhere. However, since symmetry is not a prerequisite for our subsequent analysis of DCPQ, we opt not to include it as an assumption, aiming to maintain a broader scope of applicability.

4. Notations and Preliminaries

This section introduces some notations and preliminary results that will facilitate the subsequent asymptotic analysis. All vectors and matrices are in **boldface** to differentiate from the scalars. For a sequence of random vectors $\boldsymbol{X}^n$, we use $\boldsymbol{X}^n \to \boldsymbol{X}$ a.s., $\boldsymbol{X}^n \xrightarrow{p} \boldsymbol{X}$, and $\boldsymbol{X}^n \Rightarrow \boldsymbol{X}$ to denote almost surely point-wise convergence, convergence in probability, and convergence in distribution (weak convergence), respectively. Let $\mathcal{J} := \{1, 2, \ldots, J\}$ denote the index set of the SPs. For a vector $\boldsymbol{a} \in \mathbb{R}^J$, we use $\|\boldsymbol{a}\|$ to denote the ∞-norm, so $\|\boldsymbol{a}\| := \max_{j\in\mathcal{J}} |a_j|$. For two vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^J$, we use $\langle \boldsymbol{a}, \boldsymbol{b} \rangle := \sum_{i=1}^{J} a_i b_i$ to represent the inner product, and use $\boldsymbol{a} \circ \boldsymbol{b} := (a_j b_j)_{j\in\mathcal{J}}$ to represent the Hadamard product. For a given nonnegative vector $\boldsymbol{\mu} \in \mathbb{R}^J_{++}$, we define the $\boldsymbol{\mu}$-norm as $\|\boldsymbol{a}\|^{\boldsymbol{\mu}} := \|\boldsymbol{a} \circ \boldsymbol{\mu}\|$. Note that the μ-norm is topologically equivalent to the ∞-norm. Let $\mathrm{Diag}(\boldsymbol{a})$ denote a diagonal matrix with its diagonal entries being $\boldsymbol{a}$. We use $\boldsymbol{B}(t)$ to denote a J-dimensional standard Wiener process starting at $\mathbf{0}$.

Let $D([0, +\infty), \mathbb{R}^J)$ denote the space of right-continuous functions with left limits (i.e., RCLL functions) in $\mathbb{R}^J$ with time domain $[0, +\infty)$, endowed with the usual Skorokhod topology (Jacod and Shiryaev, 1987). For any $T > 0$, we define the uniform norm $\|\cdot\|_T$ on space $D([0, +\infty), \mathbb{R}^J)$ as

$$\|\boldsymbol{y}\|_T = \sup\{\|\boldsymbol{y}(t)\|,\ s \in [0, T]\}. \tag{12}$$

We denote $\|\boldsymbol{y}\|_\infty := \sup\{\|\boldsymbol{y}(s)\|,\ s \in [0, +\infty)\}$ with a slight abuse of notations. We say that $\boldsymbol{y^n} \to \boldsymbol{y}$ uniformly on all compact sets (u.o.c.), if $\|\boldsymbol{y^n} - \boldsymbol{y}\|_T \to 0$ a.s. for all $T > 0$. When $\boldsymbol{y}$ is continuous, convergence in the topology induced by the uniform norm is equivalent to convergence in the Skorokhod topology (Chen and Yao, 2001). Therefore, when the limit process is continuous, to prove convergence with respect to the Skorokhod topology, it suffices to prove convergence with respect to the uniform topology on compact sets.

We next introduce the notations of reflection mapping, which is similar to the oblique reflection mapping defined in Chapter 7 of Chen and Yao (2001). Let $\Phi^\Omega : D([0, \infty), \mathbb{R}^J) \to D([0, \infty), \Omega)$ denote the reflection mapping with respect to a rectangular domain $\Omega := \prod_{j\in\mathcal{J}} [a_j, b_j]$, where $-\infty \le \boldsymbol{a} < \boldsymbol{b} \le +\infty$. This mapping ensures that for any given RCLL function $\boldsymbol{z}(\cdot)$ with $\boldsymbol{z}(0) \in \Omega$, $\boldsymbol{x} := \Phi^\Omega(\boldsymbol{z})$ solve the following equations for each j and $t \ge 0$,

$$\begin{aligned} x_j(t) &= z_j(t) + l_j(t) - u_j(t). \\ l_j(t) &:= \sup_{0\le s\le t} [a_j + u_j(s) - z_j(s)]^+ \\ u_j(t) &:= \sup_{0\le s\le t} [z_j(s) + l_j(s) - b_j]^+, \end{aligned} \tag{13}$$

where $l_j(\cdot)$ and $u_j(\cdot)$ are the minimal non-decreasing processes that ensure $x_j(\cdot)$ remains within the interval $[a_j, b_j]$ at all times. We allow $a_j = -\infty$ or $b_j = +\infty$, corresponding to $l_j \equiv 0$ or $u_j \equiv 0$, respectively.

Let $b_j(k)/\mu_j, b_j(2)/\mu_j, \ldots$ denote the sequence of service times of customers processed by SP j. The random variables $b_1(k), b_2(k), \ldots$, are assumed to be iid with a unit mean and a finite variance ω_j^2. From that, We define a associated renewal process $S_j(t) := \max\{k \mid \sum_{i=1}^{k} b_j(k) \le \mu_j t\}$, which gives the cumulative number of service completions at the SP j provided that the service provider has been busy during interval $[0, t]$.

To derive the fluid and diffusion limit processes, we consider a sequence of DCPQs indexed by $n = 1, 2, \ldots$. Within the n^{th} DCPQ, customers, inclusive of those opting to balk, arrive according to a time-homogeneous Poisson process with constant traffic intensity n, while the service rate at SP j is scaled up to $n\mu_j^n$, with $\mu_j^n \to \mu_j$. The associated renew process has the expression $S_j^n(t) = \{k \mid \sum_{i=1}^{k} b_j(k) \le n\mu_j^n t\}$.

The scaling for the arrival and service rates is strategically chosen based on specific considerations. In contrast to the conventional heavy traffic regime where the total arrival and service rates are asymptotically balanced (Halfin and Whitt, 1981), in DCPQ we allow the aggregate service

rate, $\sum_j \mu_j$, to differ from the aggregate arrival rate of 1. We can have this flexibility because the arrival rates in DCPQ are state-dependent, so an equilibrium exists regardless of the values of μ_j. Furthermore, instead of assuming $\mu_j^n \equiv \mu_j$, we permit $\sqrt{n}|\mu_j^n - \mu_j|$ to converge to a constant. This accommodates scenarios in which certain queues are operating either slightly under demand or over demand at equilibrium. Notably, the slightly under-demand case aligns with the quality-and-efficiency-driven (QED) regime, which has been of significant attention in the queueing literature (Garnett et al., 2002; Armony and Maglaras, 2004).

Upon arrival, a customer chooses the j^{th} SP with state-dependent probability $p_j(\boldsymbol{\tau}(t-))$, which is a deterministic function of the vector of waiting-time estimates $\boldsymbol{\tau}(t-)$ right before time t. We assume that the choice probabilities $\boldsymbol{\Lambda}(\boldsymbol{\tau}) = (p_j(\boldsymbol{\tau}(t-)))$ satisfy all the properties given in Section 3, and are invariant with respect to the system index n.

Finally, we use $\boldsymbol{X}^n(t)$ and $\boldsymbol{\tau}^n(t)$ to denote vectors of queue-lengths and waiting-time estimates in the n^{th} DCPQ at time t, respectively, and use $W_j^n(t)$ to denote the cumulative busy time of the j^{th} SP up to time t.

5. Fluid Approximation

We first derive a compact representation for the arrival process of each queue in DCPQ. In the following lemma, $N(\cdot)$ denotes a rate-one standard Poisson process. $\boldsymbol{\tau}(t-)$ denotes the left limit of $\boldsymbol{\tau}(\cdot)$ at time t, which exists because $\boldsymbol{\tau}(t) = \boldsymbol{X}(t) \circ \boldsymbol{\mu}^{-1}$ is RCLL. Note that the representation provided in Lemma 1 does not imply that the arrival process of each queue is an independent Poisson process, because the traffic intensity is state-dependent. Similar notations have been used in existing literature (e.g., Mandelbaum et al. (1998b); Weerasinghe (2014); Dong et al. (2015)) to represent state-dependent arrival or departure processes.

Lemma 1 *The total number of customers who have joined queue j during time interval $[0,t]$ in the n^{th} DCPQ has the same distribution as $N\left(\int_0^t np_j(\boldsymbol{\tau}(s-))ds\right)$.*

Although the expression provided in Lemma 1 may be intuitive, rigorous derivation of the expression relies on the Meyer's theorem (see for example, Brown and Nair (1988)) and is not straightforward. We attach the proof of Lemma 1 in Appendix B. In the subsequent discussion, we will use $\boldsymbol{\tau}(s)$ instead of $\boldsymbol{\tau}(s-)$ for brevity, causing no changes to the integral.

We next study the asymptotic behavior of the DCPQ via fluid approximation. We prove that the scaled queue-length processes in a sequence of DCPQs converge to a *fluid limit process*. Moreover, we show that the 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} DCPQ, we define the scaled queue-length

$$\boldsymbol{x}^n(t) := \frac{1}{n}\boldsymbol{X}^n(t). \tag{14}$$

We next show that the process $\boldsymbol{x}^n$ converges to a fluid limit process. From hereon, without further specification, we assume that the arrival rate function $\boldsymbol{\Lambda}(\cdot) := (p_j(\cdot))$ satisfies the CD property and the stability condition (7). As a result, the Jacobian matrix of $\boldsymbol{\Lambda}(\cdot)$ is negative definite almost everywhere over $\mathbb{R}_+^J$. We define a vector function $\boldsymbol{\Gamma}(\cdot) := (p_j(\cdot \circ \boldsymbol{\mu}^{-1}))_{j=1,\ldots,J}$, which maps the vector of queue-lengths to the vector of expected waiting times.

Theorem 1 *(Convergence to Fluid Limit) Define $\Omega := [0,+\infty)^J$. Suppose $\boldsymbol{x}^n(0) \to \boldsymbol{x}(0)$ a.s. when $n \to \infty$ with $\boldsymbol{x}(0) \geq 0$. Then for all $T > 0$,*

$$\|\boldsymbol{x}^n - \boldsymbol{x}\|_T \to 0, \ \ a.s. \tag{15}$$

where $\boldsymbol{x}$ is the unique solution to the following differential equation with reflection,

$$\boldsymbol{x}(t) = \boldsymbol{\Phi}^{\Omega}\left(\boldsymbol{x}(0) + \int_0^t (\boldsymbol{\Gamma}(\boldsymbol{x}(s)) - \boldsymbol{\mu})ds\right), \tag{16}$$

where $\boldsymbol{\Phi}^{\Omega}$ is the reflecting mapping defined in Section 4.

In our paper, we assume the arrival process to be time-homogeneous in order to derive steady-state characterization. The convergence to fluid limit process still holds for time-inhomogeneous arrivals and our proof can be adapted to that case.

We compare Theorem 1 with the findings of Mandelbaum et al. (1998b), who in Theorem 4.6, 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 Lipschitz continuous. In contrast, our customer choice model may lead to non-Lipschitz $p_j(\cdot)$. Therefore, the proof technique of Mandelbaum et al. (1998b) cannot be adapted to a proof of Theorem 1. In fact, if the drift function $\boldsymbol{\Gamma}(\cdot)$ in the differential equation (16) are non-Lipschitz, then generally speaking, the differential equation may not have a solution, or have multiple solutions; see Examples 3.1 and 3.3 in (Sideris, 2013).

Interestingly, we find that the CD property can replace the Lipschitz condition in proving Theorem 1. To that end, the next Lemma provides a new sufficient condition for the pathwise uniqueness of a solution to the following stochastic differential equation with reflection (SDER), which is a more general form of (16) by including a stochastic term[2].

$$\boldsymbol{x}(t) = \boldsymbol{x}(0) + \int_0^t \boldsymbol{b}(s, \boldsymbol{x}(s))ds + \int_0^t \boldsymbol{\sigma}(s, \boldsymbol{x}(s))d\boldsymbol{B}(s) + \boldsymbol{\ell}(t), \tag{17}$$

where the $\boldsymbol{\ell}$ is a non-decreasing process that keeps $\boldsymbol{x} \geq 0$ as defined in Equation (13).

Lemma 2 *Suppose $\boldsymbol{b}(s,\cdot)$ is absolute continuous with negative definite Jacobian matrix a.e., and $\boldsymbol{\sigma}(s,\cdot)$ is Lipschitz continuous for all s, that is, $\|\boldsymbol{\sigma}(s,\boldsymbol{x}) - \boldsymbol{\sigma}(s,\boldsymbol{y})\| \leq K\|\boldsymbol{x}-\boldsymbol{y}\|$ for some constant $K > 0$. Then the solution to SDER (17), if exists, must be pathwise unique.*

Tanaka (1979) and Dupuis and Ishii (1993) proved that there exists a pathwise unique solution to (17) if both $\boldsymbol{b}(s,\cdot)$ and $\boldsymbol{\sigma}(s,\cdot)$ are Lipschitz continuous. Swart (2002) and Yamada and Watanabe (1971) discussed pathwise uniqueness under some similar but more general conditions. While our Lemma 2 states that the Lipschitz continuity of the drift coefficient $\boldsymbol{b}(s,\cdot)$ 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 (17).

As a notable difference from the standard proof, our proof for Theorem 1 leverages the CD property instead of the Lipschitz property of the arrival rate function. To leverage the CD property, the proof invokes the inequalities of SDERs in Tanaka (1979) rather than directly applying the Gronwall's inequality.

We call $\boldsymbol{x}$ in Equation (16) the *fluid limit process* of the DCPQ. Because there is a one-to-one correspondence between $\boldsymbol{X}(t)$ and $\boldsymbol{\tau}(t)$ via equation (6), we can alternatively represent the fluid limit process using $\{\boldsymbol{\tau}(t) : t \geq 0\}$. We next define the equilibrium (stationary) state of this fluid limit process.

[2] 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.

Definition 3 *$\boldsymbol{x}^* := (\boldsymbol{x}_j^*) \in \mathbb{R}_+^J$ represent the equilibrium queue-length vector if given $\boldsymbol{x}(0) = \boldsymbol{x}^*$, the differential equation (16) has the solution $\boldsymbol{x}(t) \equiv \boldsymbol{x}^*$. The associated $\boldsymbol{\tau^*} := (\tau_j^*) = (x_j^*/\mu_j)$ represent the equilibrium waiting-time vector.*

Intuitively, a fluid limit process is at its equilibrium if and only if the net flow rate (i.e., difference between the arrival and departure rates) equals zero for each queue. This logic leads to the following characterization of the equilibrium waiting-time vector of the fluid limit process, or briefly, the fluid equilibrium.

Proposition 2 *$\boldsymbol{\tau}^*$ are equilibrium waiting-times vector of an DCPQ if and only if $\boldsymbol{\tau}^*$ is the solution to the following nonlinear complementarity problem (NCP):*

$$NCP \qquad \begin{aligned} \mu_j - p_j(\boldsymbol{\tau}) &\geq 0, \quad \text{for } j = 1, \ldots, J. \\ \tau_j &\geq 0 \quad \text{for } j = 1, \ldots, J. \\ \textstyle\sum_{j=1}^J \tau_j(\mu_j - p_j(\boldsymbol{\tau})) &= 0. \end{aligned} \tag{18}$$

The proof of Proposition 2 is attached in Appendix C.

Theorem 2 *(Existence and Uniqueness of Equilibrium) There exists a unique equilibrium waiting-time vector $\boldsymbol{\tau}^*$ for the fluid limit process in each DCPQ.*

It suffices to prove that the NCP (18) always has a unique solution. To that end, we prove that $-\boldsymbol{\Lambda}(\cdot)$ satisfies the so-called P-property (Moré and Rheinboldt, 1973), which implies uniqueness. We then construct a solution to the NCP via a tatonnement process, i.e., by adjusting the value of τ_j according to the demand-supply gap $\mu_j - p_j(\boldsymbol{\tau})$. A complete proof is provided in Appendix F.

Since our proof for the existence of an NCP solution is constructive, the tâtonnement algorithm introduced in the proof of Theorem 2 can be used to calculate the fluid equilibrium.

The (CD) property is not only sufficient for the existence and uniqueness of the equilibrium state, but also necessary in the sense that without them, these results cannot hold for *certain* parameters. Please see the following examples as an illustration of this point.

Example 5.1 *This example shows that when (CD-a) is violated, the fluid limit process may have multiple equilibria. Consider an example with $\boldsymbol{\mu} = (0.4, 0.4)^T$, arrival rate function $\boldsymbol{\Lambda}(\boldsymbol{\tau}) = (0.4 - 0.1\exp(-(\tau_1 - 1)^2), 0.4 - 0.1\exp(-(\tau_2 - 1)^2)^T$, and its Jacobian $\boldsymbol{R}(\boldsymbol{\tau}) = \begin{pmatrix} 0.2(\tau_1 - 1)\exp(-(\tau_1 - 1)^2) & 0 \\ 0 & 0.2(\tau_2 - 1)\exp(-(\tau_2 - 1)^2) \end{pmatrix}$. The j^{th} diagonal elements are positive when $\tau_j < 1$, so (CD-a) is violated. For queue $j = 1, 2$, the maximum arrival rate is attained when $\tau_j = 1$, at which time the arrival rate and service rate is balanced. Thus, $\tau_j = 1$ is an equilibrium queue length for each queue. In addition to that, $\tau_j = 0$ is also an equilibrium queue length. Thus, this DCPQ consists four equilibrium states, $(1, 1)^T$, $(0, 1)^T$, $(1, 0)^T$, $(0, 0)^T$.*

Example 5.2 *This example shows that an equilibrium state may not exist when (CD-b) is violated. Consider an example with $\boldsymbol{\mu} = (1, 0.01, 0.01)^T$, $\boldsymbol{R} \equiv \begin{pmatrix} -0.2 & -0.1 & -0.1 \\ -0.1 & -0.1 & 0.15 \\ -0.1 & 0.15 & -0.1 \end{pmatrix}$, and $\boldsymbol{\Lambda}(\boldsymbol{\tau}) = (\boldsymbol{R}\boldsymbol{\tau})^+$. This example satisfies the stability condition (7), because the arrival rate for each queue converges to zero when its length approaches to infinity, as long as the lengths of the other queues are fixed. The Jacobian also contains negative diagonals and has negative row and column sums, so (CD-a) and (CD-c) are both satisfied. However, the Jacobian matrix contains negative off-diagonal entries and therefore violates assumption (CD-b). One can check that if the fluid limit process starts from $(0, 1, 1)^T$, then we will have $\tau_1(t) \equiv 0$, and $\tau_2(t) \equiv \tau_3(t) \to \infty$ when $t \to \infty$. Consequently, no equilibrium exists.*

Example 5.3 *This example shows that when (CD-c) is violated, the fluid limit process may also have multiple equilibria. Consider an DCPQ has* $\boldsymbol{R} \equiv \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}$, $\boldsymbol{\mu} = (0.5, 0.5)^T$, $\boldsymbol{\Lambda}(\boldsymbol{\tau}) = ((0.5, 0.5)^T + \boldsymbol{R}\boldsymbol{\tau})^+$. *Then any vector in the form of* $(z, z)^T$ *with* $z \geq 0$ *can be an equilibrium state of the fluid limit process.*

The above examples show that when any one of (CD-a), (CD-b), (CD-c) fails, the fluid limit process may not have a unique equilibrium state. Because all the subsequent asymptotic characterizations for the DCPQ (e.g., convergence of the fluid limit process to the equilibrium, convergence to the diffusion limit process, and the stationary distribution of the diffusion limit process) rely on the fact that the fluid limit process has a unique equilibrium state, these characterizations would not apply to general parallel-queue systems.

The next theorem shows that given the CD property, the fluid limit of the expected delays in DCPQ must converge to the unique equilibrium state.

Theorem 3 *(Convergence to Equilibrium) Suppose* $\{\boldsymbol{x}(t) | t \geq 0\}$ *is a solution to the differential equation* (16) *with* $\boldsymbol{x}(0) \geq 0$, *and* $\boldsymbol{\tau}(t) = \boldsymbol{x}(t) \circ \boldsymbol{\mu}^{-1}$. *Then*

$$\boldsymbol{\tau}(t) \to \boldsymbol{\tau}^*, \qquad \text{when } t \to \infty. \tag{19}$$

The main idea of the proof involves showing the maximal deviation from the equilibrium queue length $\max_j \tau_j(t) - \tau_j^*$ decreases with time due to the CD properties.

6. Service Rate Decisions in DCPQ

We investigate the service rate decisions within DCPQ under the fluid approximation. When service duration has an iid exponential distribution, a multi-server queue exhibits the same asymptotic behaviour as a single-server queue with the total service rate unchanged. This asymptotic equivalence is elaborated upon in Remark 10.2.1. and Theorem 10.2.2 by Whitt (2002). This insight allows us to perceive service rate decisions as a continuous analogue to staffing level optimization, a challenge often faced in the practical management of DCPQ systems.

Mendelson (1985b) formulated the problem of determining the optimal service rate for a single-queue system serving customers with heterogeneous service valuations. This optimization problem is relevant due to the variation in social welfare with service rates. Nevertheless, our analysis reveals that if we assume that customer choices follow by a logit model in DCPQ, then social welfare remains unaffected by how service capacity is distributed among SPs that operate with non-empty queues.

We illustrate the above insight using a simple model, where the aggregate service rate is subject to a total budget, that is, $\sum_{j=1}^{J} \mu_j \leq \overline{\mu}$. Customers choice follows a conditional logit model where the service of SP j provides customer ξ with utility $u_{j,\xi} = v_j + \epsilon_{j,\xi}$, with $\epsilon_{j,\xi}$ following a standard Gumbel distribution independent of $\epsilon_{k,\xi}$ for $k \neq j$. The system manager aims to maximize the expected utility of a unitary population, or equivalently, the social welfare, at the fluid equilibrium $\boldsymbol{\tau}^*$, subject to the service capacity constraint. Leveraging the logit model, we derive the following expression for the expected utility/social welfare,

$$\begin{aligned} SW(\boldsymbol{\tau}^*) := \mathbb{E}[U_{\max}(\boldsymbol{\tau}^*)] &:= \int_{\boldsymbol{u} \in K^J} \max\{0, v_j - c\tau_j^* + \epsilon_j, j = 1, \ldots, J\} f(\boldsymbol{\epsilon}, c)\, d\boldsymbol{\epsilon} \\ &= \ln(1 + \sum_{j=1}^{J} \exp(v_j - c\tau_j^*)), \end{aligned} \tag{20}$$

where the second equality follows the conditional logit model. The service rate decision can be formulated as,

$$SW^* := \max_{\boldsymbol{\tau}^* \geq 0, \boldsymbol{\mu} \geq 0} \ln(1 + \sum_{j=1}^{J} \exp(v_j - c\tau_j^*)) \tag{21}$$
$$\text{s.t.} \quad p_j(\boldsymbol{\tau}^*) \leq \mu_j, \quad j = 1, \ldots, J \tag{22}$$
$$\sum_{j=1}^{J} \tau_j^* (\mu_j - p_j(\boldsymbol{\tau})) = 0 \tag{23}$$
$$\sum_{j=1}^{J} \mu_j \leq \overline{\mu} \tag{24}$$

The formulation of the objective function (21) follows equation (20). Constraints (22)-(23) are the NCP conditions for $\boldsymbol{\tau}^*$ to be the fluid equilibrium given service rates $\boldsymbol{\mu}$. (24) represents the budget constraint.

We next characterize the optimal solution to (21)-(24). We define the *potential demand rate* $\overline{\lambda} := \sum_{j=1}^{J} p_j(\mathbf{0})$ as the total arrival rate when all queues are vacant. Due to the CD property, the potential demand rate gives the maximum possible arrival rate across all states.

Theorem 4 *Suppose customer choice follows the conditional logit model* (3).
(1) If $\overline{\lambda} \leq \overline{\mu}$, then $SW^* = -\ln(1-\overline{\lambda})$, and a feasible solution to (21)-(24) *is optimal if and only if $\boldsymbol{\tau}^* = \mathbf{0}$.*
(2) If $\overline{\lambda} > \overline{\mu}$, then $SW^* = -\ln(1-\overline{\mu})$, and a feasible solution to (21)-(24) *is optimal if and only if $p_j(\boldsymbol{\tau}^*) = \mu_j$ for each j.*

Theorem 4 states that when the service capacity is sufficiently large to meet the potential demand rate ($\overline{\lambda} \leq \overline{\mu}$), then every SP has an empty queue with respect to the fluid scaling, that is, $\boldsymbol{\tau} = \mathbf{0}$. Perhaps the more interesting case is when the service capacity is not sufficiently large ($\overline{\lambda} > \overline{\mu}$). In that case, the optimal service rate has a simple characterization, $p_j(\boldsymbol{\tau}^*) = \mu_j$ for all j, saying that there should be no excessive service capacity at each SP. In other words, to maximize the social welfare, the system manager only needs to make sure to not waste any service capacity. This characterization relies on the logit model assumption, under which maximizing social welfare is equivalent to maximizing the throughput rate. Therefore, any throughput maximizing service rates allocation, has to maximize the social welfare; while in DCPQ, the aggregate throughput rate is a constant $\overline{\mu}$ as long as no service capacity has been wasted.

This result might be counter-intuitive as one might expect that increasing the staffing level at a more popular SP with a larger service value v_j and a larger arrival rate $p_j(\boldsymbol{\tau}^*)$ would bring in more welfare, given that expanding service capacity is equally expensive across different SPs. However, increasing service rate at a more popular site, despite affecting more customers, results in less reduction in waiting time.

If $\epsilon_{j,\xi}$ follows other distributions but remain independent of $\epsilon_{k,\xi}$ for $k \neq j$, the results of Theorem 4 does not hold exactly. However, our numerical study shows that different service rate allocations only make small variations to the total welfare so the insight from Theorem 4 remains robust. At a high level, since customer can choose which queue to join, the DCPQ as a whole is close to a pooling queue. Therefore, the performance of DCPQ is less sensitive to the allocation of service capacity across fully utilized SPs compared to that of a system consisting of separated parallel queues.

Although the result of Theorem 4 is derived under the fluid approximation, the insight remains valid in more refined approximations such as the diffusion model.

Theorem 4 implies the following result, which shows that if service rates can be reallocated to maximize the social welfare, then either all queues are empty, or no queue is empty.

Corollary 1 *(1) If $\overline{\lambda} \leq \overline{\mu}$, then all queues are empty under the socially optimal service rates.*
(2) If $\overline{\lambda} > \overline{\mu}$, then any feasible solution $(\boldsymbol{\mu}, \boldsymbol{\tau}^*)$ with $\boldsymbol{\tau}^* > 0$ must be optimal.

The proof is based on a simple observation that when $\overline{\lambda} > \overline{\mu}$, $\mathbf{0}$ cannot be a feasible solution. Then any $\boldsymbol{\tau}^* > 0$ is optimal as no queue would have excessive service capacity due to the complementary slackness condition (23).

Next, we extend the analysis to the case of unequal staffing cost, in which case the budget constraint is given by

$$\sum_{j=1}^{J} h_j \mu_j \leq \overline{\mu}, \tag{25}$$

where we assume $0 \leq h_1 \leq h_2 \leq \ldots \leq h_J$ without loss of generality. The optimal service rate decision under the above cost remains to have a simple characterization.

Proposition 3 *Suppose* $(\boldsymbol{\mu}, \boldsymbol{\tau}^*)$ *is the optimal solution to* (21) - (23) *and* (25)*, then either* $\boldsymbol{\tau}^* = \mathbf{0}$*, or* $p_j(\boldsymbol{\tau}^*) = \mu_j$ *for all* j *and the following characterization holds with* $k := \min\{j | \tau_j^* > 0\}$*,*

$$\begin{aligned} &\tau_j^* = 0 \text{ if } j < k \\ &\mu_j = 0 \text{ if } h_j > h_k. \end{aligned} \tag{26}$$

Proposition 3 states that the optimal staffing strategy within DCPQ only needs to follow two principles: (1) ensuring no wastage by never providing excessive service, that is, $\mu_j = p_j(\boldsymbol{\tau}^*)$; and (2) prioritizing staffing the most cost-effective SP. Specifically, SPs whose staffing cost falls below a threshold h_k should be fully staffed to eliminate customer backlogs at the fluid scale; conversely, SPs whose staffing costs exceed this threshold are recommended to be suspended to minimize the expense. This policy, while economically rational, might lead to infinitely large waiting times due to service unavailability.

7. Diffusion Approximation

In contrast to the fluid model, a diffusion process can capture the asymptotic behavior of the centered queue-length process at a more granular level. We show that when the queue lengths are close to the equilibrium, then its deviation from the equilibrium, under diffusion scaling, converges to a diffusion limit which is known as a reflected multi-dimensional Ornstein-Uhlenbeck (RMOU) process. We continue to examine the sequence of DCPQs defined in Section 3. In the n^{th} DCPQ, we define the *virtual equilibrium* $\boldsymbol{\tau}^{n,*}$ as the solution to the following NCP

$$\text{NCP} \quad \begin{aligned} n\mu_j^n - np_j(\boldsymbol{\tau}^{n,*}) &\geq 0, \quad \text{for } j = 1, \ldots, J. \\ \tau_j^{n,*} &\geq 0 \quad \text{for } j = 1, \ldots, J. \\ \textstyle\sum_{j=1}^{J} \tau_j^{n,*}(n\mu_j^n - np_j(\boldsymbol{\tau}^{n,*})) &= 0. \end{aligned} \tag{27}$$

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} DCPQ. Since we have assumed that $\mu_j^n \to \mu_j$, the continuity of $p_j(\boldsymbol{\tau})$ implies that the limit of $\boldsymbol{\tau}^{n,*}$ must solve the NCP (18) for the fluid model. Since the solution to (18) is unique according to Theorem 2, we deduce that $\boldsymbol{\tau}^{n,*} \to \boldsymbol{\tau}^*$. We use $\rho_j^n := \frac{p_j(\boldsymbol{\tau}^{n,*})}{\mu_j^n}$ to denote the traffic intensity at the equilibrium waiting-times. Correspondingly, we denote the traffic intensity of queue j in the fluid model by $\rho_j := \lim_{n\to\infty} \rho_j^n$. We consider four mutually exclusive cases of the limiting behaviors of the sequences (τ_j^n) and (ρ_j^n). Note that ρ_j^n is no greater than one in all queues by the NCP condition. $\tau_j^{n,*} > 0$ implies that $\rho_j^n = 1$ by complementarity slackness.

These four cases are not exhaustive, but they cover the asymptotic regimes which have been most often considered in the literature (e.g., Ward and Glynn (2003)).

$$\begin{aligned}
&\textit{Largely Under-demand Queues} && \mathcal{J}^{--} := \{j | \rho_j^n \to \rho_j < 1\} \\
&\textit{Balanced or Slightly Under-demand Queues} && \mathcal{J}^{-} := \{j | \begin{smallmatrix} \tau_j^{n,*} = 0,\ \rho_j^n \le 1 \text{ for all } n,\ \rho_j^n \to 1, \\ \sqrt{n}(\mu_j^n - p_j(\mathbf{0})) \to \theta_j \ge 0 \end{smallmatrix} \} \\
&\textit{Slightly Over-demand Queues} && \mathcal{J}^{+} := \{j | \begin{smallmatrix} \tau_j^{n,*} > 0 \text{ for all } n,\ \tau_j^{n,*} \to \tau_j^* = 0, \\ \sqrt{n}(\mu_j^n \tau_j^{n,*} - \mu_j \tau_j^*) \to \vartheta_j \ge 0 \end{smallmatrix} \} \\
&\textit{Largely Over-demand Queues} && \mathcal{J}^{++} := \{j | \begin{smallmatrix} \tau_j^{n,*} \to \tau_j^* > 0, \\ \sqrt{n}(\mu_j^n \tau_j^{n,*} - \mu_j \tau_j^*) \to \vartheta_j \end{smallmatrix} \},
\end{aligned} \tag{28}$$

where $\boldsymbol{\vartheta} := (\vartheta_j)$ and $\boldsymbol{\theta} := (\theta_j)$ are both J-dimensional vectors and have the following expressions,

$$\theta_j = \begin{cases} \lim_{n\to\infty} \sqrt{n}(\mu_j^n - p_j(\boldsymbol{\tau}^{n,*})) & \text{if } j \in \mathcal{J}^- \\ 0 & \text{otherwise,} \end{cases} \quad \vartheta_j = \begin{cases} \lim_{n\to\infty} \sqrt{n}(\mu_j^n \tau_j^{n,*} - \mu_j \tau_j^*) & \text{if } j \in \mathcal{J}^+ \cup \mathcal{J}^{++} \\ 0 & \text{otherwise.} \end{cases} \tag{29}$$

To ascertain the specific scenario of each queue, we first apply the tâtonnement algorithm in the proof of Theorem 2 to calculate the fluid equilibrium $\boldsymbol{\tau}^*$. By definition, queues with $\mu_j > p_j(\boldsymbol{\tau}^*)$ are largely under-demand and queues with $\tau_j^* > 0$ are largely over-demand. The fluid equilibrium, however, does not distinguish between "slightly under-demand" and "slightly over-demand" queues. This distinction hinges on the asymptotic regime selected by the researcher, that is, how μ_j^n converges to μ_j for queue j. In particular, the well studied QED regime (Halfin and Whitt, 1981) corresponds to the slightly under-demand queue. Whereas our model extends beyond the QED regime, noting that in numerous service systems (e.g., healthcare), expecting service providers to meet the entire potential demand (the arrival rate when all queues are empty) can be overly restrictive. More commonly, an equilibrium is achieved only after a fraction of customers balk due to congestion, which corresponds to largely over-demand and slightly over-demand queues.

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

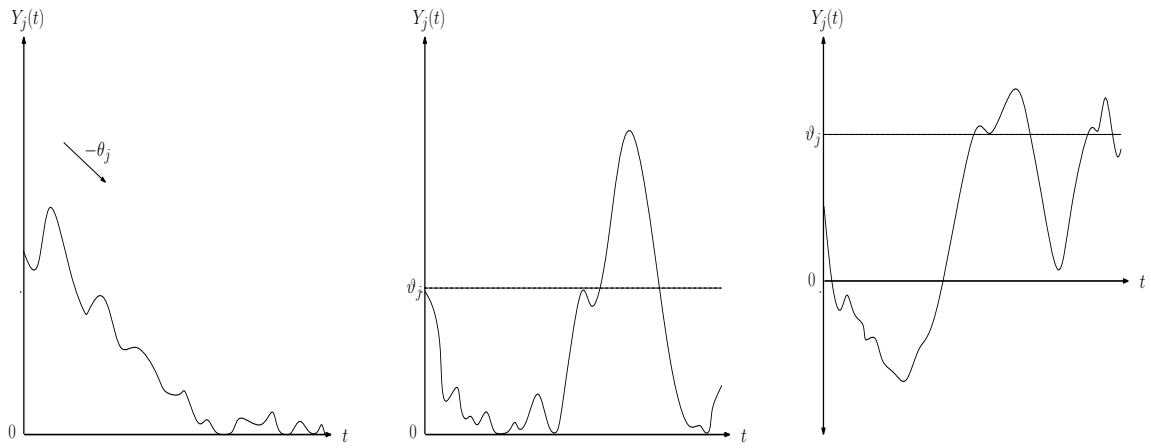
$$Q_j^n(t) := \sqrt{n}(x_j^n(t) - x_j^*), \tag{30}$$

where $\boldsymbol{x}^n$ represents the queue-lengths under fluid scaling that has been defined in Equation (14), 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., $\boldsymbol{Q}_j^n \Rightarrow 0$ (see e.g. Choudhury et al. (1997)). 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 queues in $\mathcal{J}^-$, $\mathcal{J}^+$, and $\mathcal{J}^{++}$, which can co-exist in the same system. For $j \in \mathcal{J}^- \cup \mathcal{J}^+$, we have $x_j^* = \mu_j \tau_j^* = 0$ and thus $Q_j^n(t) \ge 0$; for $j \in \mathcal{J}^{++}$, since $x_j^* > 0$, $Q_j^n(t)$ can be either positive or negative. Consequently, $\boldsymbol{Q}^n$ and its diffusion limit process $\boldsymbol{Y}$ must reside in the following domain:

$$\Omega = \otimes[0, +\infty)^{J^- + J^+} \otimes (-\infty, +\infty)^{J^{++}}. \tag{31}$$

Figure 7 depicts sample paths of the diffusion limit process $Y_j(t)$ when j is in set $\mathcal{J}^-$, $\mathcal{J}^+$, and $\mathcal{J}^{++}$, respectively.

For the diffusion limit process to exist, we need to assume that the arrival rate function to have a finite Jacobian matrix $\boldsymbol{R}^*$ at the equilibrium $\boldsymbol{\tau}^*$. This assumption, however, is without loss of generality as the choice-driven property states that a finite Jacobian exists a.e. Under this



(a)When queue j is slightly under-demand, Y_j tends to move toward the virtual equilibrium $\vartheta_j = 0$ at a constant downward drift rate θ_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 ϑ_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 ϑ_j in an unbounded domain.

Figure 3 Typical sample paths of Y_j in the cases of $j \in \mathcal{J}^-, \mathcal{J}^+, \mathcal{J}^{++}$.

assumption, we derive the diffusion limit for the queue-lengths process in DCPQ as the solution to the following SDER,

$$\boldsymbol{Y}(t) = \int_0^t \left(\boldsymbol{R}^* \text{Diag}\,(\boldsymbol{\mu}^{-1})(\boldsymbol{Y}(s) - \boldsymbol{\vartheta}) - \boldsymbol{\theta}\right) ds + \boldsymbol{\Sigma}^{1/2}\boldsymbol{B}(t) + \boldsymbol{L}(t), \tag{32}$$

where $\boldsymbol{\Sigma}^{1/2}$ is a J-by-J diagonal matrix with $\sqrt{(1+\omega_j^2)\mu_j}$ as its j^{th} diagonal entry, $\boldsymbol{B}(t)$ is a J-dimensional standard Brownian motion with covariance matrix I (identify matrix), and $\boldsymbol{L}(t)$ is a J-dimensional minimal non-decreasing process which makes $Y_j(t) \geq 0$ for all $j \in \mathcal{J}^- \cup \mathcal{J}^+$.

Theorem 5 *(Convergence to Diffusion Limit) Suppose* $\boldsymbol{Q}^n(0) \Rightarrow \boldsymbol{Y}(0)$ *and* $\mathbb{E}\|\boldsymbol{Y}(0)\| < \infty$. *We then have,*

$$\boldsymbol{Q}^n \Rightarrow \boldsymbol{Y}. \tag{33}$$

Before providing an outline of the proof , we make a few remarks. First, according to Theorem 5, the diffusion process has a reflection barrier at 0 only for $j \in \mathcal{J}^- \cup \mathcal{J}^+$, but has no reflection barrier for $j \in \mathcal{J}^{++}$. Intuitively, for $j \in \mathcal{J}^- \cup \mathcal{J}^+$, we have $Q_j^n(t) = \sqrt{n}x_j^n(t)$. Thus, $Q_j^n(t) = 0$ (so $x_j^n(t) = 0$) means that queue j is empty, at which time the server has to stop working and prevents $Q_j^n(t)$ from decreasing further. Therefore, if $j \in \mathcal{J}^- \cup \mathcal{J}^+$, 0 is a reflecting barrier for $Q_j^n(t)$. For $j \in \mathcal{J}^{++}$, since $x_j^* = \mu_j \tau_j^* > 0$, an empty queue ($x_j^n(t) = 0$) corresponds to $Q_j^n(t) = \sqrt{n}(0 - x_j^*) \to -\infty$ when $n \to \infty$. That means, if $j \in \mathcal{J}^{++}$, the reflection barrier for $Q_j^n(t)$ is at $-\infty$, which is equivalent to the case of no reflection barrier.

Second, we provide some interpretations of the two vectors $\boldsymbol{\theta}$ and $\boldsymbol{\vartheta}$ in Equation (29). For $j \in \mathcal{J}^-$, $\vartheta_j = 0$, while $-\theta_j$ represents the negative drift that brings down $Q_j^n(t)$ towards zero, due to the fact that the center of the RMOU is actually negative along the j^{th} coordinate. For $j \in \mathcal{J}^+ \cup \mathcal{J}^{++}$, $\theta_j = 0$, and ϑ_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 $\boldsymbol{\theta}$ and $\boldsymbol{\vartheta}$ in the cases when j is in $\mathcal{J}^-$, $\mathcal{J}^+$, and $\mathcal{J}^{++}$, respectively.

Finally, we want to elaborate on the relationship between our result and Theorem 7.2 in Mandelbaum et al. (1998b). Mandelbaum et al. (1998b) developed a diffusion approximation for $\sqrt{n}(\boldsymbol{x}^n(t)-\boldsymbol{x}(t))$, 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 $\boldsymbol{R}(\boldsymbol{\tau}(t))$ in the SDER (32) may have infinite values when the fluid limit $\boldsymbol{x}(t)$ passes through points at which a finite-valued Jacobian matrix does not exist. Should that happen, the sequence of $\boldsymbol{Q}^n$ may be not tight and the diffusion limit is not well defined. Thus, for our model, the diffusion limit can only be developed in a neighborhood of the fixed equilibrium state $\boldsymbol{x}^*$, at which a finite Jacobian is assumed to exist.

We next provide a high level overview for the proof of Theorem 5. To develop a diffusion approximation for $\sqrt{n}(\boldsymbol{x}^n(t)-\boldsymbol{x}^*)$, we assume that the fluid limit starts with the steady state (i.e., $\boldsymbol{x}(0)=\boldsymbol{x}^*$, or more strongly, $\boldsymbol{Q}^n(0)$ converges to a bounded random variable). Then by the definition of equilibrium, we know the fluid limit is invariant as $\boldsymbol{x}(t)\equiv\boldsymbol{x}^*$. 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 Jacobian at the equilibrium $\boldsymbol{R}^*$. So we can derive the diffusion limit as an RMOU process, which has a stationary distribution due to negative definiteness of $\boldsymbol{R}^*$. 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).

The framework introduced in Theorem 7.2 of Mandelbaum et al. (1998b) cannot be adapted to derive our Theorem 5, even by assuming $\boldsymbol{x}(0)=\boldsymbol{x}^*$ in their proof. This is because their proof framework heavily relies on the bounded derivative (or Lipschitz continuity) 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 $T>0$, $\{\boldsymbol{Q}^n(t)|t\in[0,T]\}$, as defined in (30), will be contained in a compact set with probability approaching to one when $n\to\infty$. This conclusion, nevertheless, is not valid if the arrival rate function (thus the drift function) are non-Lipschitz, where the solution to the differential equation can be unbounded. Although that differential equation is non-stochastic, adding a stochastic term will not change the boundedness of the solution. Therefore, non-Lipschitz arrival rates, if without additional constraints, may lead to a queue-lengths process that violates the compact containment condition.

To deal with the non-Lipschitz case, it suffices to prove a result analogous to Lemma 14.12 (compact containment) of Mandelbaum et al. (1998b) in Lemma 3, but for non-Lipschitz and choice-driven arrival rates. With compact containment, we can find a compact neighborhood of the equilibrium which contains the scaled stochastic processes at almost all the times for sufficiently large n. Since the drift function is Lipschitz continuous in that neighborhood, the convergence to the diffusion limit follows from Theorem 7.2 in Mandelbaum et al. (1998b).

Below we provide more details about the compact containment result. For a given $\kappa>0$, we define a compact rectangular

$$\Omega(\kappa) := [0,+\kappa]^{J^-\cup J^+}\otimes[-\kappa,+\kappa]^{J^{++}}. \tag{34}$$

Define a bounded modification of $\boldsymbol{Q}^n$ as

$$\boldsymbol{Q}^{\kappa,n}(t) = \boldsymbol{\Phi}^{\Omega(\kappa)}(\boldsymbol{Q}^n) \tag{35}$$

Intuitively, $\boldsymbol{Q}^{\kappa,n}$ is the process created from $\boldsymbol{Q}^n$ by imposing reflection barriers on the finite boundary of $\Omega(\kappa)$. We prove that in the following lemma that for any $T>0$, when $\kappa\to\infty$, with probability approaching one, $\boldsymbol{Q}^{\kappa,n}$ is contained in the bounded rectangular $\Omega(\kappa)$.

Lemma 3 *(Compact Containment) For any $T>0$, $\epsilon>0$, when $\kappa\to\infty$, we have*

$$\limsup_{n\to\infty}\Pr(\|\boldsymbol{Q}^n\|_T>\kappa)=\limsup_{n\to\infty}\Pr(\|\boldsymbol{Q}^{\kappa,n}-\boldsymbol{Q}^n\|_T\neq 0)\to 0 \tag{36}$$

To provide some intuition towards the proof of Lemma 3, we note that without the Lipschitz assumption, a small deviation of $\boldsymbol{Q}^n$ might lead to a large drift that pushes $\boldsymbol{Q}^n$ away from the equilibrium, which causes compact containment to fail. However, the choice-driven property ensures that any deviation of $\boldsymbol{Q}^n$ can only result in a drift that pulls $\boldsymbol{Q}^n$ back towards the equilibrium (even though the drift can be quite large). Thus, the choice-driven property can replace the Lipschitz condition and guarantee compact containment of $\boldsymbol{Q}^n$.

Perhaps the most useful characterization of a stochastic process is its stationary distribution. The diffusion limit process $\boldsymbol{Y}$ is an RMOU and falls into the category of multi-dimensional reflected diffusion processes, the stationary distribution of which has been studied in (Dieker and Gao, 2013; Kang and Ramanan, 2014). Based on the results of Kang and Ramanan (2014), we can derive a closed-form characterization of the stationary distribution of $\boldsymbol{Y}$ under additional assumptions.

To facilitate the presentation of the proposition, we use $\varphi(\cdot,\boldsymbol{m},\boldsymbol{\Sigma})$ to denote the density of a multivariate Gaussian distribution with mean $\boldsymbol{m}$ and covariance matrix $\boldsymbol{\Sigma}$. We use $\varphi(\cdot,\boldsymbol{m},\boldsymbol{\Sigma},\boldsymbol{r})$ to denote the density of a multivariate truncated Gaussian distribution defined within the domain $\{\boldsymbol{y}\in\mathbb{R}^J\,|\,\boldsymbol{y}\geq\boldsymbol{r}\}$, which has the following expression,

$$\varphi(\boldsymbol{z},\boldsymbol{m},\boldsymbol{\Sigma},\boldsymbol{r})=\begin{cases}\frac{\varphi(\boldsymbol{z},\boldsymbol{m},\boldsymbol{\Sigma})}{\int_{\boldsymbol{x}\geq\boldsymbol{r}}\varphi(\boldsymbol{x},\boldsymbol{m},\boldsymbol{\Sigma})\,d\boldsymbol{x}} & \text{if } \boldsymbol{z}\geq\boldsymbol{r},\\ 0 & \text{otherwise.}\end{cases} \tag{37}$$

We define $\mathrm{vec}(\boldsymbol{A})$ as the column vector obtained by stacking the columns of matrix $\boldsymbol{A}$. For two J-by-J matrices, the Kronecker product and Kronecker sum are defined respectively as follows

$$\boldsymbol{A}\otimes\boldsymbol{B}=(a_{ij}\boldsymbol{B})_{i,j=1,\ldots,J},\qquad \boldsymbol{A}\oplus\boldsymbol{B}=\boldsymbol{A}\otimes I+I\otimes\boldsymbol{B}, \tag{38}$$

where I is the J-by-J identify matrix.

Proposition 4 *(Stationary Distribution of the Diffusion Limit) In the case of $\mathcal{J}^+\cup\mathcal{J}^-=\emptyset$, the multi-dimensional O-U (MOU) process $\boldsymbol{Y}$ admits a unique stationary distribution $\pi(\cdot)=\varphi(\cdot,\boldsymbol{\vartheta},\boldsymbol{\Sigma_\infty})$, where $\boldsymbol{\Sigma_\infty}$ is determined by the equation*

$$vec(\Sigma_\infty)=-(vec(\boldsymbol{R}^*\mathrm{Diag}\,(\boldsymbol{\mu}^{-1}))\oplus vec(\boldsymbol{R}^*\mathrm{Diag}\,(\boldsymbol{\mu}^{-1})))^{-1}vec(\mathrm{Diag}\,((1+\omega_j^2)\mu_j)_{j=1,\ldots,J}). \tag{39}$$

In the case of $\mathcal{J}^+\cup\mathcal{J}^-\neq\emptyset$, provided that $\boldsymbol{R}^*$ is symmetric and $\omega_j\equiv\omega_1$ for all j, the reflected MOU (RMOU) process $\boldsymbol{Y}$ admits a unique distribution $\pi(\cdot)=\varphi(\cdot,\boldsymbol{\vartheta}+\mathrm{Diag}\,(\boldsymbol{\mu})(\boldsymbol{R}^*)^{-1}\boldsymbol{\theta},-\frac{1}{2}(1+\omega_1^2)\mathrm{Diag}\,(\boldsymbol{\mu})(\boldsymbol{R}^*)^{-1}\mathrm{Diag}\,(\boldsymbol{\mu}))$.

In scenarios where $\mathcal{J}^+\cup\mathcal{J}^-\neq\emptyset$, the symmetry of $\boldsymbol{R}^*$ combined with the condition $\omega_j\equiv\omega_1$ for all j becomes indispensable for deriving a closed-form expression for the stationary distribution, as elucidated in Example 3.10, Claim 1 by Kang and Ramanan (2014). Absent these conditions, while the stationary distribution still exists, determining its density may necessitate a numerical approach. Typical approaches include solving the associated Fokker-Planck partial differential equations with reflecting boundary conditions (e.g., Equation 4.115 of (Pavliotis, 2014)) using a finite difference method (Grossmann, 2007), or simulating the diffusion limit process using a Markov Chain Monte Carlo method (Pavliotis, 2014).

The multivariate Gaussian steady-state distribution provides the system manager with some practical insights. Since the covariance matrix of such a distribution is proportional to the inverse of the Jacobian $(\boldsymbol{R}^*)^{-1}$, the spread of the distribution is decreasing in the scale of $\boldsymbol{R}^*$. Thus if

one wishes to reduce the variability of the queue-length process of the DCPQ, one may consider increasing the scale of $\boldsymbol{R}^*$, which depends on customers' delay sensitivity. Roughly, if customers are more sensitive to the non-zero waiting times (so a larger c_ξ), then $\boldsymbol{R}^*$ 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. To increase customers' sensitivity to waiting times, system manager can disclose information about real-time waiting times to prospective customers, an approach that will be detailed in Section 8.

Proposition 4 states that characterize the stationary distribution of the limiting process $\boldsymbol{Y}$. We next investigate the asymptotic behaviour of the stationary distribution of $\boldsymbol{Q}^n(\cdot)$, the scaled queue-length process in the n^{th} DCPQ. However, $\boldsymbol{Q}^n$ is not Markovian as the probability transition depends on the remaining service time of the customer currently being served. For this reason, we study a Markov process $\boldsymbol{\Xi}^n(\cdot) := (\boldsymbol{Q}^n(\cdot), \boldsymbol{s}^n(\cdot))$ instead of $\boldsymbol{Q}^n(\cdot)$, where $\boldsymbol{s}^n(t) := (s_j^n(t))$ and $s_j^n(t)$ denotes the remaining service time of the customer currently being served by the j^{th} SP at time t in the n^{th} DCPQ. Let $\boldsymbol{\pi}^n$ denote the projection of the stationary distribution of $\boldsymbol{\Xi}^n(\cdot)$ onto the coordinates of $\boldsymbol{Q}^n(\cdot)$. Then we can prove that $\boldsymbol{\pi}^n$ weakly converges to $\boldsymbol{\pi}$ when n approaches infinity. This result is also termed as *interchange of limits* and illustrated in Figure 4.

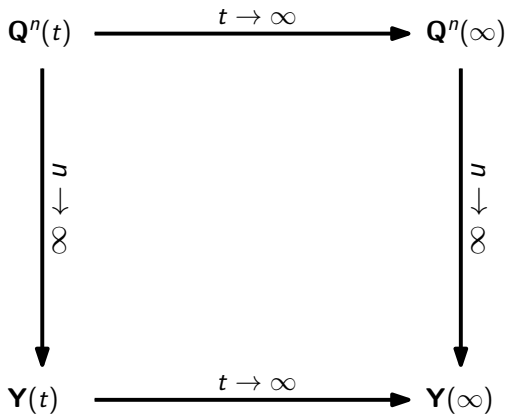


Figure 4 The interchange-of-limit result implies that the steady-State distribution of $\boldsymbol{Y}^n$, $\boldsymbol{\pi}$, can be approximated by $\boldsymbol{\pi}^n$, the projection of the steady-state distribution of $\boldsymbol{\Xi}^n$ onto the subspace of $\boldsymbol{Q}^n$.

The interchange of limits was proved when $\boldsymbol{\Xi}^n$ is the Markov process in a generalized Jackson network by Gamarnik and Zeevi (2006). We adopt their machinery and show that the interchange of limits holds for the DCPQ. The queueing network considered by Gamarnik and Zeevi (2006) assume constant arrival and service rates, while the arrival rates in our model are state-dependent and non-Lipschitz. Therefore, the adoption of their methods is not trivial and must exploit the CD property. Specifically, the choice-driven property is used to prove that a *Lyapunov function* can be constructed so that its exponential has bounded expectation.

Formally, 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 κ for a Markov process $\boldsymbol{\Xi}$ if

$$\sup_{\boldsymbol{\Xi}(0)\in\Omega:\ V(\boldsymbol{\Xi}(0))>\kappa} \{\mathbb{E}_{\boldsymbol{\Xi}(\mathbf{0})} V(\boldsymbol{\Xi}(t_0)) - V(\boldsymbol{\Xi}(0))\} \le -\gamma. \tag{40}$$

For each n, define

$$\begin{aligned} L_1(u,t,n) &:= \sup_{\boldsymbol{\Xi}^n(0)\in\Omega} \mathbb{E}[\exp(u(V(\boldsymbol{\Xi}^n(t)) - V(\boldsymbol{\Xi}^n(0))))|\boldsymbol{\Xi}^n(0)] \\ L_2(u,t,n) &:= \sup_{\boldsymbol{\Xi}^n(0)\in\Omega} \mathbb{E}[(V(\boldsymbol{\Xi}^n(t)) - V(\boldsymbol{\Xi}^n(0)))^2 \exp(u(V(\boldsymbol{\Xi}^n(t)) - V(\boldsymbol{\Xi}^n(0)))^+)|\boldsymbol{\Xi}^n(0)] \end{aligned} \tag{41}$$

for any $u>0$, $t\geq 0$. We then have the following proposition.

Proposition 5 *Let $V(\boldsymbol{\Xi}^n(t)) := \|\boldsymbol{Q}^n(t)\|^{\boldsymbol{\mu}^{-1}}$. Then for sufficiently large n, $V(\cdot)$ is a Lyapunov function with drift size parameter -1, drift time parameter t_0, and exception parameter κ for some $\kappa, t_0>0$. In addition, there exists u_0 such that*

$$\begin{aligned} \limsup_{n\to\infty} L_1(u_0,t_0,n) &< \infty \\ \limsup_{n\to\infty} L_2(u_0,t_0,n) &< \infty \end{aligned} \tag{42}$$

The above proposition is in analogue to Proposition 3 in Gamarnik and Zeevi (2006), but deals with the DCPQ case in which the drift function is not Lipschitz. Note that we have used different notations from those used in Gamarnik and Zeevi (2006): our $\boldsymbol{Q}^n(t)$ corresponds to the notation "$\frac{1}{\sqrt{n}}\boldsymbol{Q}^n(nt)$" in their paper. Because we have used a different scale, the bound we derived with respect to the $\|\cdot\|_{t_0}$ norm is exactly the bound derived in their paper the interval $[0, nt_0]$.

Theorem 6 *(Interchange of Limit) The sequence of stationary distributions, π^n, weakly converges to π.*

The main idea of the proof is to construct a Lyapunov function with the properties given in Proposition 5. Those properties allow us to prove uniform tightness of the sequences $(\boldsymbol{\pi}^n)$, which then yields the existence of a limiting distribution $\hat{\boldsymbol{\pi}}$. The interchange of limits can then be proved by arguing that any such $\hat{\boldsymbol{\pi}}$ must coincide with the unique stationary distribution of $\boldsymbol{Y}$, $\boldsymbol{\pi}$.

8. Value of Waiting Time Information in DCPQ

In the previous sections, we assume that all customers in DCPQ have access to real-time information on queue lengths or waiting times upon arrival. We now turn our attention to a variant where only a subset of customers observe the real-time waiting times, $\boldsymbol{\tau}(\cdot)$. Those without access to current waiting times rely instead on steady-state expected waiting times as their decision-making basis, an assumption commonly adopted in the literature (Guo and Zipkin, 2007; Ata et al., 2021).

We introduce the concept of *η-informed DCPQ*, characterized by a proportion, $\eta\in[0,1]$, of customers who have access to real-time waiting time information. We consider a sequence of η-informed DCPQs, similar to that in Section 7, with arrival and service rates scaled by $n=1,2,\ldots$ We extend our notations to include an η subscript to denote variables and processes specific to η-informed DCPQs. For instance, $\boldsymbol{\tau}^{\eta,*}$ denotes the fluid equilibrium for the original η-informed DCPQ, whereas $\boldsymbol{\tau}^{\eta,n,*}$ represents the virtual equilibrium in the n^{th} η-informed DCPQ.

The next proposition shows that the fluid equilibrium in η-informed DCPQ remains consistent with $\boldsymbol{\tau}^*$, the equilibrium in the traditional DCPQ where $\eta=1$. The consistency is rooted in the observation that the expected waiting times at virtual equilibrium align with those in steady state at the fluid scale. Therefore, the decision-making processes of both informed and uninformed customers are identical at the virtual equilibrium. The proof is omitted as it is straightforward following the above logic.

Proposition 6 *For all $\eta\in[0,1]$, $\boldsymbol{\tau}^{\eta,*}\equiv\boldsymbol{\tau}^*$ and $\boldsymbol{\tau}^{\eta,*,n}\equiv\boldsymbol{\tau}^{\eta,*}$ for all n.*

We delve into the nuances of waiting time disclosure at the diffusion level, focusing on the scaled queue-lengths process within the n^{th} η-informed DCPQ:

$$\boldsymbol{Q}^{\eta,n}(t) := \sqrt{n}\left(\frac{1}{n}X_j^{\eta}(t) - \tau_j^*\mu_j\right). \tag{43}$$

To derive the diffusion limit of $\boldsymbol{Q}^{\eta,n}(t)$, we have to make an assumption that all queues are largely over-demand, that is, $\boldsymbol{\tau}^* > \mathbf{0}$, or equivalently, $\mathcal{J}^- \cup \mathcal{J}^+ = \emptyset$. Under this assumption, we will show that $\boldsymbol{Q}^{\eta,n}$ weakly converges to a diffusion limit $\boldsymbol{Y}^\eta$, the solution to the following SDER,

$$\boldsymbol{Y}^\eta(t) = \int_0^t \eta \left(\boldsymbol{R}^* \text{Diag}\left(\boldsymbol{\mu}^{-1}\right)(\boldsymbol{Y}(s) - \boldsymbol{\vartheta})\right) ds + \boldsymbol{\Sigma}^{1/2}\boldsymbol{B}(t) + \boldsymbol{L}(t), \tag{44}$$

with $\boldsymbol{\Sigma}^{1/2}$ retaining its definition in Section 7. Notably, the only difference between $\boldsymbol{Y}^\eta$ and $\boldsymbol{Y}$, the diffusion limit in the original system, lies in the drift of $\boldsymbol{Y}^\eta$ having been scaled by η.

We have to assume $\boldsymbol{\tau}^* > 0$, because otherwise, queues in $\mathcal{J}^- \cup \mathcal{J}^+$ have nonzero expected waiting times at the diffusion scaling, which deviates from their fluid equilibrium $\tau_j^* = 0$. This deviation presents challenges in modeling customers' decisions and deriving the diffusion limit. However, the assumption of $\boldsymbol{\tau}^* > 0$ aligns well with the goal of welfare maximization in DCPQ systems. This is supported by Corollary 1, which suggests that the system manager allocates the limited service capacity to achieve a state where $\boldsymbol{\tau}^* > 0$.

Proposition 7 *Suppose the fluid equilibrium $\boldsymbol{\tau}^* > 0$. For any $\eta \in [0,1]$, if $\boldsymbol{Q}^{\eta,n}(0) \Rightarrow \boldsymbol{Y}^\eta(0)$ and $\mathbb{E}\|\boldsymbol{Y}^\eta(0)\| < \infty$, we have*

$$\boldsymbol{Q}^{\eta,n} \Rightarrow \boldsymbol{Y}^\eta. \tag{45}$$

According to Proposition 7, the dynamics of $\boldsymbol{Y}^\eta$ differ by the value of η. When $\eta \in (0,1]$, $\boldsymbol{Y}^\eta$ is an MOU process, which has a unique stationary distribution due to its mean-reverting property. In contrast, when $\eta = 0$, the absence of a linear drift transforms $\boldsymbol{Y}^0$ into a multi-dimensional Brownian motion, which is not positively recurrent and does not have a stationary distribution.

The theoretical foundations laid out in Proposition 4 and Theorem 6, which address the interchange of limits and the derivation of a closed-form stationary distribution, remain applicable to the η-informed DCPQ scenario with a simple adjustment. By substituting $\boldsymbol{R}^*$ with $\eta\boldsymbol{R}^*$ in previous proofs, these results can be seamlessly extended to accommodate the varying degrees of customer information.

Corollary 2 *When $\boldsymbol{\tau}^* > 0$ and $\eta \in (0,1]$, we have $\boldsymbol{Q}^{\eta,n}(\infty) \Rightarrow \boldsymbol{Y}^\eta(\infty)$, where $\boldsymbol{Y}^\eta(\infty)$ follows a multivariate Gaussian distribution, possessing a mean vector $\boldsymbol{\vartheta}$ and a covariance matrix $\eta^{-1}\boldsymbol{\Sigma}_\infty$, with $\boldsymbol{\Sigma}_\infty$ defined in* (39).

Knowing the stationary distribution of diffusion limit in the η-informed DCPQ, we are ready to calculate the expected social welfare at steady state. Let $SW^{\eta,n}$ denote the expected social welfare associated with $\boldsymbol{Y}^{\eta,n}(\infty)$. Let $SW(\boldsymbol{\tau}) := \mathbb{E}[U_{max}(\boldsymbol{\tau})]$ denote the expected social welfare when the DCPQ is at state $\boldsymbol{\tau}$. The subsequent theorem quantifies the asymptotic influence of the information level η on DCPQ's social welfare.

Theorem 7 *Suppose $\boldsymbol{\tau}^* > 0$. For all $\eta \in (0,1]$, we have*

$$SW^{\eta,n} - SW(\boldsymbol{\tau}^*) \equiv \frac{C}{n} + o(\frac{1}{n}). \tag{46}$$

for some constant $C > 0$.

Theorem 7 reveals an intriguing insight: in a DCPQ system with no empty queues, the proportion of customers informed about real-time waiting time has almost no impact on social welfare ($o(1/n)$). This outcome emerges because uninformed customers default to the fluid equilibrium for decision-making, rendering the incremental benefit of real-time information marginal (expressed as C/n). Although having a greater proportion of informed customers might intuitively seem advantageous,

it actually diminishes queue-length variability, inadvertently impacting social welfare negatively due to its convexity with queue lengths. This finding diverges notably from existing results on single-queue systems, where disseminating real-time waiting time information typically enhances social welfare compared to scenarios where customers rely on equilibrium waiting times for decision-making, when the system is congested (Hassin, 1986a; Chen and Frank, 2004; Guo and Zipkin, 2007).

9. DCPQ with Reneging Customers

The previous results on DCPQ without reneging customers can be extended to the case when customers may renege (or abandon) after an exponentially distributed time before getting served. Note that in most past studies on the queues with customer choice, the reneging feature was not considered due to the reason that a customer's decision to join was made based on the expected service utility. We incorporate reneging, as it is a feature in our motif dating examples. For example, in health care settings, death or unexpected changes in medical conditions may lead to abandonment of the service by patients. Since the analysis with reneging is similar to the one in earlier sections, we only elaborate the results where the technical differences are significant.

We assume that customers renege after an exponentially distributed period with mean of $1/d$. We consider a Markovian system in which the inter-arrival times, reneging times, and service times are all exponentially distributed. In this case, the offered waiting time (i.e., waiting time conditional on that the customer receives service before reneging) can be estimated using the following asymptotic formula adapted from Eq. (19) of Zenios (1999),

$$\tau_j(t) = \frac{1}{d}\log(1 + \frac{X_j(t)d}{\mu_j}). \tag{47}$$

We assume that all customers use (47) 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 $\mathbf{\Lambda}(\boldsymbol{\tau})$. Because our proof for Proposition 1 does not rely on the functional form of τ_j with respect to X_j, the proof can be adapted to establishing the choice-driven property of the arrival rate function in the presence of reneging customers.

Corollary 3 *With the customer choice model defined in Section 3, even if customers renege after an exponentially distributed time with mean $1/d$ before service, the arrival rate function still satisfies the CD property, and its Jacobian is symmetric.*

With reneging, the DCPQ is always stable. So the stability condition (7) is no longer necessary.

We next prove that the fluid process in a DCPQ 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 8 *The equilibrium state of the fluid limit process in DCPQ with reneging is the unique solution to the following Nonlinear Complementary Problem (NCP).*

$$\begin{aligned} NCP \qquad & Z_j := \mu_j \exp(\tau_j d) - p_j(\boldsymbol{\tau}) \geq 0, \quad \textit{for } j = 1, \ldots, J. \\ & \tau_j \geq 0, \quad \textit{for } j = 1, \ldots, J. \\ & \tau_j Z_j = 0, \quad \textit{for } j = 1, \ldots, J. \end{aligned} \tag{48}$$

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

Proof. By defining $\hat{p}_j := \mu_j \exp(\tau_j d) - p_j(\boldsymbol{\tau})$, the above NCP can be rewritten into a similar form as in(18) by replacing the arrival function $\boldsymbol{\Lambda}(\cdot)$ with $\hat{\boldsymbol{\Lambda}}(\cdot) := (\hat{p}_j(\cdot))_{j=1,\ldots,J}$. Note that the Jacobian for $\hat{\boldsymbol{\Lambda}}(\boldsymbol{\tau})$ has the form $\hat{R} = \sigma(\boldsymbol{\tau}) + R$, where R is the Jacobian of $p(\boldsymbol{\tau})$ and is a symmetric negative definite matrix by Corollary 3, and $\sigma(\boldsymbol{\tau})$ is a diagonal matrix with the j^{th} entry $\sigma_{jj}(\boldsymbol{\tau}) = \mu_j d \exp(\boldsymbol{\tau}_j d) > 0$. Because of the extra term $\sigma(\boldsymbol{\tau})$, we are now able to prove that $\hat{p}_j$ satisfies the uniform P-property, i.e.,

$$\text{Uniform P-Property: } \forall \boldsymbol{\tau}^1, \boldsymbol{\tau}^2 \in \mathbb{R}^J_+,\ \boldsymbol{\tau}^1 \neq \boldsymbol{\tau}^2,\ \min_{j=1}^{J}\ (\tau_j^1 - \tau_j^2)(\hat{p}_j(\boldsymbol{\tau}^1) - \hat{p}_j(\boldsymbol{\tau}^2)) < c\|\boldsymbol{\tau}^1 - \boldsymbol{\tau}^2\|^2, \quad (49)$$

with $c > d \max_j \mu_j > 0$. Thus, the classical theorem by Cottle (1966) implies the existence of a unique solution to the NCP (48).

To prove $\boldsymbol{\tau}(t) \to \boldsymbol{\tau}^*$, we define $\overline{\Delta}\boldsymbol{\tau}(t) = \max_j(\tau_j(t) - \tau_j^*)$, and $\underline{\Delta}\boldsymbol{\tau}(t) = \min_j(\tau_j(t) - \tau_j^*)$. We want to prove that $\overline{\Delta}\boldsymbol{\tau}'(t) \leq \kappa(\delta)$ for some constant $\kappa(\delta) > 0$ whenever $\overline{\Delta}\boldsymbol{\tau}(t) \geq \delta$. Without loss of generality, assume that $\tau_j(t) - \tau_j^* = \overline{\Delta}\boldsymbol{\tau}'(t)$, then $\tau_j(t) > 0$ and (47) imply that

$$\tau_j'(t) = \frac{p_j(\boldsymbol{\tau}) - \mu_j}{X_j(t)d + \mu_j} \leq \frac{p_j - \mu_j}{p_j(\boldsymbol{\tau}^*)}, \quad (50)$$

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

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

The proof of the convergence to the diffusion limit is a simple extension of Theorem 5 by including an extra term $-dI$ in the drift matrix as a result of reneging. We summarize the result below and the notations follow from the definitions in the previous sections.

We next study the diffusion approximation for $\boldsymbol{Q}^n(\cdot) := (Q_j^n(\cdot))$, where $Q_j^n(\cdot)$ is the scaled queue-length process in the n^{th} DCPQ with its expression given in Equation (30). As before we partition the index set of queues into four subsets $\mathcal{J}^{--}$, $\mathcal{J}^-$, $\mathcal{J}^+$, and $\mathcal{J}^{++}$ according to (28) and assume $\mathcal{J}^{--} = \emptyset$. We redefine $\boldsymbol{\theta}$ and $\boldsymbol{\vartheta}$ as follows,

$$\theta_j = \begin{cases} \lim_{n\to\infty} \sqrt{n}(\mu_j^n - p_j(\boldsymbol{\tau}^{n,*})) & \text{if } j \in \mathcal{J}^- \\ 0 & \text{otherwise,} \end{cases} \quad \vartheta_j = \begin{cases} \lim_{n\to\infty} \sqrt{n}(x_j^{n,*} - X_j^*) & \text{if } j \in \mathcal{J}^+ \cup \mathcal{J}^{++} \\ 0 & \text{otherwise,} \end{cases} \quad (51)$$

where $X_j^{n,*} = \frac{\mu_j^n}{d}(\exp(d\tau_j^{n,*}) - 1)$ represents the queue length when the expected waiting time is $\tau_j^{n,*}$, and simiarly $x_j^* := \frac{\mu_j}{d}(\exp(d\tau_j^*) - 1)$ denotes the equilibrium queue length in the fluid model. The scaled-queue length process $\boldsymbol{Q}^n(\cdot)$ and its diffusion limit $\boldsymbol{Y}(\cdot)$ thus reside in the following domain,

$$\Omega = [0, +\infty)^{J^- \cup J^+} \otimes (-\infty, +\infty)^{J^{++}}. \quad (52)$$

The next Corollary, which is analogous to Theorem 5, states that $\boldsymbol{Q}^n$ converges to a J-dimensional diffusion process $\boldsymbol{Y}$ which is the solution to the following stochastic-differential-equation,

$$\boldsymbol{Y}(t) = \int_0^t \left((\boldsymbol{R}^* \text{Diag}\left((\mathbf{exp}(\boldsymbol{\tau}^*\boldsymbol{d}) \circ \boldsymbol{\mu})^{-1}\right) - dI)(\boldsymbol{Y}(s) - \boldsymbol{\vartheta}) - \boldsymbol{\theta}\right) ds + \int_0^t \boldsymbol{\Sigma}^{R,1/2} d\boldsymbol{B}(s) + \boldsymbol{L}(t), \quad (53)$$

where I is an J-by-J identity matrix, $\boldsymbol{\Sigma}^{R,1/2}$ is a J-by-J diagonal matrix with $\sqrt{(\omega_j^2 + \exp(\tau_j^* d))\mu_j}$ as its j^{th} diagonal entry, $\boldsymbol{B}(t)$ is a J-dimensional Brownian motion, and $\boldsymbol{L}(t)$ is a J-dimensional minimal non-decreasing process which makes $Y_j(t) \geq 0$ for all $j \in \mathcal{J}^- \cup \mathcal{J}^+$.

Corollary 4 *Suppose* $\boldsymbol{Q}^n(0) \Rightarrow \boldsymbol{Y}(0)$ *and* $\mathbb{E}\|\boldsymbol{Y}(0)\| < \infty$. *Then we have*

$$\boldsymbol{Q}^n \Rightarrow \boldsymbol{Y}. \tag{54}$$

The proof for Corollary 4 is provided in Appendix R.

10. Case Study

We illustrate the applications of the theoretical results in a real life parallel-queue system and caliber our model using real data. We consider automobile queues at the two U.S.-Canada border-crossing ports of entries at the west coast, i.e., Peace Arch and Pacific crossings. The two ports are located within 2 miles of each other and an automobile can cross the border via either port by choosing the corresponding exit to leave the highway. Figure 5 visualizes the geographic locations of the two ports.

Figure 5 The Peace Arch and Pacific Border-Crossings

To cross the border, every vehicle needs to be screened by an officer at an inspection booth. This process takes a few minutes and creates a bottleneck or a queue for the border-crossing traffic. There are a maximum of eight booths at each port of entry and the number of open booths varies across a day. Since these booths are located next to each other, a vehicle can choose one of the open booths to cross the border. Thus, all vehicles at the same port of entry are in a pooled queue, regardless which booth they actually go through. However, the vehicles at one port of entry cannot switch to the other, so vehicles at the two ports of entry form two separated parallel queues.

The up-to-date waiting times are estimated from the data gathered by nearby loop detectors, capturing vehicle queue lengths, moving speed, and crossing durations (WCOG, 2019). Rajbhandari et al. (2012) provided a comprehensive review on the detailed methodologies for estimating waiting time at the U.S.-Mexico border. These waiting time estimates are updated every 5 minutes, ensuring timely information for potential passengers. Access to these updates is facilitated through

various channels: the website (WSDOT, 2024), which posts updates every 5 minutes, the in-vehicle radio broadcasts at intervals of 10 minutes or less, and digital message boards along the highway (Interstate-5) near the exists to the two crossings.

Travellers can then cross the border through either Peace Arch or Pacific. Thus, the vehicle queues at the two crossings can be modeled as DCPQ with $J = 2$. We next use the historical border-crossing traffic data to calibrate our customer choice model.

Our data is collected from the public website (WCOG, 2019). It records the number of arrivals in five minute intervals at each port of entry, denoted by $a_{pe}(t)$ and $a_{pa}(t)$, and waiting-time (delay) estimates at the beginning of every five-minute interval, denoted by $\tau_{pe}(t)$ and $\tau_{pa}(t)$. Here $t = 1, 2, \ldots$ denotes the index of each five-minute interval. Commercial trucks and vehicles with a special dedicated fast lane such as NEXUS, go through separated lanes and are not included in this tally. Anecdotal evidence suggests that some vehicles indeed balk upon observing a long queue at the ports. However, the exact number of balked vehicles cannot be tracked because a vehicle can balk anywhere on its way to the crossing.

Our data calibration is based on the northbound border-crossing traffic data in a one-year study period from February 2018 to January 2019. To control the potential seasonal effect, we divide the study period into four seasons: Feb-Apr, May-July, August-October, and November-January. Months with similar intra-day arrival patterns are grouped into the same season. To control the day-of-week effect, we only use traffic data on Tuesday, Wednesday, and Thursday, because the arrival patterns in these days are very similar (Yu et al., 2016). Since travellers may pay more attention to the waiting time estimates when there is a substantial delay, we focus on traffics during the peak hours. To that end, we select a fixed 2.5-hours time window among those days in each season, during which the arrival rate reaches a plateau. See Appendix S for how the 2.5-hour time windows are selected.

After a traveller learns about the waiting-time estimates, it typically takes her less than ten minutes till his vehicle joins the queue at a port of entry and is counted as an arrival. Thus, when predicting the choice probability at the beginning of the t^{th} five-minute slot, we should use the waiting time estimates at the beginning of the $(t-\Delta)^{th}$ slot. In our numerical experiments, as a robustness check we have tested $\Delta = 0, 1, 2$ to capture the possible time lags of 0, 5, and 10 minutes, respectively.

We want to study the effect of waiting time on travellers' queue-joining behavior. A simple model – the conditional logit model – is sufficient to serve that purpose. In fact, given the limited number of explanatory variables in the dataset, employing a random coefficient model could lead to overfitting.

According to the logit model, the probability for a passenger to choose Peace Arch instead of Pacific, conditional on that the passenger would not balk, can be calculated as follows,

$$\begin{aligned} \frac{p_{pe}(t)}{p_{pe}(t)+p_{pa}(t)} &= \frac{\exp(v_{pe}-c\tau_{pe}(t-\Delta))}{\exp(v_{pe}-c\tau_{pe}(t-\Delta))+\exp(v_{pa}-c\tau_{pa}(t-\Delta))} \\ &= \frac{1}{1+\exp((v_{pa}-v_{pe})-c(\tau_{pa}(t-\Delta)-\tau_{pe}(t-\Delta)))}. \end{aligned} \tag{55}$$

where $p_{pe}(t)$ and $p_{pa}(t)$ denote the proportion of travellers who choose Peace Arch and Pacific crossing at time t, respectively, v_{pe} and v_{pa} denote the expected service utility, excluding the waiting cost, at Pacific and Peace Arch, respectively, and c denotes the waiting cost per minute. Although we do not have data on the number of balking vehicles, we can derive the nonlinear least square estimator for $\hat{v}_{pa} - \hat{v}_{pe}$ and $\hat{c}$ as

$$(\hat{v}_{pa} - \hat{v}_{pe}, \hat{c}) := \arg\min\left(\frac{1}{1+\exp((v_{pa}-v_{pe})-c(\tau_{pa}(t-\Delta)-\tau_{pe}(t-\Delta)))} - \frac{a_{pe}(t)}{a_{pe}(t)+a_{pa}(t)}\right)^2. \tag{56}$$

The coefficients estimated from the conditional logit model, log-likelihood, and number of observations are summarized in Table 1. For all the four seasons and time lags $\Delta = 0, 1, 2$, the estimator

of waiting cost $\hat{c}$ is consistently positive at a 0.001 significance level. Moreover, the odds ratio of c spans from 1.003 to 1.014, indicating that each minute decrease in waiting time increases the odds of joining that queue by $3\% - 14\%$. That provides strong evidence that travellers have paid attention to the waiting time estimates and tried to avoid longer queues during the peak hours.

We also find that travellers' preference, excluding waiting cost effect, changes between Peace Arch and Pacific differ from season to season. From August till January, the coefficient estimator $\hat{v}_{pa} - \hat{v}_{pe}$ stays negative at a 0.001 significance level, suggesting that more travellers prefer Peace Arch to Pacific during those months. However, from February till May, Pacific becomes the preferred crossing. From June till August, the two ports are equally preferred. The dataset size varies slightly across different Δ scenarios, as we exclude the initial Δ-minute period from each analysis to accommodate the delay. The fluctuations in log-likelihood values across various scenarios align with the respective adjustments in the size of the dataset.

Table 1 Estimation Results from the Logit Model

		Coefficients (Standard Error)			Odds ratio		
Time Lag		0 min	5 min	10 min	0 min	5 min	10 min
Nov-Jan	$\hat{v}_{pa} - \hat{v}_{pe}$	-0.124***	-0.118 ***	-0.118***	0.883	0.889	0.888
		(0.008)	(0.008)	(0.008)			
	$\hat{c}$	0.003***	0.006***	0.006***	1.004	1.006	1.006
		(0.001)	(0.001)	(0.001)			
	Log-Lik	-3503.3	-3483.2	-3483.99			
	N	1139	1100	1062			
Feb-Apr	$\hat{v}_{pa} - \hat{v}_{pe}$	0.032***	0.034***	0.034***	1.032	1.034	1.034
		(0.008)	(0.008)	(0.008)			
	$\hat{c}$	0.013***	0.016***	0.017***	1.014	1.016	1.017
		(0.001)	(0.001)	(0.001)			
	Log-Lik	-3298.8	-3235.4	-3206.7			
	N	1020	986	952			
May-Jul	$\hat{v}_{pa} - \hat{v}_{pe}$	-0.003	0.005	0.010	0.997	1.005	1.010
		(0.007)	(0.007)	(0.007)			
	$\hat{c}$	0.008***	0.010***	0.012***	1.008	1.010	1.012
		(0.001)	(0.001)	(0.001)			
	Log-Lik	-3942.1	-3871.8	-3820.1			
	N	1200	1160	1120			
Aug-Oct	$\hat{v}_{pa} - \hat{v}_{pe}$	-0.227***	-0.217 ***	-0.214***	0.797	0.805	0.807
		(0.009)	(0.009)	(0.009)			
	$\hat{c}$	0.003***	0.005***	0.006***	1.003	1.005	1.006
		(0.001)	(0.001)	(0.001)			
	Log-Lik	-3185.1	-3171.1	-3163.1			
	N	1050	1015	980			

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

For robustness check, we also estimate a conditional logit model and obtain similar coefficient estimates and log-likelihood values; see Table 2.

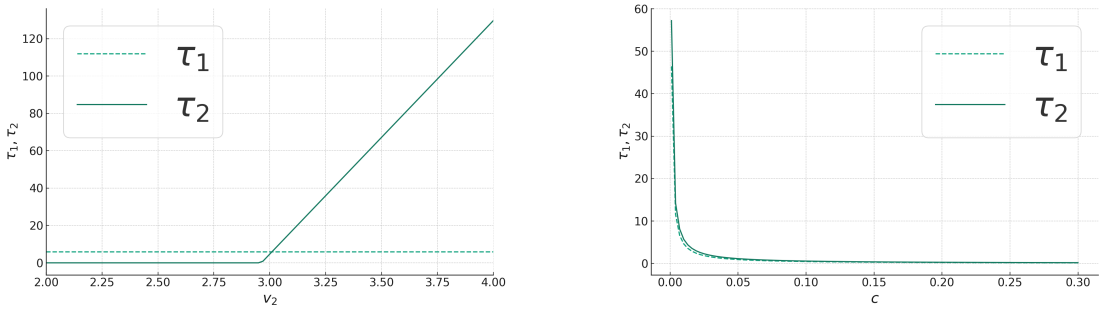
In our forthcoming analysis, we delve into numerical illustrations derived from the choice model configured for the period of Feb-Apr with $\Delta = 0$, with the other inferred from data, if possible. The numerical examples validate the robustness of our theoretical results in the fluid and diffusion analysis, while also shedding light on additional observations. Specifically, we explore variations

Table 2 Estimation Results from the Probit Model

		Coefficients (Standard Error)			Odds ratio		
Time Lag		0 min	5 min	10 min	0 min	5 min	10 min
Nov-Jan	$\hat{v}_{pa} - \hat{v}_{pe}$	-0.078***	-0.074 ***	-0.074***	0.925	0.929	0.929
		(0.005)	(0.005)	(0.005)			
	$\hat{c}$	0.002***	0.004***	0.003***	1.002	1.004	1.004
		(0.000)	(0.000)	(0.000)			
	Log-Lik	-3503.3	-3483.2	-3484.0			
	N	1139	1100	1062			
Feb-Apr	$\hat{v}_{pa} - \hat{v}_{pe}$	0.020***	0.021 ***	0.021***	1.020	1.021	1.022
		(0.005)	(0.005)	(0.005)			
	$\hat{c}$	0.008***	0.010***	0.010***	1.008	1.010	1.010
		(0.000)	(0.000)	(0.000)			
	Log-Lik	-3298.8	-3235.4	-3206.8			
	N	1020	986	952			
May-Jul	$\hat{v}_{pa} - \hat{v}_{pe}$	-0.002	0.003	0.006	0.998	1.003	1.006
		(0.005)	(0.005)	(0.005)			
	$\hat{c}$	0.005***	0.006***	0.007***	1.005	1.007	1.007
		(0.000)	(0.000)	(0.000)			
	Log-Lik	-3942.1	-3871.9	-3820.2			
	N	1200	1160	1120			
Aug-Oct	$\hat{v}_{pa} - \hat{v}_{pe}$	-0.142***	-0.136 ***	-0.134***	0.867	0.873	0.875
		(0.005)	(0.005)	(0.005)			
	$\hat{c}$	0.002***	0.003***	0.004***	1.002	1.003	1.004
		(0.001)	(0.001)	(0.001)			
	Log-Lik	-3185.1	-3171.1	-3163.1			
	N	1050	1015	980			

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

in the fluid equilibrium $\boldsymbol{\tau}^*$ in response to alteration in key model parameters, including c and $v_{pa} - v_{pe}$; see Figure 6(b) for a graphical illustration.



(a)$\boldsymbol{\tau}^*$ varies by with v_2 (fix $v_1 = 3, c = 0.008$) (b)$\boldsymbol{\tau}^*$ varies by with c (fix $v_1 = 3$, $v_2 = 3.02$)

Figure 6 Sensitivity of Fluid Equilibrium to Model Coefficients

Next we delve into how the expected social welfare (SW) changes to different service rates μ_1 and μ_2, subject to the budget constraint $\mu_1 + \mu_2 \leq \overline{\mu} = 0.9747$. The budget constraint is pertinent

to planning at border crossings, where the allocation of security check personnel, which is the service bottleneck, between two ports does not entail additional expenses. By computing the fluid equilibrium from the NCP (18), we ascertain that $\boldsymbol{\tau}^* > 0$ when $0.463 < \mu_1 < 0.508$. This range correlates with the plateau on top of the $SW_{logit}(\mu_1)$ curve in Figure 7, aligning with Theorem 4, which states that social welfare reaches its peak and remains invariant for $\mu_1 \in (0.0463, 0.508)$ where $\boldsymbol{\tau}^* > 0$.

We also observe a not entirely smooth plateau on top of the $SW_{probit}(\mu_1)$ curve in Figure 7, a phenomenon not attributed to stochastic variance as we have run 400,000 simulations to precisely estimate the probit function. This irregularity attributes to the inapplicability of Theorem 4 to the probit model, leading to minor the social welfare fluctuations with service rate adjustments. Nonetheless, provided the service rate modifications remain within bounds ensuring $\boldsymbol{\tau}^* > 0$, social welfare hovers near the optimum. This empirical finding underscores the robustness of Theorem 4's insight across diverse utility models.

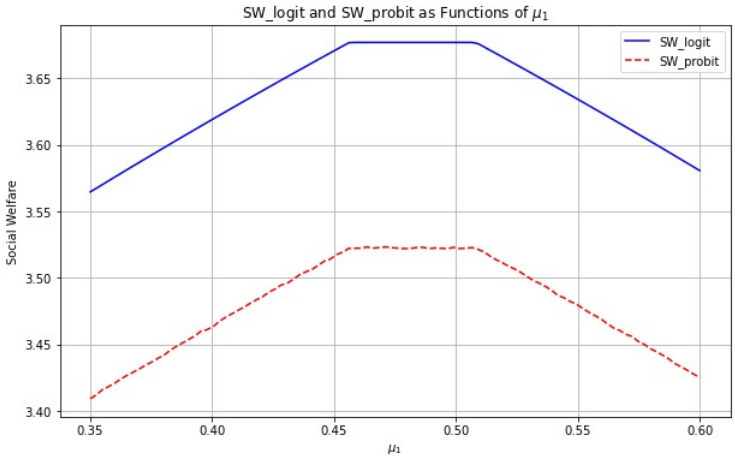
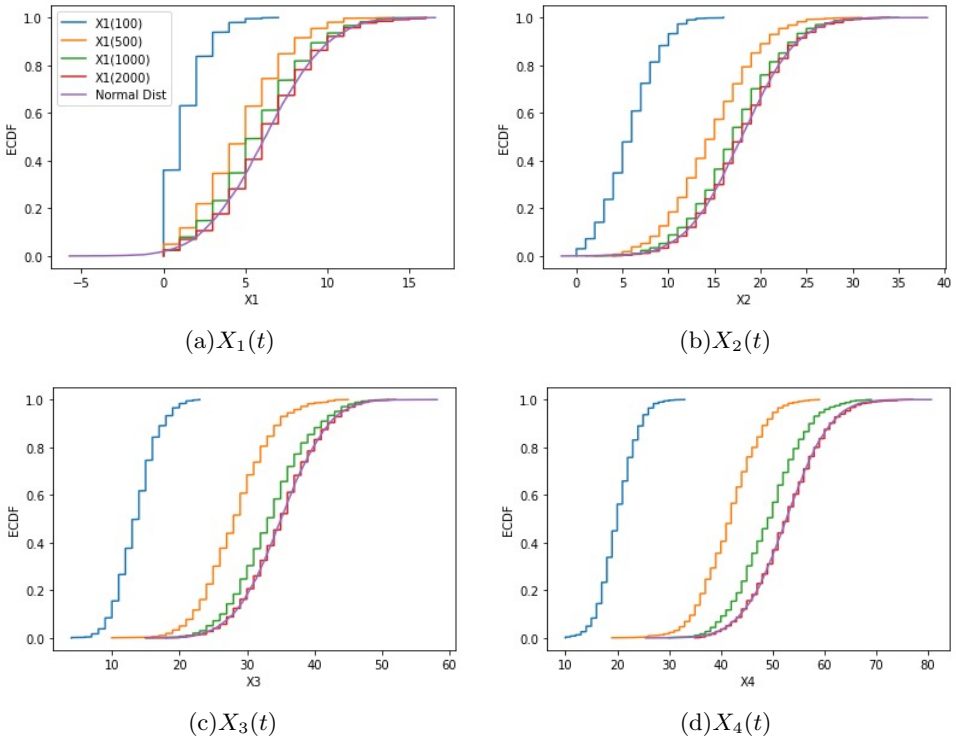


Figure 7 Social Welfare versus Service Rate Adjustment

Finally, we examine the convergence behaviour of the stochastic process towards its diffution limit. We a more complicated DCPQ with $J = 4$ SPs. The service duration at each SP is modeled to follow an independent Gamma distribution, characterized by shape parameters $\{2, 2, 3, 3\}$ and scale parameters $\{2, 2, 2, 2\}$. Customer utility function follows the form (1), with random coefficients $u_{\xi,j}$ for service valuation and c_ξ for waiting cost. We assume that $u_{\xi,j}$ follows a normal distribution with mean and standard deviation both set to $j = 1, 2, 3, 4$, and that c_ξ follows an exponential distribution with a mean 0.05. Each simulation iteration spans $T = 4000$ time units and the entire simulation is replicated 1000 times to construct the probabilistic profiles of the transient transient queue lengths $X_j(100)$, $X_j(500)$, $X_j(1000)$, $X_j(2000)$ for $j = 1, 2, 3, 4$. The fluid analysis shows $\boldsymbol{\tau}^* > 0$. Therefore, Proposition 4 states that the drift limit process should follow an MOU with no reflection barrier.

Figure 8 depicts the cumulative distributions for each $X_j(\cdot)$ and compared them with a normal distribution whose mean and variance is estimated based on the states at $T = 4000$ when the system presumably has reached the steady state. (We choose to estimate them based on simulation as computing the Jacobian is equally challenging due to the random coefficient model). We observe that after $T = 2000$, all queue lengths have reached the stationary distribution. More interestingly, even at $T = 500, 1000$, when the system has not reached the stationary distribution yet, the cumulative distribution is still close to the Gaussian distribution. This observation is in harmony with Theorem 5 which shows the diffusion limit is an MOU whose transient distribution is multivariate Gaussian.



(a)$X_1(t)$ (b)$X_2(t)$

(c)$X_3(t)$ (d)$X_4(t)$

Figure 8 Simulated Queue-Lengths Process in DCPQ

11. Conclusions and Future Research

Our paper is the first to apply heavy traffic approximations to the general DCPQ problem and derive properties of its steady sate. As mentioned, we not only make theoretical contributions but also address some of the managerial issues of interest that arise in practice. For example, our results can directly help practitioners implement system evaluation metrics for controlling these types of stochastic systems. Future work can evaluate the value of information by comparing the social welfare achieved in DCPQ versus a parallel-queue system without waiting time announcements. Another important question would be to evaluate the discrepancy between the social welfare optimization to that of a customer's self-interest maximization in a DCPQ. The results can also be applied to evaluate the performance of DCPQs under different staffing policies, which we were unable to do for the border-crossing queues due to the lack of data on customer balking.

Our analytical framework can be extended to a DCPQ in which all customers renege after an identically and exponentially distributed random period. However, if reneging is endogenous (state-dependent), then the problem is known to be hard (Ata and Peng, 2018). Also, in some situations, waiting customers may abandon the current queue and join a different queue. Usually, when a customer abandons the current queue, she has to lose her priority in that queue and has to wait at the end of the new queue. Such a 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 this will not change the behavior of the DCPQ and thus will not affect the asymptotic characterization significantly. Relaxing some of the technical assumptions, such as Poisson arrival and exponential reneging times, can be an interesting but challenging and is left for future research.

Future studies could delve into the implications of delays in disseminating real-time waiting time information and the effects of inaccuracies in such data on the performance of DCPQ systems. The majority of existing research on this topic focuses on single-queue models (Ibrahim, 2018), indicating a significant research gap for DCPQ environments. This highlights a compelling need for

further exploration within the context of parallel-queue systems, where the dynamics of information dissemination and its reliability may have distinct impacts on system efficiency and social welfare.

Acknowledgments
The authors are grateful to Jim Dai, Armann Ingolfsson, Peter Glynn, Itai Gurvich, Avishai Mandelbaum, Kavita Ramanan, Amy Ward, and Assaf Zeevi for their valuable suggestions on this work. The authors appreciate for the help from Yifeng Cao with the Case Study. The research of Yichuan Ding is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grants (RGPIN 2019-05539) and grants from the Key Program of NSFC-FRQSC Joint Project (NSFC No. 72061127002, FRQSC No. 295837). The Research of Mahesh Nagarajan is partially supported by NSERC RGPIN-2014-03901. The Research of George Zhang is partially supported by NSERC RGPIN-2019-06364.

References

Abouee-Mehrizi, Hossein, Opher Baron. 2016. State-dependent m/g/1 queueing systems. *Queueing Systems* **82**(1-2) 121–148.
Adiri, I., U. Yechiali. 1974. Optimal priority-purchasing and pricing decisions in nonmonopoly and monopoly queues. *Operations Research* **22**(5) pp. 1051–1066. URL `http://www.jstor.org/stable/169658`.
Afèche, Philipp, Barış Ata. 2013. Bayesian dynamic pricing in queueing systems with unknown delay cost characteristics. *Manufacturing & Service Operations Management* **15**(2) 292–304.
Armony, Mor, Constantinos Maglaras. 2004. Contact centers with a call-back option and real-time delay information. *Operations Research* **52**(4) 527–545.
Armony, Mor, Nahum Shimkin, Ward Whitt. 2009. The impact of delay announcements in many-server queues with abandonment. *Operations Research* **57**(1) 66–81.
Arrow, Kenneth J, Henry D Block, Leonid Hurwicz. 1959. On the stability of the competitive equilibrium, ii. *Econometrica: Journal of the Econometric Society* 82–109.
Ata, Baris, Yichuan Ding, Stefanos Zenios. 2021. An achievable-region-based approach for kidney allocation policy design with endogenous patient choice. *Manufacturing & Service Operations Management* **23**(1) 36–54.
Ata, Baris, Xiaoshan Peng. 2018. An equilibrium analysis of a multiclass queue with endogenous abandonments in heavy traffic. *Operations Research* **66**(1) 163–183.
Berry, Arthur, James Levinsohn, Ariel Pakes. 1995. Your use of the jstor archive indicates your acceptance of the terms & conditions of use, available at http://about. jstor. org/terms. *Econometrica* **63**(4) 841–890.
Brémaud, Pierre. 1981. *Point Processes and Queues. Martingale Dynamics., Berlin – Heidelberg – New York 1981, 373 S., 31 Abb., DM 88,–*. Springer.
Brown, Timothy C, M Gopalan Nair. 1988. A simple proof of the multivariate random time change theorem for point processes. *Journal of Applied Probability* 210–214.
Cao, Ping, Shuangchi He, Junfei Huang, Yunan Liu. 2019. To pool or not to pool: Queueing design for large-scale service systems. *working paper* .
Chen, Hong, Murray Frank. 2004. Monopoly pricing when customers queue. *IIE Transactions* **36**(6) 569–581.
Chen, Hong, David D Yao. 2001. *Fundamentals of queueing networks: Performance, asymptotics, and optimization*, vol. 46. Springer.
Choudhury, GL, A Mandelbaum, MI Reiman, W Whitt. 1997. Fluid and diffusion limits for queues in slowly changing environments. *Stochastic Models* **13**(1) 121–146.
Cottle, Richard W. 1966. Nonlinear programs with positively bounded jacobians. *SIAM Journal on Applied Mathematics* **14**(1) pp. 147–158. URL `http://www.jstor.org/stable/2946183`.
Delasay, Mohammad, Armann Ingolfsson, Bora Kolfal. 2016. Modeling load and overwork effects in queueing systems with adaptive service rates. *Operations Research* .

Dieker, Antonius Bernardus, Xuefeng Gao. 2013. Positive recurrence of piecewise ornstein–uhlenbeck processes and common quadratic lyapunov functions. *The Annals of Applied Probability* **23**(4) 1291–1317.
Dong, Jing, Pnina Feldman, Galit B Yom-Tov. 2015. Service systems with slowdowns: Potential failures and proposed solutions. *Operations Research* **63**(2) 305–324.
Dong, Jing, Elad Yom-Tov, Galit B Yom-Tov. 2019. The impact of delay announcements on hospital network coordination and waiting times. *Management Science* **65**(5) 1969–1994.
Dupuis, Paul, Hitoshi Ishii. 1993. Sdes with oblique reflection on nonsmooth domains. *The annals of Probability* 554–580.
Edelson, Noel M, David K Hilderbrand. 1975. Congestion tolls for poisson queuing processes. *Econometrica: Journal of the Econometric Society* 81–92.
Eschenfeldt, Patrick, David Gamarnik. 2018. Join the shortest queue with many servers. the heavy-traffic asymptotics. *Mathematics of Operations Research* **43**(3) 867–886.
Ethier, Stewart N, Thomas G Kurtz. 2009. *Markov processes: characterization and convergence*, vol. 282. John Wiley & Sons.
Frutos, Isabel Parra, Joaquin Aranda Gallego. 1999. Multiproduct monopoly: a queueing approach. *Applied Economics* **31**(5) 565–576.
Gamarnik, David, Assaf Zeevi. 2006. Validity of heavy traffic steady-state approximations in generalized jackson networks. *The Annals of Applied Probability* 56–90.
Garnett, Ofer, Avi Mandelbaum, M Reiman. 2002. Designing a call center with impatient customers. *Manufacturing & Service Operations Management* **4**(3) 208–227.
Grossmann, Christian. 2007. *Numerical treatment of partial differential equations*. Springer.
Guo, Pengfei, Paul Zipkin. 2007. Analysis and comparison of queues with different levels of delay information. *Management Science* **53**(6) 962–970.
Gupta, Varun, Jiheng Zhang. 2014. Approximations and optimal control for state-dependent limited processor sharing queues. *arXiv preprint arXiv:1409.0153* .
Haddad, Jean-Paul, Ravi R Mazumdar. 2012. Heavy traffic approximation for the stationary distribution of stochastic fluid networks. *Queueing Systems* **70**(1) 3–21.
Halfin, Shlomo, Ward Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. *Operations research* **29**(3) 567–588.
Harrison, J Michael, Martin I Reiman. 1981. Reflected brownian motion on an orthant. *The Annals of Probability* 302–308.
Harrison, J Michael, Assaf Zeevi. 2004. Dynamic scheduling of a multiclass queue in the halfin-whitt heavy traffic regime. *Operations Research* **52**(2) 243–257.
Hassin, Refael. 1986a. Consumer information in markets with random product quality: The case of queues and balking. *Econometrica* **54**(5) pp. 1185–1195. URL `http://www.jstor.org/stable/1912327`.
Hassin, Refael. 1986b. Consumer information in markets with random product quality: The case of queues and balking. *Econometrica: Journal of the Econometric Society* 1185–1195.
Hassin, Refael. 2009. Equilibrium customers choice between fcfs and random servers.
Hassin, Refael, Moshe Haviv, Shimon Hassin. 2006. To queue or not to queue: Equilibrium behavior in queueing systems. *International Series in Operations Research & Management Science, Springer (hardcover)*. Elsevier, 1109–1186.
Hassin, Refael, Ricky Roet-Green. 2020. On queue-length information when customers travel to a queue. *Manufacturing & Service Operations Management* **23**(4) 989–1004.
Hu, Ming, Yang Li, Jianfu Wang. 2018. Efficient ignorance: Information heterogeneity in a queue. *Management Science* **64**(6) 2650–2671.
Hua, Zhen, When Chen, George Zhe Zhang. 2014. Two-tier service systems. *working paper* .
Ibrahim, Rouba. 2018. Sharing delay information in service systems: a literature survey. *Queueing Systems* **89**(1) 49–79.

Ibrahim, Rouba, Mor Armony, Achal Bassamboo. 2016. Does the past predict the future? the case of delay announcements in service systems. *Management Science* **63**(6) 1762–1780.
Jacod, Jean, Albert N Shiryaev. 1987. *Limit theorems for stochastic processes*, vol. 288. Springer-Verlag Berlin.
Kang, Weining, Kavita Ramanan. 2014. Characterization of stationary distributions of reflected diffusions. *The Annals of Applied Probability* **24**(4) 1329–1374.
Karamardian, Stepan. 1969. The nonlinear complementarity problem with applications, part 1. *Journal of Optimization Theory and Applications* **4**(2) 87–98.
Larsen, Christian. 1998. Investigating sensitivity and the impact of information on pricing decisions in an m/m/1/∞ queueing model. *International journal of production economics* **56** 365–377.
Lee, Chihoon, Anatolii A Puhalskii. 2015. Non-markovian state-dependent networks in critical loading. *Stochastic Models* **31**(1) 43–66.
Leeman, Wayne A. 1964. The reduction of queues through the use of price. *Operations Research* **12**(5) pp. 783–785. URL `http://www.jstor.org/stable/167784`.
Leite, Saul C, Marcelo D Fragoso. 2008. Diffusion approximation of state dependent g-networks under heavy traffic. *Decision and Control, 2008. CDC 2008. 47th IEEE Conference on*. IEEE, 1495–1500.
Li, Lode, Yew Sing Lee. 1994. Pricing and delivery-time performance in a competitive environment. *Management Science* **40**(5) 633–646.
Littlechild, SC. 1974. Optimal arrival rate in a simple queueing system. *International Journal of Production Research* **12**(3) 391–397.
Lowther, George. 2011. The general theory of semimartingales. *Stochastic Calculus*. URL `https://almostsuremath.com/2011/12/27/compensators-of-counting-processes/`.
Luski, Israel. 1976. On partial equilibrium in a queuing system with two servers. *The Review of Economic Studies* **43**(3) 519–525. URL `http://www.jstor.org/stable/2297230`.
Maglaras, Constantinos, Assaf Zeevi. 2004. Diffusion approximations for a multiclass markovian service system with "guaranteed" and "best-effort" service levels. *Mathematics of Operations Research* **29**(4) 786–813.
Maglaras, Costis, John Yao, Assaf Zeevi. 2016. Optimal price and delay differentiation in queueing systems. *Management Science* .
Mandelbaum, Avi, William A Massey, Martin I Reiman. 1998a. Strong approximations for markovian service networks. *Queueing Systems* **30**(1-2) 149–201.
Mandelbaum, Avi, Gennady Pats, et al. 1998b. State-dependent stochastic networks. part i. approximations and applications with continuous diffusion limits. *The Annals of Applied Probability* **8**(2) 569–646.
McFadden, Daniel, et al. 1973. Conditional logit analysis of qualitative choice behavior .
Megiddo, Nimrod, Masakazu Kojima. 1977. On the existence and uniqueness of solutions in nonlinear complementarity theory. *Mathematical Programming* **12**(1) 110–130.
Mendelson, Haim. 1985a. Pricing computer services: queueing effects. *Communications of the ACM* **28**(3) 312–321.
Mendelson, Haim. 1985b. Pricing computer services: Queueing effects. *Communications of the ACM* **28**(3) 312–321.
Meucci, Attilio. 2009. Review of statistical arbitrage, cointegration, and multivariate ornstein-uhlenbeck. *Cointegration, and Multivariate Ornstein-Uhlenbeck (May 14, 2009)* .
Moré, J, Werner Rheinboldt. 1973. On p-and s-functions and related classes of n-dimensional nonlinear mappings. *Linear Algebra and its Applications* **6** 45–68.
Moré, Jorge J. 1974a. Classes of functions and feasibility conditions in nonlinear complementarity problems. *Mathematical Programming* **6**(1) 327–338.
Moré, Jorge J. 1974b. Coercivity conditions in nonlinear complementarity problems. *Siam Review* **16**(1) 1–16.

Naor, P. 1969. The regulation of queue size by levying tolls. *Econometrica* **37**(1) 15–24. URL http://econpapers.repec.org/RePEc:ecm:emetrp:v:37:y:1969:i:1:p:15-24.
Nevo, Aviv. 2000. A practitioner's guide to estimation of random-coefficients logit models of demand. *Journal of economics & management strategy* **9**(4) 513–548.
Park, Eric, Huiyin Ouyang, Jingqi Wang, Sergei Savin, Siu Chung Leung, Timothy H Rainer. 2023. Patient sensitivity to emergency department waiting time announcements. *Manufacturing & Service Operations Management* .
Pavliotis, Grigorios A. 2014. Stochastic processes and applications. *Texts in Applied Mathematics* **60**.
Pender, Jamol, Richard Rand, Elizabeth Wesson. 2020. A stochastic analysis of queues with customer choice and delayed information. *Mathematics of Operations Research* .
Plemmons, RJ, A Berman. 1979. Nonnegative matrices in the mathematical sciences. *Academic, New York* .
Rajbhandari, Rajat, Juan Villa, Roberto Macias, William Tate, et al. 2012. Measuring border delay and crossing times at the us-mexico border: part ii. guidebook for analysis and dissemination of border crossing time and wait time data. Tech. rep., United States. Federal Highway Administration.
Reed, Josh, Amy R. Ward. 2004. A Diffusion Approximation for a Generalized Jackson Network with Reneging. *Proceedings of the 42nd Annual Allerton Conference on Communication, Control, and Computing, Sept. 29-Oct. 1* .
Reiman, Martin I. 1984. Open queueing networks in heavy traffic. *Mathematics of operations research* **9**(3) 441–458.
Sideris, Thomas C. 2013. *Ordinary differential equations and dynamical systems*, vol. 2. Springer.
Singh, Siddharth Prakash, Mohammad Delasay, Alan Scheller-Wolf. 2023. Real-time delay announcement under competition. *Production and Operations Management* **32**(3) 863–881.
Stidham Jr, Shaler. 1978. Socially and individually optimal control of arrivals to a gi/m/1 queue. *Management Science* **24**(15) 1598–1610.
Su, Xuanming, Stefanos A. Zenios. 2006. Recipient choice can address the efficiency-equity trade-off in kidney transplantation: A mechanism design model. *Manage. Sci.* **52**(11) 1647–1660. doi: http://dx.doi.org/10.1287/mnsc.1060.0541.
Swart, JM. 2002. Pathwise uniqueness for a sde with non-lipschitz coefficients. *Stochastic processes and their applications* **98**(1) 131–149.
Tanaka, Hiroshi. 1979. Stochastic differential equations with reflecting boundary condition in convex regions. *Hiroshima Mathematical Journal* **9**(1) 163–177. URL http://projecteuclid.org/euclid.hmj/1206135203.
Train, Kenneth. 1986. *Qualitative choice analysis: Theory, econometrics, and an application to automobile demand*, vol. 10. MIT press.
Vatiwutipong, Pat, Nattakorn Phewchean. 2019. Alternative way to derive the distribution of the multivariate ornstein–uhlenbeck process. *Advances in Difference Equations* **2019** 1–7.
Walras, Leon. 2013. *Elements of pure economics*. Routledge.
Wang, Jianfu, Ming Hu. 2020. Efficient inaccuracy: User-generated information sharing in a queue. *Management Science* **66**(10) 4648–4666.
Ward, Amy R, Mor Armony. 2013. Blind fair routing in large-scale service systems with heterogeneous customers and servers. *Operations Research* **61**(1) 228–243.
Ward, Amy R., Peter W. Glynn. 2003. A diffusion approximation for a markovian queue with reneging. *Queueing Syst. Theory Appl.* **43**(1-2) 103–128.
WCOG. 2019. Cascade gateway border data warehouse URL http://www.cascadegatewaydata.com.
Weerasinghe, Ananda. 2014. Diffusion approximations for g/m/n+ gi queues with state-dependent service rates. *Mathematics of Operations Research* **39**(1) 207–228.

Whitt, Ward. 2002. *Stochastic-process limits: an introduction to stochastic-process limits and their application to queues*. Springer.
WSDOT. 2024. Border crossing wait times URL `https://wsdot.com/travel/real-time/border-crossings`.
Yamada, Keigo. 1995. Diffusion approximation for open state-dependent queueing networks in the heavy traffic situation. *The Annals of Applied Probability* 958–982.
Yamada, Toshio, Shinzo Watanabe. 1971. On the uniqueness of solutions of stochastic differential equations. *J. Math. Kyoto Univ.* **11**(1) 155–167. doi:10.1215/kjm/1250523691. URL `http://dx.doi.org/10.1215/kjm/1250523691`.
Yu, Mengqiao, Yichuan Ding, Robin Lindsey, Cong Shi. 2016. A data-driven approach to manpower planning at us–canada border crossings. *Transportation Research Part A: Policy and Practice* **91** 34–47.
Zenios, Stefanos A. 1999. Modeling the transplant waiting list: A queueing model with reneging. *Queueing Syst. Theory Appl.* **31**(3-4) 239–251.

Appendix A: Proof of Proposition 1

Proof. We first prove that the Jacobian of the arrival rate function exists and is continuous a.e.

For all $j \neq i$ and $i, j \neq 0$, if the partial derivative $\frac{\partial p_j(\boldsymbol{\tau})}{\partial \tau_i}$ exists, then it must equal to the following limit

$$\lim_{t\to 0} \frac{1}{t}(p_j(\boldsymbol{\tau} + t\boldsymbol{e_i}) - p_j(\boldsymbol{\tau})). \tag{57}$$

Note that $\boldsymbol{\tau} + t\boldsymbol{e_i}$ and $\boldsymbol{\tau}$ differs only in the i^{th} component. Thus, if a customer of type ξ chooses to join queue j at $\boldsymbol{\tau} + t\boldsymbol{e_i}$, but not to join queue j at $\boldsymbol{\tau}$, then he must have chosen queue i at $\boldsymbol{\tau}$. Because his utility of joining other queues is not changed. Those customers must have their parameters $(\boldsymbol{u}_\xi, c_\xi)$ contained in the set $S^1 \cap S^2(t)$, where

$$\begin{aligned} S^1 &:= \left\{ (\boldsymbol{u}, c) \,|\, \begin{array}{l} u_j - c\tau_j > \max\{0, u_k - c\tau_k,\ k \neq i, j\} \\ u_i - c\tau_i > \max\{0, u_k - c\tau_k,\ k \neq i, j\} \end{array} \right\} \\ S^2(t) &:= \left\{ (\boldsymbol{u}, c) |\, c(\tau_i - \tau_j) \leq u_i - u_j < c(\tau_i - \tau_j + t) \right\} \end{aligned} \tag{58}$$

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

If $\tau_i \neq \tau_j$, the limit (57) can be calculated as

$$\begin{aligned} &\lim_{t\to 0} \frac{1}{t}(p_j(\boldsymbol{\tau} + t\boldsymbol{e_i}) - p_j(\boldsymbol{\tau})) \\ =& \lim_{t\to 0} \frac{1}{t} \Pr\left((\boldsymbol{u}, c) \in S(t)\right) \\ =& \lim_{t\to 0} \frac{1}{t} \int_{(\boldsymbol{u},c)\in S^2(t)} I((\boldsymbol{u}, c) \in S_1) f(\boldsymbol{u}, c)\, d\boldsymbol{u} dc \\ =& \lim_{t\to 0} \int [\frac{1}{t} \int_{\frac{u_i-u_j}{\tau_i-\tau_j+t}}^{\frac{u_i-u_j}{\tau_i-\tau_j}} f_{c|\boldsymbol{u}}(c) I((\boldsymbol{u}, c) \in S_1)\, dc] f(\boldsymbol{u})\, d\boldsymbol{u} \\ =& \int \lim_{t\to 0} [\frac{1}{t} \int_{\frac{u_i-u_j}{\tau_i-\tau_j+t}}^{\frac{u_i-u_j}{\tau_i-\tau_j}} f_{c|\boldsymbol{u}}(c) I((\boldsymbol{u}, c) \in S_1)\, dc] f(\boldsymbol{u})\, d\boldsymbol{u} \qquad (59) \\ =& \int I((\boldsymbol{u}, \frac{u_i - u_j}{\tau_i - \tau_j}) \in S_1) f(\boldsymbol{u}, \frac{u_i - u_j}{\tau_i - \tau_j})\, d\boldsymbol{u}. \qquad (60) \end{aligned}$$

Equality (59) is due to dominated convergence. To see that, note that the term inside $[\cdot]$ has the following limit

$$\lim_{t\to 0}[\frac{1}{t}\int_{\frac{u_i-u_j}{\tau_i-\tau_j+t}}^{\frac{u_i-u_j}{\tau_i-\tau_j}} f_{c|\boldsymbol{u}}(c)I((\boldsymbol{u},c)\in S_1)\,dc] = f_{c|\boldsymbol{u}}(\frac{u_i-u_j}{\tau_i-\tau_j})\,I((\boldsymbol{u},\frac{u_i-u_j}{\tau_i-\tau_j})\in S_1). \tag{61}$$

Thus, for sufficiently small t, the term inside $[\cdot]$ is upper bounded by $2f_{c|\boldsymbol{u}}(\frac{u_i-u_j}{\tau_i-\tau_j})I((\boldsymbol{u},\frac{u_i-u_j}{\tau_i-\tau_j})\in S_1)$, whose integral with respect to $\boldsymbol{u}$ is upper bounded by the marginal density $2f_c(\frac{u_i-u_j}{\tau_i-\tau_j})$.

Therefore, if $\tau_i\neq\tau_j$, the partial derivative $\frac{\partial p_j(\boldsymbol{\tau})}{\partial\tau_i}$, as the limit of $\frac{1}{t}(p_j(\boldsymbol{\tau}+t\boldsymbol{e_i})-p_j(\boldsymbol{\tau}))$, exists and has the following expression,

$$\frac{\partial p_j(\boldsymbol{\tau})}{\partial\tau_i} = \int I((\boldsymbol{u},\frac{u_i-u_j}{\tau_i-\tau_j})\in S_1)f(\boldsymbol{u},\frac{u_i-u_j}{\tau_i-\tau_j})\,d\boldsymbol{u}. \tag{62}$$

Since the RHS of above equation is a continuous function of τ_i and τ_j when $\tau_i\neq\tau_j$, the partial derivative $\frac{\partial p_j(\boldsymbol{\tau})}{\partial\tau_i}$ must be continuous across all $\boldsymbol{\tau}$, except at a zero-measured set of points with $\tau_i=\tau_j$.

The above argument proves that if $j\neq i$, then $\frac{\partial p_j(\boldsymbol{\tau})}{\partial\tau_i}$ exists and is continuous for all $\boldsymbol{\tau}\notin\mathcal{K}^J$. It remains to prove the above property of $\frac{\partial p_j(\boldsymbol{\tau})}{\partial\tau_i}$ for the $j=i$ case. Because $\sum_{i=0}^J p_j(\boldsymbol{\tau})\equiv 1$, we know that

$$\tfrac{\partial p_i(\boldsymbol{\tau})}{\partial\tau_i} = -\textstyle\sum_{j\neq i,\,j=0,1,\ldots,J}\tfrac{\partial p_j(\boldsymbol{\tau})}{\partial\tau_i} \tag{63}$$

Note that the summation at the RHS consists of $\frac{\partial p_j(\boldsymbol{\tau})}{\partial\tau_i}$ for all $j\neq i$ (including $j=0$). $p_0(\boldsymbol{\tau})$ 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, using the previous argument for the $i\neq j$ case, we can show that $\frac{\partial p_0(\boldsymbol{\tau})}{\partial\tau_i}$ exists and is continuous a.e. Because for all $j\neq i$ (including $j=0$), $\frac{\partial p_j(\boldsymbol{\tau})}{\partial\tau_i}$ exists and is continuous a.e., Equation (63) implies that $\frac{\partial p_i(\boldsymbol{\tau})}{\partial\tau_i}$ exists and is continuous a.e..

So far, we have proved that the arrival rate function $p_j(\boldsymbol{\tau})$ has continuous derivatives except at a zero-measured set with $\tau_j=\tau_k$. Next we show that even at points with $\tau_j=\tau_k$, $p_j(\boldsymbol{\tau})$ remains continuous, though it may not have finite derivatives. Thus, $p_j(\boldsymbol{\tau})$ is absolute continuous. Formally,

$$\begin{aligned}
&\lim_{t\to 0} p_j(\boldsymbol{\tau}+t\boldsymbol{e_i})-p_j(\boldsymbol{\tau}) \;=\; \lim_{t\to 0}\Pr\left((\boldsymbol{u},c)\in S(t)\right)\\
&= \lim_{t\to 0}\int\int[\int_0^{ct} f_{u_i|\boldsymbol{u}_{-i},c}(u_j+x)I((\boldsymbol{u},c)\in S_1)\,du_i]f_{\boldsymbol{u}_{-i},c}(\boldsymbol{u}_{-i},c)\,dc\,d\boldsymbol{u}_{-i} &(64)\\
&= 0 &(65)
\end{aligned}$$

where equality (64) follows from Equality (65) follows from that $\lim_{t\to 0}\int_0^{ct} f_{u_i|\boldsymbol{u}_{-i},c}(u_j+x)I((\boldsymbol{u},c)\in S_1)\,du_i=0$. In the case that c has a discrete distribution, one can replace the integral $\int\cdot dc$ in the above equation with summation $\sum_k\cdot\Pr(c=k)$ without changing the result.

We next prove (CD-a)-(CD-c).

(CD-a): Suppose $\tau_k^2>\tau_k^1$, and $\tau_l^2=\tau_l^1$ for $j\neq k$. For a customer indexed by ξ, if his choice is queue $j\neq k$, then

$$u_k-c\tau_k^2<u_k-c\tau_k^1\leq u_j-c\tau_j^1=u_j-c\tau_j^2, \tag{66}$$

where the first inequality is due to $\tau_k^2>\tau_k^1$, 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 since $\tau_j^1=\tau_j^2$. Therefore, if a customer's initial choice is queue j, then his choice remains the same when the waiting-time vector is changed from $\boldsymbol{\tau}^1$ to $\boldsymbol{\tau}^2$. We thus deduce that $p_j(\boldsymbol{\tau})$ is non-decreasing in τ_k.

(CD-b): Note that $p_j(\boldsymbol{\tau})$ must be non-increasing with τ_j as a result of (CD-a) and $\sum_{k=0}^{J} p_k = 1$. So it suffices to prove $p_j(\boldsymbol{\tau})$ is strictly decreasing when $\boldsymbol{\tau}^1$ has been replaced by $\boldsymbol{\tau}^2$, where $\tau_j^2 > \tau_j^1$ but $\tau_k^2 = \tau_k^1$ for $k \neq j$. A customer ξ will choose to join queue j given expected waiting-times vector $\boldsymbol{\tau}^1$, but not join queue j when the waiting-time vector is changed to $\boldsymbol{\tau}^2$, if and only if

$$(\boldsymbol{u}_\xi, c_\xi) \in \{(\boldsymbol{u}, c) | \begin{matrix} u_j - c\tau_j^1 > \max\{0, u_k - c\tau_k^1,\ k \neq j\} \\ u_j - c\tau_j^2 < \max\{0, u_k - c\tau_k^2,\ k \neq j\} \end{matrix} \} \tag{67}$$

Because the parameter c has positive conditional pdf $f_{c|\boldsymbol{u}}$ 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 τ_j^1 to τ_j^2. Therefore, $p_j(\boldsymbol{\tau})$ is strictly decreasing in τ_j.

(CD-c): Given $\boldsymbol{\tau}^2 := \boldsymbol{\tau}^1 + t\boldsymbol{e}$, the linear form of $U_{\xi,j}$ implies that if $U_{\xi,j} \geq U_{\xi,k}$ for all $k \neq j$ (including $k = 0$) at $\boldsymbol{\tau}^2$, then the same inequalities must hold at $\boldsymbol{\tau}^1$. Therefore, we deduce that $p_j(\boldsymbol{\tau}^1) \geq p_j(\boldsymbol{\tau}^2)$ for all $j \neq 0$. To prove the strict inequality in (10), we notice that a customer of type ξ joins some queue at $\boldsymbol{\tau}^1$, but balks at $\boldsymbol{\tau}^2$ if

$$(\boldsymbol{u}, c) \in \left\{ (\boldsymbol{u}, c) \,|\, \begin{matrix} 0 < \max\{u_k - c\tau_k^1,\ k = 1, \ldots, J\} \\ 0 > \max\{u_k - c(\tau_k^2 + t),\ k = 1, \ldots, J\} \end{matrix} \right\}. \tag{68}$$

Because the parameter c has positive conditional pdf $f_{c|\boldsymbol{u}}$ over $\mathbb{R}_+$, the above set must have a positive probability mass, so the strict inequality (10) is proved, which implies row strict diagonally dominance of the Jacobian matrix. Inequality (11) and the column strict diagonally dominance follow from symmetry of the Jacobian matrix, a result that will be proved in the end of this proof.

We next prove the stability condition (7). A customer will joint queue j only if $u_j - c\tau > 0$. Therefore, when $\tau_j \to \infty$,

$$p_j(\boldsymbol{\tau}) \leq \Pr(u_j - c\tau > 0) = \int [\int_0^{u_j/\tau_j} f_{c|\boldsymbol{u}}(c)\, dc]\, f_{\boldsymbol{u}}(\boldsymbol{u}) d\boldsymbol{u} \to 0. \tag{69}$$

where the convergence follows from $u_j/\tau_j \to 0$ and our assumption that $f_{c|\boldsymbol{u}}(\cdot)$ is bounded. Equation (69) leads to (7).

Finally, we prove that the Jacobian matrix is symmetric whenever it exists. Equation (59) implies that

$$\begin{aligned} \tfrac{\partial p_j(\boldsymbol{\tau})}{\partial \tau_i} &= \lim_{t\to 0} \tfrac{1}{t}(p_j(\boldsymbol{\tau} + t\boldsymbol{e_i}) - p_j(\boldsymbol{\tau})) \\ &= \lim_{t\to 0} \tfrac{1}{t} \Pr\left(\left\{ (\boldsymbol{u}, c) | \begin{matrix} u_j - c\tau_j > \max\{0, u_k - c\tau_k,\ k \neq i, j\} \\ u_i - c\tau_i > \max\{0, u_k - c\tau_k,\ k \neq i, j\} \\ u_j - c\tau_j > u_i - c(\tau_i + t) \\ u_j - c\tau_j < u_i - c\tau_i \end{matrix} \right\} \right) \end{aligned} \tag{70}$$

Similarly,

$$\begin{aligned} \tfrac{\partial p_i(\boldsymbol{\tau})}{\partial \tau_j} &= \lim_{t\to 0} \tfrac{1}{t}(p_i(\boldsymbol{\tau}) - p_i(\boldsymbol{\tau} - t\boldsymbol{e_j})) \\ &= \lim_{t\to 0} \tfrac{1}{t} \Pr\left(\left\{ (\boldsymbol{\alpha}, c, \boldsymbol{\epsilon}) | \begin{matrix} u_j - c\tau_j > \max\{0, u_k - c\tau_k,\ k \neq i, j\} \\ u_i - c\tau_i > \max\{0, u_k - c\tau_k,\ k \neq i, j\} \\ u_j - c(\tau_j - t) > u_i - c\tau_i \\ u_j - c\tau_j < u_i - c\tau_i \end{matrix} \right\} \right) \end{aligned} \tag{71}$$

Notice that the set at the RHS of Equation (70) and (71) are identical. The intuition is that it is the same group of customers who will switch to queue j, when either τ_j has been decreased by t, or τ_j has been increased by t. We thus have $\partial p_j(\boldsymbol{\tau})/\partial \tau_i = \partial p_i(\boldsymbol{\tau})/\partial \tau_j$ and symmetry is proved. ■

The above proof also leads to an example that the arrival rate function $\boldsymbol{\Lambda}(\cdot)$ does not have to be (even locally) Lipschitz continuous. In particular, its partial derivative may be infinite at some point $\boldsymbol{\tau}$. Let $\boldsymbol{u}_{-i}$ denote the vector obtained by removing the i^{th} entry from $\boldsymbol{u}$. Equation (64) then implies that

$$\begin{aligned}
&\liminf_{t\to 0}\frac{1}{t}(p_j(\boldsymbol{\tau}+t\boldsymbol{e_i})-p_j(\boldsymbol{\tau}))\\
&=\liminf_{t\to 0}\int\int[\frac{1}{t}\int_{u_j}^{u_j+ct} f_{u_i|\boldsymbol{u}_{-i},c}(u_i)I((\boldsymbol{u},c)\in S_1)\,du_i]f_{\boldsymbol{u}_{-i},c}(\boldsymbol{u}_{-i},c)\,dcd\boldsymbol{u}_{-i} \qquad (72)\\
&\geq \iint\liminf_{t\to 0}[\frac{1}{t}\int_0^{ct} f_{u_i|\boldsymbol{u}_{-i},c}(u_j+x)I((\boldsymbol{u},c)\in S_1)\,dx]f_{\boldsymbol{u}_{-i},c}(\boldsymbol{u}_{-i},c)\,dcd\boldsymbol{u}_{-i} \qquad (73)\\
&=\iint cf_{u_i|\boldsymbol{u}_{-i},c}(u_j)I((\boldsymbol{u},c)\in S_1)f_{\boldsymbol{u}_{-i},c}(\boldsymbol{u}_{-i},c)\,d\boldsymbol{u}_{-i}\,dc\\
&=\int c[\int f_{\boldsymbol{u}|c}(u_j,\boldsymbol{u}_{-i})I((\boldsymbol{u},c)\in S_1)\,d\boldsymbol{u}_{-i}]\,f_c(c)\,dc
\end{aligned}$$

where inequality (73) follows from the Fatou's Lemma. Note that the integral inside $[\cdot]$ can be infinitely large because $\int f_{\boldsymbol{u}|c}(u_j,\boldsymbol{u}_{-i})\,d\boldsymbol{u}_{-i}=f_{u_i|c}(u_j)$ can be infinitely large when $u_i=u_j$; while we can always properly select the parameters such that the constraint $I((\boldsymbol{u},c)\in S_1)$ is satisfied by $\boldsymbol{u}_{-i}$s in a positive-measured set. Consequently, the partial derivative $\frac{\partial p_j(\boldsymbol{\tau})}{\partial \tau_i}$ can be infinitely large (i.e., not exist) at $\boldsymbol{\tau}$, and can be unbounded near those points. That means, the arrival rate function $\boldsymbol{\Lambda}(\cdot)$ does not have to be (even locally) Lipschitz continuous.

Appendix B: Proof of Lemma 1

Given $\boldsymbol{\tau}(t-)\in\mathbb{R}_+^J$, define the following partition over the domain of $(\boldsymbol{u},c)$ (i.e., $\mathbb{R}_+^{J+1}$):

$$\begin{aligned}
&\pi_0(\boldsymbol{\tau}(t-)) := \{(\boldsymbol{u},c)\in\mathbb{R}_+^{J+1}\,|\,0>u_k-c\tau_k(t-) \text{ for all } k=1,\ldots,J\},\\
&\pi_j(\boldsymbol{\tau}(t-)) := \{(\boldsymbol{u},c)\in\mathbb{R}_+^{J+1}\,|\,u_j-c\tau_j(t-)>\max\{0,u_k-c\tau_k(t-),\ k\neq j\}\}.
\end{aligned} \qquad (74)$$

According to the above definition, a customer, by observing waiting-time estimates $\boldsymbol{\tau}(t-)$, will join queue $j(=0,1,\ldots,J)$ if his parameter vector $(\boldsymbol{u},c)\in\pi_j(\boldsymbol{\tau}(t-))$. Since a tie happens with probability zero, the probability for a customer to join queue j is given by

$$p_j(\boldsymbol{\tau}(t))=\int_{(\boldsymbol{u},c)\in\mathbb{R}_+^{J+1}} 1((\boldsymbol{u},c)\in\pi_j(\boldsymbol{\tau}(t-)))f(\boldsymbol{u},c)\,d\boldsymbol{u}dc. \qquad (75)$$

Let $A_j(t)$ denote the cumulative number of arrivals at queue j by time t. Let $(\boldsymbol{u}^k,c^k)$ denote the parameters of the We have

$$A_j(t)=\int_0^t 1\{(\boldsymbol{u}^{N(s)},c^{N(s)})\in\pi_j(\boldsymbol{\tau}(s-))\}\,dN(s), \qquad (76)$$

where $N(\cdot)$ denotes a standard rate-one Poisson process. Thus, $(\boldsymbol{u}^{N(s)},c^{N(s)})$ denote the parameters of the customer who arrive at time s. Let $\hat{A}_j(t):=\int_0^t p_j(\boldsymbol{\tau}(s-))\,ds$ denote the mean of $A(t)$. Let $\mathcal{H}$ denote the σ-field of the common probabilistic space where all the random events are defined. We then define a filtration for the arrival process as

$$\mathcal{F}(t):=\sigma(N(s),\ 0\leq s\leq t)\vee\sigma((\boldsymbol{u}^{\ell\cap N(t)},c^{\ell\cap N(t)}),\ell=0,1,\ldots)\vee\sigma(\mathcal{N}^0). \qquad (77)$$

where $\sigma(\cdot)$ denotes the sigma-field generated by the random variables inside $(\cdot)$, and $\mathcal{N}^0$ consists of all null sets in $\mathcal{H}$. We define stochastic processes $\boldsymbol{M}^1:=(M_j^1)$ and $\boldsymbol{M}^2:=(M_j^2)$ as follows,

$$\begin{aligned}
M_j^1(t) &:= A_j(t)-\textstyle\int_0^t p_j(\boldsymbol{\tau}(s-))\,dN(s)\\
&= \textstyle\int_0^t(1\{(\boldsymbol{u}^{N(s)},c^{N(s)})\in\pi_j(\boldsymbol{\tau}(s-)\}-p_j(\boldsymbol{\tau}(s-)))\,dN(s),\\
M_j^2(t) &:= \textstyle\int_0^t p_j(\boldsymbol{\tau}(s-))\,dN(s)-\hat{A}_j(t).
\end{aligned} \qquad (78)$$

We next show that M_j^1 and M_j^2 are both $\mathcal{F}(t)$-martingales. For any $t > t_0 \geq 0$, the following identify holds due to Equation (75),

$$\begin{aligned} &\mathbb{E}[M_j^1(t)|\mathcal{F}(t_0)] \\ &= M_j^1(t_0) + \mathbb{E}\left[\int_{t_0}^t [1\{(\boldsymbol{u}^{N(s)}, c^{N(s)}) \in \pi_j(\boldsymbol{\tau}(s-))\} - p_j(\boldsymbol{\tau}(s-))]\, dN(s) \,|\, \mathcal{F}(t_0)\right] \\ &= M_j^1(t_0) + \textstyle\sum_{\ell=1}^{\infty} \mathbb{E}\left[\left(1\{(\boldsymbol{u}^{N(t_0)+\ell}, c^{N(t_0)+\ell}) \in \pi_j(\boldsymbol{\tau}(t_\ell-))\} - p_j(\boldsymbol{\tau}(t_\ell-))\right) 1\{t_\ell \leq t\} \,|\, \mathcal{F}(t_0)\right] \\ &= M_j^1(t_0). \end{aligned} \tag{79}$$

where $t_\ell := N^{-1}(N(t_0)+\ell)$ denotes the arrival time of the $(N(t_0)+\ell)^{th}$ customer. The last equality follows that the random variables $(\boldsymbol{u}^{N(t_0)+\ell}, c^{N(t_0)+\ell})$ $(\ell = 1, 2, \ldots)$ are independent of $\mathcal{F}(t_0)$, t_ℓ, and $\boldsymbol{\tau}(t_\ell-)$. Thus, $\boldsymbol{M}^1$ is an $\mathcal{F}(t)$-martingale.

For $\boldsymbol{M^2}$, since $N(t)$ is a Poisson process, $N(t)-t$ is an $\mathcal{F}(t)$-martingale. Moreover, since $p_j(\boldsymbol{\tau}(t-))$ is left-continuous, and thus is an $\mathcal{F}(t)$-predictable process with respect to $\mathcal{F}(t)$. We then invoke the integration theorem part (β) (T8 Page 27, Brémaud (1981)), in which $X_s = p_j(\boldsymbol{\tau}(s-))$, $\lambda_u \equiv 1$, and $M_s = N(s) - s$ in the theorem. It then implies that $M_j^2(t) := \int_0^t p_j(\boldsymbol{\tau}(s-))\, dN(s) - \int_0^t p_j(\boldsymbol{\tau}(s-))\, ds$ is an $\mathcal{F}(t)$-martingale for each $j = 1, \ldots, J$. Therefore, both $\boldsymbol{M}^1$ and $\boldsymbol{M}^2$ are vector-valued $\mathcal{F}(t)$-martingale, and so is $\boldsymbol{A} - \hat{\boldsymbol{A}} = \boldsymbol{M}^1 + \boldsymbol{M}^2$.

Since $\boldsymbol{A} - \hat{\boldsymbol{A}}$ is an $\boldsymbol{F}(t)$-martingale, it must be also an $\boldsymbol{F}(t)$-local martingale. Furthermore, $\hat{\boldsymbol{A}}(0) = \boldsymbol{0}$. Thus, $\hat{\boldsymbol{A}}$ satisfies the definition as being a compensator of the counting process $\boldsymbol{A}(\cdot)$, i.e., the unique right-continuous and increasing process with $\hat{\boldsymbol{A}}(0) = \boldsymbol{0}$ such that $\boldsymbol{A} - \hat{\boldsymbol{A}}$ is a local martingale (Lowther, 2011). Furthermore, $\hat{\boldsymbol{A}}$ is a continuous compensator of $\boldsymbol{A}$ because for each $j = 1, 2, \ldots, J$, $\hat{A}_j(t) = \int_0^t p_j(\boldsymbol{\tau}(s-))\, ds$ has continuous paths (Brown and Nair, 1988). We also know that with probability 1, $\boldsymbol{A}(\cdot)$ does not have simultaneous jumps. We can then invoke Meyer's theorem (Brown and Nair, 1988) and deduce that $A_j(\hat{A}_j^{-1}(t))$, $j = 1, \ldots, J$ are independent rate-one Poisson processes, i.e.,

$$A_j(\hat{A}_j^{-1}(\cdot)) \stackrel{d}{=} N_j(\cdot), \tag{80}$$

where each $N_j(\cdot)$ $(j = 1, 2, \ldots, J)$ is an independent rate-one standard Poisson process. Note that the inverse function $\hat{A}_j^{-1}(\cdot)$ is well defined since $\hat{A}_j(\cdot)$ is strictly and continuously increasing. Consequently, for $0 < t_1 < \ldots \leq t_m$, we define $z_k = \hat{A}_j(t_k) = \int_0^{t_k} p_j(\boldsymbol{\tau}(s-))\, ds$ for $k = 1, 2, \ldots, m$. Then for all Borel sets $B_1, B_2, \ldots, B_m$, we have

$$\begin{aligned} &\Pr(A_j(t_1) \in B_1, A_j(t_2) \in B_2, \ldots, A_j(t_m) \in B_m) \\ &= \Pr(A_j(\hat{A}_j^{-1}(z_1)) \in B_1, A_j(\hat{A}_j^{-1}(z_2)) \in B_2, \ldots, A_j(\hat{A}_j^{-1}(z_m)) \in B_m) \\ &= \Pr(N_1(z_1) \in B_1, N_2(z_2) \in B_2, \ldots, N_m(z_m) \in B_m) \\ &= \Pr(N_1(\textstyle\int_0^{t_1} p_j(\boldsymbol{\tau}(s-))\, ds) \in B_1, N_2(\int_0^{t_2} p_j(\boldsymbol{\tau}(s-))\, ds) \in B_2, \\ &\quad \ldots, N_m(\textstyle\int_0^{t_m} p_j(\boldsymbol{\tau}(s-))\, ds) \in B_m) \end{aligned} \tag{81}$$

where the second equality follows from (80) (finite dimensional distribution equivalence). The above equality therefore proves the equivalence between $A_j(\cdot)$ and $N_j(\int_0^{\cdot} p_j(\boldsymbol{\tau}(s-))\, ds)$ with respect to finite dimensional distribution.

Appendix C: Proof of Proposition 2

Proof. If $\boldsymbol{\tau}^*$ 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 $p_j(\boldsymbol{\tau}^*)$. For queues with excessive service capacity, we must have $\mu_j - p_j(\boldsymbol{\tau}^*) > 0$, and that queue must be empty so $\tau_j^* = 0$; for other queues, we have $\mu_j - p_j(\boldsymbol{\tau}) = 0$. We thus proved the complementary slackness condition in (18). The other inequality constraints can be proved straightforwardly.

Suppose $\boldsymbol{\tau}^*$ is a solution to (18). For queues with $\tau_j^* > 0$, by the complementary slackness condition in (18), we have $\mu_j - p_j(\boldsymbol{\tau}) = 0$, 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 $\mu_j - p_j(\boldsymbol{\tau}) \geq 0$. Since those queues are empty, the arrival and departure rates must be balanced. Thus, the drift coefficient in equation (16) must equal to zero at $\boldsymbol{\tau}^*$, which implies $\boldsymbol{\tau}(t) \equiv \boldsymbol{\tau}^*$ provided that $\boldsymbol{\tau}(t)$ is a solution to (16) with $\boldsymbol{\tau}(0) = \boldsymbol{\tau}^*$. ■

Appendix D: Proof of Lemma 2

Proof. Suppose $\boldsymbol{x}$ and $\boldsymbol{y}$ are both solutions to SDER (17). Then by the first equation in the proof of Theorem 4.1 in (Tanaka (1979), page 175), we have

$$\begin{aligned} &\|\boldsymbol{x}(t) - \boldsymbol{y}(t)\|^2 \\ &\leq \|\int_0^t (\boldsymbol{\sigma}(s, \boldsymbol{x}(s)) - \boldsymbol{\sigma}(s, \boldsymbol{y}(s)))d\boldsymbol{B}(s)\|^2 + 2\int_0^t \langle \boldsymbol{x}(s) - \boldsymbol{y}(s), \boldsymbol{b}(s, \boldsymbol{x}(s)) - \boldsymbol{b}(s, \boldsymbol{y}(s))\rangle ds + \text{ the remainder.} \end{aligned} \tag{82}$$

where the remainder has zero expectation. We thus have

$$\begin{aligned} &\mathbb{E}\|\boldsymbol{x}(t) - \boldsymbol{y}(t)\|^2 \\ &\leq \mathbb{E}\int_0^t \|\boldsymbol{\sigma}(s, \boldsymbol{x}(s)) - \boldsymbol{\sigma}(s, \boldsymbol{y}(s))\|^2 ds + 2\int_0^t \langle \boldsymbol{x}(s) - \boldsymbol{y}(s),\ \boldsymbol{b}(s, \boldsymbol{x}(s)) - \boldsymbol{b}(s, \boldsymbol{y}(s))\rangle\, ds \\ &\leq K^2 \mathbb{E}\int_0^t \|\boldsymbol{x}(s) - \boldsymbol{y}(s)\|^2 ds + 2\int_0^t \langle \boldsymbol{x}(s) - \boldsymbol{y}(s),\ \boldsymbol{b}(s, \boldsymbol{x}(s)) - \boldsymbol{b}(s, \boldsymbol{y}(s))\rangle\, ds \end{aligned} \tag{83}$$

where the inequality follows from Lipschitz continuity of $\boldsymbol{\sigma}(s, \cdot)$. By absolute continuity of $\boldsymbol{b}(s, \cdot)$, we have

$$\boldsymbol{b}(s, \boldsymbol{x}(s)) - \boldsymbol{b}(s, \boldsymbol{y}(s)) = \int_0^1 \boldsymbol{R}(\boldsymbol{y}(s) + \xi(\boldsymbol{x}(s) - \boldsymbol{y}(s)))(\boldsymbol{x}(s) - \boldsymbol{y}(s))\, d\xi, \tag{84}$$

with the Jacobian matrix $\boldsymbol{R}(\boldsymbol{y}(s) + \xi(\boldsymbol{x}(s) - \boldsymbol{y}(s)))$ negative definite for almost all $\xi \in [0, 1]$. Consequently,

$$\begin{aligned} \langle \boldsymbol{x}(s) - \boldsymbol{y}(s), \boldsymbol{b}(s, \boldsymbol{x}(s)) - \boldsymbol{b}(s, \boldsymbol{y}(s))\rangle &= \langle \boldsymbol{x}(s) - \boldsymbol{y}(s),\ \int_0^1 \boldsymbol{R}(\boldsymbol{y}(s) + \xi(\boldsymbol{x}(s) - \boldsymbol{y}(s)))(\boldsymbol{x}(s) - \boldsymbol{y}(s))\, d\xi\rangle \\ &= \int_0^1 \langle \boldsymbol{x}(s) - \boldsymbol{y}(s),\ \boldsymbol{R}(\boldsymbol{y}(s) + \xi(\boldsymbol{x}(s) - \boldsymbol{y}(s)))(\boldsymbol{x}(s) - \boldsymbol{y}(s))\rangle\, d\xi \\ &\leq 0 \end{aligned} \tag{85}$$

which, together with Equation (83), leads to

$$\mathbb{E}\|\boldsymbol{x}(t) - \boldsymbol{y}(t)\|^2 \leq K^2 \int_0^t \mathbb{E}\|\boldsymbol{x}(s) - \boldsymbol{y}(s)\|^2 ds. \tag{86}$$

Then by the Gronwall's inequality (e.g., Ethier and Kurtz (2009), page 498), we have $\|\boldsymbol{x}(t) - \boldsymbol{y}(t)\| = 0$. ■

Appendix E: Proof of Theorem 1

Proof. By Lemma 1, the length of queue j is described by the following equation,

$$\begin{aligned} x_j^n(t) &= x_j^n(0) + \tfrac{1}{n} N(\textstyle\int_0^t np_j(\boldsymbol{\tau}^n(s))ds) - \tfrac{1}{n} S_j^n(W_j^n(t)) \\ &= x_j^n(0) + \tfrac{1}{n} Z_j^n(t) + \textstyle\int_0^t (p_j(\boldsymbol{x}^n(s) \circ (\boldsymbol{\mu}^n)^{-1}) - p_j(\boldsymbol{x}^n(s) \circ \boldsymbol{\mu}^{-1}))ds \\ &\quad + \textstyle\int_0^t (p_j(\boldsymbol{x}^n(s) \circ \boldsymbol{\mu}^{-1}) - \mu_j^n)ds + \ell_j^n(t) \end{aligned} \tag{87}$$

where $x_j^n(t)$ was defined in (14), $\ell_j^n(t) := \mu_j^n(t - W_j^n(t))$ is the minimal non-decreasing process which ensures $x_j^n(t) \geq 0$, and

$$\begin{aligned} Z_j^n(t) := &\left(N(\textstyle\int_0^t np_j(\boldsymbol{X}^n(s) \circ (n\boldsymbol{\mu}^n)^{-1})ds) - \textstyle\int_0^t np_j(\boldsymbol{X}^n(s) \circ (n\boldsymbol{\mu}^n)^{-1})ds\right) \\ &+ \left(n\mu_j^n W_j^n(t) - S_j^n(W_j^n(t))\right) \end{aligned} \tag{88}$$

represents a mean-zero centered process. We also define $\boldsymbol{\Gamma}(\boldsymbol{x}) := \boldsymbol{\Lambda}(\boldsymbol{x} \circ \boldsymbol{\mu}^{-1})$ and

$$\tilde{\boldsymbol{z}}^n(t) := \tfrac{1}{n}\boldsymbol{Z}^n(t) + \int_0^t (\boldsymbol{\Lambda}(\boldsymbol{x}^n(s) \circ (\boldsymbol{\mu}^n)^{-1}) - \boldsymbol{\Lambda}(\boldsymbol{x}^n(s) \circ (\boldsymbol{\mu})^{-1}))ds. \tag{89}$$

Then we can express $\boldsymbol{x}(t)$ and $\boldsymbol{x}^n(t)$ as

$$\begin{aligned} \boldsymbol{x}^n(t) &= \boldsymbol{x}^n(0) + \int_0^t \boldsymbol{\Gamma}(\boldsymbol{x}^n(s))ds - t\boldsymbol{\mu}^n + \tilde{\boldsymbol{z}}^n(t) + \boldsymbol{\ell}^n(t), \\ \boldsymbol{x}(t) &= \boldsymbol{x}(0) + \int_0^t \boldsymbol{\Gamma}(\boldsymbol{x}(s))ds - t\boldsymbol{\mu} + \boldsymbol{\ell}(t). \end{aligned} \tag{90}$$

where $\boldsymbol{\ell}(\cdot) := (\ell_j(\cdot))_{j=1,\ldots,J}$ and $\boldsymbol{\ell}^n(\cdot)$ denote the minimal non-decreasing processes that keep $\boldsymbol{x}(t)$ and $\boldsymbol{x}^n(t)$ staying non-negative.

We invoke the first inequality in Remark 2.2 of (Tanaka, 1979), in which we plug in the following quantity $\xi(t) := \boldsymbol{x}(t)$, $\tilde{\xi}(t) := \boldsymbol{x}^n(t)$, $w(t) := \boldsymbol{x}(0) - t\boldsymbol{\mu}$ and $\tilde{w}(t) := \boldsymbol{x}^n(0) + \tilde{\boldsymbol{z}}^n(t) - t\boldsymbol{\mu}^n$, $a(t) = \boldsymbol{\Gamma}(\boldsymbol{x}(t))$ and $\tilde{a}(t) = \boldsymbol{\Gamma}(\boldsymbol{x}^n(t))$. Since $\boldsymbol{\Gamma}(\cdot)$ is absolutely continuous, $a(\cdot)$ and $\tilde{a}(\cdot)$ are both right continuous and have bounded variation, which satisfy the conditions specified in (Tanaka, 1979). The first inequality in Remark 2.2 of (Tanaka, 1979) then leads to following inequality,

$$\begin{aligned} &\|\boldsymbol{x}^n(t) - \boldsymbol{x}(t)\|^2 \\ \le\ &\|\boldsymbol{x}^n(0) - \boldsymbol{x}(0) + \tilde{\boldsymbol{z}}^n(t) - t(\boldsymbol{\mu}^n - \boldsymbol{\mu})\|^2 + 2\int_0^t \langle \boldsymbol{x}^n(s) - \boldsymbol{x}(s), \boldsymbol{\Gamma}(\boldsymbol{x}^n(s)) - \boldsymbol{\Gamma}(\boldsymbol{x}(s))\rangle ds \\ &+ \int_0^t \langle \tilde{\boldsymbol{z}}^n(t) - \tilde{\boldsymbol{z}}^n(s) - (\boldsymbol{\mu}^n - \boldsymbol{\mu})(t-s), d\tilde{a}(s) - da(s) + d\tilde{\boldsymbol{\ell}}(s) - d\boldsymbol{\ell}(s)\rangle ds \end{aligned} \tag{91}$$

Later, we will prove that $\|\tilde{\boldsymbol{z}}^n\|_T \to 0$ for all $T > 0$. Since $\|\boldsymbol{x}^n(0) - \boldsymbol{x}(0)\| \to 0$ and $\|\boldsymbol{\mu}^n - \boldsymbol{\mu}\| \to 0$, the first and the third terms on the right-hand-side of Equation (91) both converge to zero. The second term is non-positive because

$$\begin{aligned} &\langle \boldsymbol{x}^n(s) - \boldsymbol{x}(s), \boldsymbol{\Gamma}(\boldsymbol{x}^n(s)) - \boldsymbol{\Gamma}(\boldsymbol{x}(s))\rangle \\ =\ &\langle \boldsymbol{x}^n(s) - \boldsymbol{x}(s), \int_0^1 \boldsymbol{R}(\boldsymbol{x}^n(s) + \xi(\boldsymbol{x}^n(s) - \boldsymbol{x}(s)))((\boldsymbol{x}^n(s) - \boldsymbol{x}(s)) \circ \boldsymbol{\mu}^{-1})\rangle \\ =\ &\langle (\boldsymbol{x}^n(s) - \boldsymbol{x}(s)) \circ \boldsymbol{\mu}^{-1/2}, \int_0^1 \boldsymbol{R}(\boldsymbol{x}^n(s) + \xi(\boldsymbol{x}^n(s) - \boldsymbol{x}(s)))((\boldsymbol{x}^n(s) - \boldsymbol{x}(s)) \circ \boldsymbol{\mu}^{-1/2})\rangle \\ \le\ &0, \end{aligned} \tag{92}$$

where the last inequality follows from that the Jacobian matrix $\boldsymbol{R}(\boldsymbol{x}^n(s) + \xi(\boldsymbol{x}^n(s) - \boldsymbol{x}(s)))$ is negative semidefinite a.e. The inequality (91) thus implies that $\|\boldsymbol{x}^n(t) - \boldsymbol{x}(t)\|^2 \to 0$.

It remains to show that $\|\tilde{\boldsymbol{z}}^n\|_T \to 0$ for all fixed $T > 0$. By the functional strong law of large number (e.g., Theorem 5.10 in Chen and Yao (2001)), and $\boldsymbol{\mu^n} \to \boldsymbol{\mu}$, we have

$$\begin{aligned} \tfrac{1}{n}\|N(n\textstyle\int_0^t p_j(\frac{\boldsymbol{X}_j^n(s)}{n\mu_j^n})ds) - \int_0^t np_j(\frac{\boldsymbol{X}_j^n(s)}{n\mu_j^n})ds\|_T &\to 0 \\ \tfrac{1}{n}\|n\mu_j^n W_j^n(t) - S_j^n(W_j^n(t))\|_T &\to 0. \end{aligned} \tag{93}$$

We thus conclude that

$$\|\frac{1}{n}\boldsymbol{Z}^n\|_T \to 0. \tag{94}$$

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

$$\|\int_0^t (\boldsymbol{\Lambda}(\boldsymbol{x}^n(s) \circ (\boldsymbol{\mu}^n)^{-1}) - \boldsymbol{\Lambda}(\boldsymbol{x}^n(s) \circ (\boldsymbol{\mu})^{-1}))ds\|_T \le \int_0^T \|\boldsymbol{\Lambda}(\boldsymbol{x}^n(s) \circ (\boldsymbol{\mu}^n)^{-1}) - \boldsymbol{\Lambda}(\boldsymbol{x}^n(s) \circ (\boldsymbol{\mu})^{-1}))\|ds \to 0 \tag{95}$$

Equations (94) and (95) imply that $\|\tilde{\boldsymbol{z}}^n\|_T \to 0$.

■

Appendix F: Proof of Theorem 2

Proof. We first use (CD-a) and (CD-c) to prove that $-\mathbf{\Lambda}(\cdot) := -(p_j(\cdot))_{j=1,\dots,J}$ satisfies the so-called P-property (Moré and Rheinboldt (1973)). Then by Theorem 2.3 of Moré (1974a) or the comments after Theorem 1.6 of Megiddo and Kojima (1977), the P-property of $-\mathbf{\Lambda}(\boldsymbol{\tau})$ ensures that the solution to the NCP (18) is unique, if exists.

$$\text{P-Property: } \forall \boldsymbol{\tau}^1, \boldsymbol{\tau}^2 \in \mathbb{R}_+^J,\ \boldsymbol{\tau}^1 \neq \boldsymbol{\tau}^2,\ \min_{j=1}^{J}\ (\tau_j^1 - \tau_j^2)(p_j(\boldsymbol{\tau}^1) - p_j(\boldsymbol{\tau}^2)) < 0. \tag{96}$$

Without loss of generality, we assume that $\tau_{j^*}^1 - \tau_{j^*}^2 = \max_j(\tau_j^1 - \tau_j^2) > 0$ for some j^*, and define

$$\overline{\Delta}\tau := \tau_{j^*}^1 - \tau_{j^*}^2. \tag{97}$$

Then to prove (96), it suffices to prove that $p_{j^*}(\boldsymbol{\tau}^1) < p_{j^*}(\boldsymbol{\tau}^2)$. By the definition of $\overline{\Delta}\tau$, we have $\boldsymbol{\tau}^1 \leq \boldsymbol{\tau}^2 + \overline{\Delta}\tau e$, but $\boldsymbol{\tau}_{j^*}^1 = \boldsymbol{\tau}_{j^*}^2 + \overline{\Delta}\tau$. Therefore, (CD-a) implies that

$$p_{j^*}(\boldsymbol{\tau}^1) \leq p_{j^*}(\boldsymbol{\tau}^2 + \overline{\Delta}\tau e). \tag{98}$$

If we define a univariate function $f(x) := p_{j^*}(\boldsymbol{\tau}^2 + x\boldsymbol{e})$ and apply the mean value theorem to $f(\cdot)$[3], we get

$$f(\overline{\Delta}\tau) - f(0) = \overline{\Delta}\tau f'(\zeta). \tag{99}$$

for some $\zeta \in [0, \overline{\Delta}\tau]$. That implies

$$\begin{aligned} p_{j^*}(\boldsymbol{\tau}^2 + \overline{\Delta}\tau\boldsymbol{e}) - p_{j^*}(\boldsymbol{\tau}^2) &= \overline{\Delta}\tau \textstyle\sum_i R_{j^*i}(\boldsymbol{\tau}^2 + \zeta\boldsymbol{e}) \\ &= \overline{\Delta}\tau R_{j^*j^*}(\boldsymbol{\tau}^2 + \zeta\boldsymbol{e}) + \overline{\Delta}\tau \textstyle\sum_{i\neq j^*} R_{j^*i}(\boldsymbol{\tau}^2 + \zeta\boldsymbol{e}) \\ &< 0 \end{aligned} \tag{100}$$

for some $\zeta \in [0, \overline{\Delta}\tau]$, where $R_{ji}(\boldsymbol{\tau}^2 + \zeta\boldsymbol{e})$ represents the entry at the j^{th} row and i^{th} column of the Jacobian matrix evaluated at $\boldsymbol{\tau}^2 + \zeta\boldsymbol{e}$, and the last inequality follows from (CD-c). Inequalities (98) and (100) together imply that $p_{j^*}(\boldsymbol{\tau}^1) < p_{j^*}(\boldsymbol{\tau}^2)$, which leads to the P-property.

We next prove the existence of a solution to the NCP. The most well known sufficient conditions for existence is that the Jacobian of $-\mathbf{\Lambda}(\boldsymbol{\tau})$ is positively bounded, i.e., every principle minor of the Jacobian of $-\mathbf{\Lambda}(\boldsymbol{\tau})$ is bounded between $[\delta, \delta^{-1}]$ for all $\boldsymbol{\tau}$ (Cottle (1966)), or that $-\mathbf{\Lambda}(\boldsymbol{\tau})$ is a uniform P-function, i.e., $\min(\tau_j^1 - \tau_j^2)(p_j(\boldsymbol{\tau}^1) - p_j(\boldsymbol{\tau}^2)) \leq -c\|\boldsymbol{\tau}^1 - \boldsymbol{\tau}^2\|^2$ for some $c > 0$ (Karamardian (1969); Moré (1974b)). Unfortunately, neither condition is satisfied by our $-\mathbf{\Lambda}(\boldsymbol{\tau})$, as its Jacobian can be arbitrarily close to a singular matrix when $\|\boldsymbol{\tau}\| \to \infty$.

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 (18), i.e., (CD-a), (CD-b), and the stability condition (7). Note that (CD-c) is only needed to prove the uniqueness of the solution, but not the 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 $\boldsymbol{\tau}$. This adjustment process is referred to as a tatonnement process in the economics literature Arrow et al. (1959); Walras (2013). We start with $\boldsymbol{\tau} = \mathbf{0}$. In each iteration, we check sequentially if $\mu_j - p_j(\boldsymbol{\tau}) < 0$ for each $j = 1, 2, \dots, J$. Suppose for some j, $\mu_j - p_j(\boldsymbol{\tau}) < 0$, then we increase the value of τ_j and keep the other components of $\boldsymbol{\tau}$ unchanged until $\mu_j - p_j(\boldsymbol{\tau}) = 0$. Such a $\boldsymbol{\tau}$ always exists because $\liminf \mu_j - p_j(\boldsymbol{\tau}) > 0$ by the stability condition (7), and $\mu_j - p_j(\boldsymbol{\tau})$ increases continuously in τ_j by (CD-b). We repeat the above procedure sequentially for $j = 1, 2, \dots, J$ until at some j, $\mu_k - p_k(\boldsymbol{\tau}) \geq 0$ for $k > j$. Note that after τ_j being increased, the value of $\mu_l - p_l(\boldsymbol{\tau})$ can only decrease and turn

[3] The mean value theorem holds even if at some point x, the derivative $f'(x)$ may equal to $+\infty$ or $-\infty$, as long as $f'(x)$ has no jumps.

negative again for some $\ell < j$ due to (CD-a). Therefore, we have to run the above algorithm for another iteration, that is, checking if $\mu_j - p_j(\boldsymbol{\tau}) < 0$ for some j and increase τ_j to make the equality to hold.

According to the above discussion, either at the very beginning $\mu_j - p_j(\boldsymbol{\tau}) > 0$, or $\mu_j - p_j(\boldsymbol{\tau}) \le 0$ throughout the entire algorithm. We use $\boldsymbol{\tau}^N$ to denote the updated value of $\boldsymbol{\tau}$ in the N^{th} iteration. If in some iteration N, $\mu_j - p_j(\boldsymbol{\tau}^N) \ge 0$ for all j, then $\boldsymbol{\tau}^N$ is a solution to the NCP because $\mu_j - p_j(\boldsymbol{\tau}) > 0$ only if the value of τ_j has never been updated (so $\tau_j = 0$); otherwise, we obtain a sequence of waiting-time vectors $\{\boldsymbol{\tau}^N | N = 1, 2, \ldots\}$. We next show that $\boldsymbol{\tau}^N \to \boldsymbol{\tau}^* < \infty$ and $\boldsymbol{\tau}^*$ is the unique solution to the NCP (18).

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

$$\mu_j - p_j(\tau_1^{N_l}, \ldots, \tau_j^{N_l}, \tau_{j+1}^{N_l-1}, \ldots, \tau_J^{N_l-1}) = 0. \tag{101}$$

Since the value of τ_j^N can only increase after each iteration, the monotone convergence theorem implies that $\tau_j \to \tau_j^*$. By the stability condition (7), τ_j^* must be a finite number, otherwise we have $\mu_j - p_j(\boldsymbol{\tau}^N) \to \mu_j - 0 > 0$, which contradicts the complementarity slackness condition. By letting $l \to \infty$ and taking the limit on both sides of equation (101), we get $\mu_j - p_j(\boldsymbol{\tau}^*) = 0$. By repeatedly applying this argument for $j = 1, 2, \ldots, J$, we prove that $(\boldsymbol{\mu} - \boldsymbol{\Lambda}(\boldsymbol{\tau}^*), \boldsymbol{\tau}^*)$ is a solution to the NCP (18). ■

Appendix G: Proof of Theorem 3

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

Suppose $\tau_{j^*}(t) - \tau_{j^*}^* = \overline{\Delta}\boldsymbol{\tau}(t) \ge \delta$. Since $\tau_{j^*}(t) > 0$, the complementarity slackness condition implies that $\mu_j = p_j(\boldsymbol{\tau}^*)$. Thus,

$$\tau'_{j^*}(t) = \frac{X'_{j^*}(t)}{\mu_{j^*}} = \frac{p_{j^*}(\boldsymbol{\tau}(t))}{\mu_{j^*}} - 1 = \frac{p_{j^*}(\boldsymbol{\tau}(t))}{p_{j^*}(\boldsymbol{\tau}^*)} - 1. \tag{102}$$

Note that $\tau'_{j^*}(t)$ exists a.e., because $X_{j^*}(t)$ can be expressed as integrals from 0 to t (See e.g., Equation (87)) and is therefore absolute continuous.

With the above equality, to show that $\tau'_{j^*}(t) \le -h(\delta)$, it suffices to show that

$$\frac{p_{j^*}(\boldsymbol{\tau}(t)) - p_{j^*}(\boldsymbol{\tau}^*)}{p_{j^*}(\boldsymbol{\tau}^*)} \le -h(\delta). \tag{103}$$

We prove the above inequality using a similar argument as in the proof of P-property of Theorem 2. By substituting $\boldsymbol{\tau}^1 = \boldsymbol{\tau}(t)$ and $\boldsymbol{\tau}^2 = \boldsymbol{\tau}^*$ into inequality (98) and (100), we get

$$\begin{aligned} p_{j^*}(\boldsymbol{\tau}(t)) - p_{j^*}(\boldsymbol{\tau}^*) &\le p_{j^*}(\boldsymbol{\tau}^* + \overline{\Delta}\boldsymbol{\tau}(t)e) - p_{j^*}(\boldsymbol{\tau}^*) \\ &\le p_{j^*}(\boldsymbol{\tau}^* + \delta e) - p_{j^*}(\boldsymbol{\tau}^*) \\ &= \delta R_{j^*j^*}(\boldsymbol{\tau}^* + \zeta e) + \delta \textstyle\sum_{i \ne j^*} R_{j^*i}(\boldsymbol{\tau}^* + \zeta e) \end{aligned} \tag{104}$$

for some $\zeta \in [0, \delta]$. In Equation (104), the first inequality follows from inequality (98) (which uses property (CD-a)), and the second inequality follows from $\overline{\Delta}\boldsymbol{\tau}(t) \ge \delta$ and property (CD-c). We then define

$$h(\delta) := \frac{-\delta}{p_j(\boldsymbol{\tau}^*)} \left(\max\{z \in [0, \delta] \mid R_{j^*j^*}(\boldsymbol{\tau}^* + ze) + \sum_{i \ne j^*} R_{j^*i}(\boldsymbol{\tau}^* + ze)\} \right). \tag{105}$$

Using (CD-c), we deduce that $R_{j^*j^*}(\boldsymbol{\tau}^* + ze) + \sum_{i \neq j^*} R_{j^*i}(\boldsymbol{\tau}^* + ze) < 0$ for all $z \in [0, \delta]$. Therefore, $h(\delta)$ is a positive constant that is independent of $\boldsymbol{\tau}(t)$. With $h(\delta)$ defined as in (105), inequality (104) directly implies (103). Therefore, $\tau'_{j^*}(t) = \overline{\Delta}\boldsymbol{\tau}'(t) \leq -h(\delta)$ whenever $\overline{\Delta}\boldsymbol{\tau}(t) \geq \delta$. An analogous argument can be used to prove that $\underline{\Delta}\boldsymbol{\tau}'(t) \geq h(\delta)$ whenever $\underline{\Delta}\boldsymbol{\tau}(t) \leq -\delta$. Therefore, whenever the maximum deviation of $\boldsymbol{\tau}(t)$ from $\boldsymbol{\tau}^*$ has to decrease at a rate of at least $h(\delta)$ whenever it is greater than δ. This guarantees that the maximum deviation must drop below δ after a finite period. The conclusion of Theorem 3 then follows by letting $\delta \to 0$. ■

Appendix H: Proof of Theorem 4

Proof. The optimization problem (21)-(24) is equivalent to the following problem,

$$\max_{\boldsymbol{\tau}^* \geq 0} \quad \ln(1 + \textstyle\sum_{j=1}^{J} \exp(v_j - c\tau_j^*)) \tag{106}$$
$$\text{s.t. } \textstyle\sum_{j=1}^{J} p_j(\boldsymbol{\tau}^*) \leq \overline{\mu}, \quad j = 1, \ldots, J. \tag{107}$$
$$\tag{108}$$

Because if $\boldsymbol{\tau}^*$ is an optimal solution to the above problem, by letting $\mu_j = p_j(\boldsymbol{\tau}^*)$, $(\boldsymbol{\mu}, \boldsymbol{\tau}^*)$ must be an optimal solution to (21)-(24).

Furthermore, we have $\ln(1 + \sum_{j=1}^{J} \exp(v_j - c\tau_j^*)) = -\ln(p_0(\boldsymbol{\tau}))$ as $p_0(\boldsymbol{\tau}) = (1 + \sum_{j=1}^{J} \exp(v_j - c\tau_j^*))^{-1}$, and $\sum_{j=1}^{J} p_j(\boldsymbol{\tau}^*) = 1 - p_0(\boldsymbol{\tau})$. Then the optimization problem (106)-(107) is equivalent to the following problem,

$$\max_{\boldsymbol{\tau}^* \geq 0} \quad -\ln(p_0(\boldsymbol{\tau}^*)) \tag{109}$$
$$\text{s.t. } p_0(\boldsymbol{\tau}^*) \geq 1 - \overline{\mu}, \quad j = 1, \ldots, J. \tag{110}$$
$$\tag{111}$$

Note that $\overline{\lambda} = 1 - p_0(\mathbf{0}) \leq 1 - p_0(\boldsymbol{\tau}^*)$, we have $p_0(\boldsymbol{\tau}^*) = \max\{1 - \overline{\mu}, 1 - \overline{\lambda}\}$, and the optimal objective value is given by $-\ln(\max\{1 - \overline{\mu}, 1 - \overline{\lambda}\})$.

In the case of $\overline{\lambda} \leq \overline{\mu}$, the optimal value is given by $-\ln(1 - \overline{\lambda})$, in which case $\overline{\lambda} = 1 - p_0(\boldsymbol{\tau}^*)$ implies $\boldsymbol{\tau}^* = 0$.

In the case of $\overline{\lambda} < \overline{\mu}$, $p_0(\boldsymbol{\tau}^*) = 1 - \overline{\mu}$, indicating that $\sum_j p_j(\boldsymbol{\tau}^*) = \overline{\mu}$ and thus $p_j(\boldsymbol{\tau}^*) = \mu_j$. In fact, any feasible $(\boldsymbol{\mu}, \boldsymbol{\tau}^*)$ that satisfies $p_j(\boldsymbol{\tau}^*) = \mu_j$ for each j will have the same objective value and thus be an optimal solution. ■

Appendix I: Proof of Proposition 3

Proof. The first-order necessary conditions for the optimization problem (21) - (23) and (25) imply

$$\begin{aligned} -cp_j(\boldsymbol{\tau}^*)(1 - hh_j p_0(\boldsymbol{\tau}^*) + t_j p_0(\boldsymbol{\tau}^*)) &= -s_j \leq 0 \\ s_j = 0 \quad &\text{if } \tau_j^* > 0 \\ t_j = 0 \quad &\text{if } \mu_j > 0, \end{aligned} \tag{112}$$

where h denotes the shadow price for the constraint (25), and $s_j \geq 0$ and $t_j \geq 0$ denotes the shadow prices for the non-negative constraints $\tau_j^* \geq 0$ and $\mu_j \geq 0$, respectively.

If $\boldsymbol{tau}^* \neq \mathbf{0}$, then by defining $k := \min\{j | \tau_j^* > 0\}$, we have $\tau_j^* = 0$ for all $j < k$ by definition. For $j \geq k$, we have $\tau_j^* > 0$ and thus $s_j = 0$. Then the first equation in (112) implies

$$(1 - hh_j + t_j)p_0(\boldsymbol{\tau}^*)) = 0, \text{ for all } j \geq k. \tag{113}$$

Given that $h_k \leq h_{k+1} \leq \ldots \leq h_J$, the above condition implies $t_k \leq t_{k+1} \leq \ldots \leq t_J$. If $h_k < h_j$, then we have $t_j > t_k \geq 0$, implying $\mu_j = 0$. ■

Appendix J: Proof of Lemma 3

We define $n^{1/2}\Delta\boldsymbol{\tau}^n(s)$ for a given $\boldsymbol{Q}^n(s)$ as

$$\begin{aligned} n^{1/2}\Delta\boldsymbol{\tau}^n(s) &= n^{1/2}(n^{1/2}\boldsymbol{Q}^n(s) + n\boldsymbol{\tau}^* \circ \boldsymbol{\mu}) \circ (n\boldsymbol{\mu}^n)^{-1} - \boldsymbol{\tau}^{n,*} \\ &= \boldsymbol{Q}^n(s) \circ (\boldsymbol{\mu}^n)^{-1} + (\boldsymbol{\tau}^* \circ \boldsymbol{\mu} - \boldsymbol{\tau}^{n,*} \circ \boldsymbol{\mu}^n) \circ (\boldsymbol{\mu}^n)^{-1} \end{aligned} \tag{114}$$

Note that the second term at the RHS of (114) converges to $-\boldsymbol{\vartheta} \circ \boldsymbol{\mu}^{-1}$, so the second term must be bounded for all n. Also, the sequence $\{\boldsymbol{\mu}^n\}$ is bounded as it converges to $\boldsymbol{\mu}$. Thus, there exists $\epsilon > 0$, such that for sufficiently large n,

$$n^{1/2}\Delta\boldsymbol{\tau}_j^n(s) + \frac{\vartheta_j}{\mu_j} - \epsilon \leq \frac{\boldsymbol{Q}_j^n(s)}{\mu_j} \leq n^{1/2}\Delta\boldsymbol{\tau}_j^n(s) + \frac{\vartheta_j}{\mu_j} + \epsilon, \tag{115}$$

which implies that $n^{1/2}\Delta\boldsymbol{\tau}^n(s)$ is bounded if and only if $\boldsymbol{Q}^n(s)$ is bounded. We let $\overline{\Delta}\boldsymbol{\tau}^n(t)$ and $\underline{\Delta}\boldsymbol{\tau}^n(t)$ denote the maximal and minimal entries in the vector $\Delta\boldsymbol{\tau}^n(t)$, respectively. To prove Lemma 3, it suffices to prove that for any fixed $T > 0$, when $\kappa \to \infty$,

$$\begin{aligned} &\limsup_n \Pr(\sup\{n^{1/2}\overline{\Delta}\boldsymbol{\tau}^n(t) \mid t \in [0,T]\} > \kappa) \ \to 0 \\ &\limsup_n \Pr(\inf\{n^{1/2}\underline{\Delta}\boldsymbol{\tau}^n(t) \mid t \in [0,T]\} < -\kappa) \to 0 \end{aligned} \tag{116}$$

To prove (116), we first derive an expression for $\boldsymbol{Q}^n$ in analogue to the expression for $\boldsymbol{Q}^{\kappa,n}$ in (35) by ignoring the reflection barrier at $\pm\kappa$,

$$Q_j^n(t) = Q_j^n(0) + \int_0^t \Gamma_j^n(\boldsymbol{\tau}^{n,*} + \Delta\boldsymbol{\tau}^n(s))ds + n^{-1/2}Z_j^n(t) + n^{-1/2}L_j^n(t), \tag{117}$$

where $\Delta\boldsymbol{\tau}^n(s)$ is defined as in (114) for a given $\boldsymbol{Q}^n(s)$, $\Gamma_j^n(\boldsymbol{\tau}) := n^{1/2}\left(p_j(\boldsymbol{\tau}) - \mu_j^n\right)$ represents the deterministic drift that can be non-Lipschitz, and $Z_j^n(t)$ represents a mean-zero stochastic process which was defined in Equation (88).

We next consider the scenario when $n^{1/2}\overline{\Delta}\boldsymbol{\tau}^n(s) = n^{1/2}(\tau_{j^*}^n(s) - \tau_{j^*}^{n,*}) > \delta$ in some interval $[a_1, b_1]$ and for some fixed $j^* \in \{j = 1, \ldots, J\}$. That means, $\boldsymbol{\tau}^n$ has the largest positive deviation from the equilibrium $\boldsymbol{\tau}^{n,*}$ along dimension j^* over $[a_1, b_1]$. Then using the choice-driven property of $\boldsymbol{\Gamma}^n(\boldsymbol{\tau})$ (whose Jacobian is $\boldsymbol{R}(\boldsymbol{\tau})$ so it inherits the choice-driven property), we can prove that over $[a_1, b_1]$, the drift term would be upper bounded by a negative constant (See (122) below), and consequently the deviation $\overline{\Delta}\boldsymbol{\tau}^n(s)$ would decrease by at least an amount proportional to $b_1 - a_1$ (See (125)).

Formally, we have

$$\begin{aligned} \Gamma_{j^*}^n(\boldsymbol{\tau}^{n,*} + \Delta\boldsymbol{\tau}^n(s)) &= n^{1/2}\left(p_{j^*}(\boldsymbol{\tau}^{n,*} + \Delta\boldsymbol{\tau}^n(s)) - \mu_{j^*}^n\right) \\ &= n^{1/2}\left(p_{j^*}(\boldsymbol{\tau}^{n,*} + \Delta\boldsymbol{\tau}^n(s)) - p_{j^*}(\boldsymbol{\tau}^{n,*})\right) + n^{1/2}(p_{j^*}(\boldsymbol{\tau}^{n,*}) - \mu_{j^*}^n). \end{aligned} \tag{118}$$

We next provide an upper bound for the RHS of Equation (118). In inequality (103) (which builds on the choice-driven property), by replacing $\boldsymbol{\tau}(t)$ with $\boldsymbol{\tau}^{n,*} + \Delta\boldsymbol{\tau}^n(s)$, and by noting that $\overline{\Delta}\tau^n(s) \geq n^{-1/2}\delta$, we get

$$p_{j^*}(\boldsymbol{\tau}^{n,*} + \Delta\tau^n(s)) - p_{j^*}(\boldsymbol{\tau}^{n,*}) \leq -n^{-1/2}h^n(\delta). \tag{119}$$

where $h^n(\cdot)$ follows a similar functional form of $h(\cdot)$ as given in Equation (105), that is,

$$h^n(\delta) := \frac{-\delta}{p_j(\boldsymbol{\tau}^{n,*})}\left(\max\{z \in [0, n^{-1/2}\delta] \mid R_{j^*j^*}(\boldsymbol{\tau}^{n,*} + ze) + \sum_{i \neq j^*} R_{j^*i}(\boldsymbol{\tau}^{n,*} + ze)\}\right) (> 0) \tag{120}$$

Inequality (119) allows us to upper bound the RHS of (118) as

$$\begin{aligned} \Gamma_{j^*}^n(\boldsymbol{\tau}^{n,*} + \Delta\boldsymbol{\tau}^n(s)) \ &\leq \ -h^n(\delta) + n^{1/2}(p_{j^*}(\boldsymbol{\tau}^{n,*}) - \mu_{j^*}^n) \\ &\to \ \tfrac{\delta}{p_j(\boldsymbol{\tau}^{n,*})}\left(R_{j^*j^*}(\boldsymbol{\tau}^{n,*}) + \textstyle\sum_{i \neq j^*} R_{j^*i}(\boldsymbol{\tau}^{n,*})\right) - \theta_{j^*} \end{aligned} \tag{121}$$

That means, for sufficiently large n,

$$\Gamma^n_{j^*}(\boldsymbol{\tau}^{n,*}+\Delta\boldsymbol{\tau}^n(s)) < \frac{\delta}{p_j(\boldsymbol{\tau}^{n,*})}\left(R_{j^*j^*}(\boldsymbol{\tau}^{n,*})+\sum_{i\neq j^*}R_{j^*i}(\boldsymbol{\tau}^{n,*})\right)-\theta_{j^*} := -\Delta_n < 0 \tag{122}$$

where $R_{j^*j^*}(\boldsymbol{\tau}^{n,*})+\sum_{i\neq j^*}R_{j^*i}(\boldsymbol{\tau}^{n,*})<0$ by the choice-driven property. By looking into the sequence $\{\Delta_n\}$, we deduce that it converges to some positive constant, $\Delta>0$. Inequalities (117) and (122) imply that

$$Q^n_{j^*}(b_1)-Q^n_{j^*}(a_1) \le -\Delta_n(b_1-a_1)+n^{-1/2}(Z^n_{j^*}(b_1)-Z^n_{j^*}(a_1))+n^{-1/2}(L_{j^*}(b_1)-L_{j^*}(a_1)) \tag{123}$$

If $j^*\in\mathcal{J}^-\cup\mathcal{J}^+$, then $\tau^n_{j^*}(s)-\tau^{n,*}_{j^*}>0$ implies that $Q^n_{j^*}(s)>0$ over $[a_1,b_1)$. Consequently, $L_{j^*}(b_1)-L_{j^*}(a_1)=0$. If $j^*\in\mathcal{J}^{++}$, then there is no reflection barrier along dimension j^*, so $L_{j^*}\equiv 0$. Thus in either case, $L_i(b_1)-L_i(a_1)=0$ and inequality (123) implies that

$$Q^n_{j^*}(b_1)-Q^n_{j^*}(a_1) \le -\Delta_n(b_1-a_1)+n^{-1/2}(Z^n_i(b_1)-Z^n_i(a_1)). \tag{124}$$

which leads to

$$\begin{aligned} n^{1/2}(\overline{\Delta}\tau^n(b_1)-\overline{\Delta}\tau^n(a_1)) &= n^{1/2}(\tau^n_{j^*}(b_1)-\tau^n_{j^*}(a_1)) \\ &= \tfrac{1}{\mu^n_j}\left(Q^n_{j^*}(b_1)-Q^n_{j^*}(a_1)\right). \\ &\le \tfrac{1}{\mu^n_j}\left(-\Delta_n(b_1-a_1)+n^{-1/2}(Z^n_{j^*}(b_1)-Z^n_{j^*}(a_1))\right) \end{aligned} \tag{125}$$

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

$$\begin{aligned} n^{1/2}(\overline{\Delta}\tau^n(b)-\overline{\Delta}\tau^n(a)) &= \textstyle\sum_{i=1}^\infty n^{1/2}(\overline{\Delta}\tau^n(b_i)-\overline{\Delta}\tau^n(a_i)) \\ &\le \textstyle\sum_{i=1}^\infty \frac{1}{\mu^n_{j^i}}\left(-\Delta_n(b_i-a_i)+n^{-1/2}(Z^n_{j^i}(b_i)-Z^n_{j^i}(a_i))\right) \\ &\le \tfrac{1}{\min_j\mu^n_j}\left(-\Delta_n(b-a)+n^{-1/2}\|\boldsymbol{Z}^n(b-a)\|\right) \end{aligned} \tag{126}$$

Now let $\delta=\frac{\kappa}{2}$. If $\overline{\Delta}\tau^n(\cdot)$ has ever exceeded $\frac{\kappa}{2}$ over $[0,t]$, then we let $a=\sup\{s\in[0,t]:\overline{\Delta}\tau^n(s)\le\frac{\kappa}{2}\}$ and $b=t$. The selection of a and b guarantees that $\overline{\Delta}\tau^n(a)=\frac{\kappa}{2}$ and $\overline{\Delta}\tau^n(s)\ge\frac{\kappa}{2}$ for all $s\in[a,b]$. Thus, Equation (126) implies that[4]

$$\begin{aligned} n^{1/2}\overline{\Delta}\tau^n(t)-\tfrac{\kappa}{2} &= n^{1/2}(\overline{\Delta}\tau^n(b)-\overline{\Delta}\tau^n(a)) \\ &\le \tfrac{1}{\min_j\mu^n_j}\left(n^{-1/2}\|\boldsymbol{Z}^n\|_t\right). \end{aligned} \tag{127}$$

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

$$\begin{aligned} n^{1/2}\sup\{\overline{\Delta}\tau^n(t)\mid t\in[0,T]\} &\le \tfrac{\kappa}{2}+\tfrac{1}{\min_j\mu^n_j}n^{-1/2}\sup\{\|\boldsymbol{Z}^n(t)\|\mid t\in[0,T]\} \\ &= \tfrac{\kappa}{2}+\tfrac{1}{\min_j\mu^n_j}n^{-1/2}\|\boldsymbol{Z}^n\|_T. \end{aligned} \tag{128}$$

[4] To derive (127), we have only used a weaker upper bound (126) for $\overline{\Delta}\tau^n(b)-\overline{\Delta}\tau^n(a)$ by ignoring the negative drift $-\Delta_n(b-a)$. The original upper bound (126) including $-\Delta_n(b-a)$, however, is needed in the later proof for Proposition 5.

When $\kappa \to \infty$, we deduce that

$$\begin{aligned}
&\limsup_n \Pr(\sup\{n^{1/2}\overline{\Delta}\tau^n(t) \mid t\in[0,T]\} > \kappa) \\
&\le \limsup_n \Pr(\sup\{n^{1/2}\overline{\Delta}\tau^n(t) \mid t\in[0,T]\} > \kappa \mid n^{1/2}\overline{\Delta}\tau^n(0) \le \tfrac{\kappa}{2}) \Pr(n^{1/2}\overline{\Delta}\tau^n(0) \le \tfrac{\kappa}{2}) \\
&\quad + \limsup_n \Pr(n^{1/2}\overline{\Delta}\tau^n(0) > \tfrac{\kappa}{2}) \\
&\to \limsup_n \Pr(\sup\{n^{1/2}\overline{\Delta}\tau^n(t) \mid t\in[0,T]\} > \kappa \mid n^{1/2}\overline{\Delta}\tau^n(0) \le \tfrac{\kappa}{2}) \cdot 1 + 0 \\
&\le \limsup_n \Pr(\sup_{t\in[0,T]} \tfrac{1}{\min_j \mu_j^n} n^{-1/2} \|\boldsymbol{Z}^n\|_T > \tfrac{\kappa}{2}) \\
&\le \sup_n 2c_1 \exp(-\tfrac{c_2}{4}\kappa^2) + 2n^{c_3}\exp(-\tfrac{c_4}{2}\kappa\sqrt{n})
\end{aligned} \tag{129}$$

for some positive constants c_i $(i=1,2,3,4)$. In Equation (129), the convergence result follows from $\limsup_n \Pr(n^{1/2}\overline{\Delta}\tau^n(0) > \frac{\kappa}{2}) \to 0$ as $\boldsymbol{Q}^n(0)$ (so $n^{1/2}\Delta\boldsymbol{\tau}^n(0)$) is assumed to have finite expectation; the second inequality follows from (128), and the last inequality follows from the upper bound (149) for the tail probability of $n^{-1/2}\|\boldsymbol{Z}^n\|_T$ (See Lemma 4 in Appendix O). Note that the second term of RHS in Equation (129) is dominated by $\exp(-\frac{c_4}{4}\kappa\sqrt{n})$ when n is large, so the RHS has to converge to zero when $\kappa\to\infty$, which leads to the first convergence equation in (116).

The second convergence in (116) can be proved using an analogous argument and is omitted here.

Appendix K: Proof of Theorem 5

Since $\|\boldsymbol{Q}^{\kappa,n}\|_T \le \kappa$, if we define the waiting-time vector associated with $\boldsymbol{Q}^{\kappa,n}$ as

$$\boldsymbol{\tau}^{\kappa,n}(t) = (n^{1/2}\boldsymbol{Q}^{\kappa,n}(t) + n\boldsymbol{\tau}^* \circ \boldsymbol{\mu}^*) \circ (n\boldsymbol{\mu}^n)^{-1}, \tag{130}$$

then $\|\boldsymbol{\tau}^{\kappa,n} - \boldsymbol{\tau}^*\|_T \to 0$. We can then select a neighborhood $\mathcal{N}$ of $\boldsymbol{\tau}^*$, such that $\boldsymbol{\tau}^{\kappa,n} \in \mathcal{N}$ for all sufficiently large n, and the arrival rate function $\boldsymbol{\Lambda}(\cdot)$ is Lipschitz continuous in $\mathcal{N}$. The latter holds because $\boldsymbol{\Lambda}(\cdot)$ has bounded Jacobian $\boldsymbol{R}^*$ at $\boldsymbol{\tau}^*$, and the Jacobian is continuous everywhere. Therefore, the state-dependent arrival rate of the process $\boldsymbol{Q}^{\kappa,n}$ is Lipschitz continuous over its domain. Hence, we can invoke Theorem 7.2 in Mandelbaum et al. (1998b) and show that

$$\{\boldsymbol{Q}^{\kappa,n}(t) | 0 \le t \le T\} \Rightarrow \{\boldsymbol{Y}^{\kappa}(t) | 0 \le t \le T\}. \tag{131}$$

Finally, for all bounded, continuous real-valued function f with domain $D([0,T),\mathbb{R}^J)$, when $\kappa\to\infty$, we have

$$\begin{aligned}
&\limsup_{n\to\infty} |\mathbb{E}f(\boldsymbol{Q}^n) - \mathbb{E}f(\boldsymbol{Y})| \\
&\le \limsup_{n\to\infty} |\mathbb{E}f(\boldsymbol{Q}^n) - \mathbb{E}f(\boldsymbol{Q}^{\kappa,n})| + \limsup_{n\to\infty} |\mathbb{E}f(\boldsymbol{Q}^{\kappa,n}) - \mathbb{E}f(\boldsymbol{Y}^{\kappa,n})| + |\mathbb{E}f(\boldsymbol{Y}^{\kappa}) - \mathbb{E}f(\boldsymbol{Y})| \\
&\le \limsup_{n\to\infty} 2\overline{f}\Pr(\|\boldsymbol{Q}^n - \boldsymbol{Q}^{\kappa,n}\|_T \ne 0) + 0 + |\mathbb{E}f(\boldsymbol{Y}^{\kappa}) - \mathbb{E}f(\boldsymbol{Y})| \\
&\to 0
\end{aligned} \tag{132}$$

where $\overline{f}$ represents an upper bound for $|f|$, $\limsup_{n\to\infty} |\mathbb{E}f(\boldsymbol{Q}^{\kappa,n}) - \mathbb{E}f(\boldsymbol{Y}^{\kappa,n})| = 0$ follows from Equation (131), $\limsup_{n\to\infty} 2\overline{f}\Pr(\|\boldsymbol{Q}^n - \boldsymbol{Q}^{\kappa,n}\|_T \ne 0) \to 0$ follows from Lemma 3, and $|\mathbb{E}f(\boldsymbol{Y}^{\kappa}) - \mathbb{E}f(\boldsymbol{Y})| \to 0$ follows from bounded convergence and the continuous mapping theorem. Equation (132) implies that $\boldsymbol{Q}^n \Rightarrow \boldsymbol{Y}$. ■

Appendix L: Proof of Proposition 4

Proof. When reflections are absent, the density of the stationary distribution of $\boldsymbol{Y}$ follows the classical results pertaining to the O-U process (Meucci, 2009; Vatiwutipong and Phewchean, 2019). For the case involving reflection barriers, as described by Example 3.10, Claim 1 in the work of Kang and Ramanan (2014), the situation is as follows: If the diffusion limit is a solution to an SDER with affine drift coefficient $\boldsymbol{Cx}$, and if $\boldsymbol{C}^* := [\boldsymbol{A} - \overline{\boldsymbol{N}}^{-1}\overline{\boldsymbol{Q}}]^{-1}\boldsymbol{C}$ (see definitions in Kang and Ramanan

(2014)) is symmetric, then $p(\boldsymbol{x}) = e^{\boldsymbol{x}^T \boldsymbol{C}_* \boldsymbol{x}}$, after normalization, gives the stationary distribution of the diffusion limit. We next check whether with the parameters in our setting, $\boldsymbol{C}_*$ is symmetric and $p(\boldsymbol{x})$ is proportional to $\pi(\boldsymbol{z})$ as defined in the proposition. Because in our model the reflection direction is always normal, it has zero component tangential to the boundary. Thus, we have $\overline{\boldsymbol{Q}} = 0$, because its rows are exactly the tangential components of the reflection direction according to the comments after Theorem 3 in Kang and Ramanan (2014). Consequently, by comparing the SDER in Kang and Ramanan (2014) to Equation (32), we have $\boldsymbol{A} = \boldsymbol{\Sigma} = (1+\omega_1^2)\text{Diag}\,(\boldsymbol{\mu})$, $\boldsymbol{x} = \boldsymbol{z} - \boldsymbol{\vartheta} - (\text{Diag}\,(\boldsymbol{\mu})\boldsymbol{R}^*)^{-1}\boldsymbol{\theta}$ and $\boldsymbol{C} = \boldsymbol{R}^*\text{Diag}\,(\boldsymbol{\mu}^{-1})$. Thus, $\boldsymbol{C}^* := \boldsymbol{A}^{-1}\boldsymbol{C} = (1+\omega_1^2)^{-1}\text{Diag}\,(\boldsymbol{\mu}^{-1})\boldsymbol{R}^*\text{Diag}\,(\boldsymbol{\mu}^{-1})$ is symmetric and negative definite as $\boldsymbol{R}^*$ is symmetric and negative definite. We thus conclude that

$$\begin{aligned} & p(\boldsymbol{x}) = \exp(\boldsymbol{x}^T \boldsymbol{C}_* \boldsymbol{x}) \\ &= \exp((\boldsymbol{z} - \boldsymbol{\vartheta} - (\text{Diag}\,(\boldsymbol{\mu})\boldsymbol{R}^*)^{-1}\boldsymbol{\theta})^T((1+\omega_1^2)^{-1}\text{Diag}\,(\boldsymbol{\mu}^{-1})\boldsymbol{R}^*\text{Diag}\,(\boldsymbol{\mu}^{-1}))(\boldsymbol{z} - \boldsymbol{\vartheta} - (\text{Diag}\,(\boldsymbol{\mu})\boldsymbol{R}^*)^{-1}\boldsymbol{\theta})) \\ &= \exp(-\tfrac{1}{2}(\boldsymbol{z} - \boldsymbol{\vartheta} - (\text{Diag}\,(\boldsymbol{\mu})\boldsymbol{R}^*)^{-1}\boldsymbol{\theta})^T(-\tfrac{1}{2}(1+\omega_1^2)\text{Diag}\,(\boldsymbol{\mu})(\boldsymbol{R}^*)^{-1}\text{Diag}\,(\boldsymbol{\mu}))^{-1} \\ &\qquad (\boldsymbol{z} - \boldsymbol{\vartheta} - (\text{Diag}\,(\boldsymbol{\mu})\boldsymbol{R}^*)^{-1}\boldsymbol{\theta})) \end{aligned} \tag{133}$$

is proportional to the density of the stationary distribution of the diffusion limit, $\pi_{\boldsymbol{Y}}(\boldsymbol{z})$. By looking into the above expression, we find that $p(\boldsymbol{x})$ is proportional to the density of a multivariate Gaussian random variable with mean $\boldsymbol{\vartheta} + (\text{Diag}\,(\boldsymbol{\mu})\boldsymbol{R}^*)^{-1}\boldsymbol{\theta}$ and covariance matrix $-\frac{1}{2}(1+\omega_1^2)\text{Diag}\,(\boldsymbol{\mu})(\boldsymbol{R}^*)^{-1}\text{Diag}\,(\boldsymbol{\mu})$, which is denoted by $\pi(\boldsymbol{z})$. Therefore, $\pi_{\boldsymbol{Y}}(\boldsymbol{z})$ is proportional to $\pi(\boldsymbol{z})$. Normalizing $\pi(\boldsymbol{z})$ thus leads to an exact expression for $\pi_{\boldsymbol{Y}}(\boldsymbol{z})$ in (37). ■

Appendix M: Proof of Proposition 5

Equation (115) implies that when n is sufficiently large, the difference between $V(\boldsymbol{\Xi}^n(t))(= \|\boldsymbol{Q}^n\|^{\boldsymbol{\mu}^{-1}})$ and $\|n^{1/2}\Delta\boldsymbol{\tau}^n(t)\|$ is almost a constant (i.e., within $\pm\epsilon$). So proving Equation (42) is equivalent to proving the same bounded condition for $\|n^{1/2}\Delta\boldsymbol{\tau}^n(t)\|$, that is, for some $u_0 > 0$, $t_0 \geq 0$,

$$\begin{aligned} &\limsup_{n\to\infty}\sup_{\boldsymbol{\Xi}^n(0)\in\Omega} \mathbb{E}[\exp\left(u_0(\|n^{1/2}\Delta\boldsymbol{\tau}^n(t_0)\| - \|n^{1/2}\Delta\boldsymbol{\tau}^n(0)\|)^+\right) \mid \boldsymbol{\Xi}^n(0)] < \infty \\ &\limsup_{n\to\infty}\sup_{\boldsymbol{\Xi}^n(0)\in\Omega} \mathbb{E}[(\|n^{1/2}\Delta\boldsymbol{\tau}^n(t_0)\| - \|n^{1/2}\Delta\boldsymbol{\tau}^n(0)\|)^2 \\ &\qquad\qquad \exp\left(u(\|n^{1/2}\Delta\boldsymbol{\tau}^n(t_0)\| - \|n^{1/2}\Delta\boldsymbol{\tau}^n(0)\|)^+\right) \mid \boldsymbol{\Xi}^n(0)] < \infty \end{aligned} \tag{134}$$

To prove (134), we first consider the case when $\|n^{1/2}\Delta\boldsymbol{\tau}^n(s)\| > \frac{\kappa}{2}$ for all $s \in [0,T]$. By Equation (126) (which builds on the choice-driven properties of the arrival rate) and by plugging into $a = 0$ and $b = t_0$, we have

$$n^{1/2}\|\Delta\tau^n(t)\| - n^{1/2}\|\Delta\tau^n(0)\| \leq \tfrac{1}{\min_j \mu_j^n}\left(-\Delta_n t_0 + n^{-1/2}\|\boldsymbol{Z}^n(t_0)\|\right) \tag{135}$$

where Δ_n was defined in (122), which converges to a positive constant $\Delta > 0$. By choosing

$$t_0 = \frac{\min_j \mu_j^n}{\Delta}\left(n^{1/2}\|\Delta\tau^n(0)\| - \frac{\kappa}{2}\right)^+, \tag{136}$$

for sufficiently large n, Equation (135) implies that

$$n^{1/2}\|\Delta\tau^n(t)\| \leq \frac{\kappa}{2} + \frac{1}{\min_j \mu_j^n} n^{-1/2}\|\boldsymbol{Z}^n(t_0)\|. \tag{137}$$

In the other case when $\|n^{1/2}\Delta\boldsymbol{\tau}^n(s)\| \leq \frac{\kappa}{2}$ for some $s \in [0,T]$, we can also deduce (137) using a similar argument as we establish inequality (128) in the proof for Lemma 3.

In view of (137), we deduce that there exists $u_0 > 0$ such that

$$\begin{aligned} &\limsup_{n\to\infty} \sup_{\boldsymbol{\Xi}^n(0)\in\Omega} \mathbb{E}[\exp(u_0(\|n^{1/2}\Delta\boldsymbol{\tau}^n(t_0)\| - \|n^{1/2}\Delta\boldsymbol{\tau}^n(0)\|)^+) \mid \boldsymbol{\Xi}^n(0)] \\ \le &\limsup_{n\to\infty} \sup_{\boldsymbol{\Xi}^n(0)\in\Omega} \mathbb{E}[\exp(u_0\|n^{1/2}\Delta\boldsymbol{\tau}^n(t_0)\|) \mid \boldsymbol{\Xi}^n(0)] \\ \le &\limsup_{n\to\infty} \sup_{\boldsymbol{\Xi}^n(0)\in\Omega} \mathbb{E}[\exp(u_0(\tfrac{\kappa}{2} + \tfrac{1}{\min_j \mu_j^n} n^{-1/2}\|\boldsymbol{Z}^n(t_0)\|))|\boldsymbol{\Xi}^n(0)] \\ < &+\infty, \end{aligned} \tag{138}$$

where the last inequality follows from (144) in Lemma (4) (See Appendix O). Similarly, there exists $u_0 > 0$, such that

$$\begin{aligned} &\limsup_{n\to\infty} \sup_{\boldsymbol{\Xi}^n(0)\in\Omega} \mathbb{E}[(\|n^{1/2}\Delta\boldsymbol{\tau}^n(t_0)\| - \|n^{1/2}\Delta\boldsymbol{\tau}^n(0)\|)^2 \\ &\qquad \exp(u(\|n^{1/2}\Delta\boldsymbol{\tau}^n(t_0)\| - \|n^{1/2}\Delta\boldsymbol{\tau}^n(t_0)\|)^+) \mid \boldsymbol{\Xi}^n(0)] \\ \le &\limsup_{n\to\infty} \sup_{\boldsymbol{\Xi}^n(0)\in\Omega} \mathbb{E}[(\max\{n^{1/2}\Delta\boldsymbol{\tau}^n(0), \tfrac{\kappa}{2} + \tfrac{1}{\min_j \mu_j^n} n^{-1/2}\|\boldsymbol{Z}^n(t_0)\|\})^2 \\ &\qquad \exp(u_0(\tfrac{\kappa}{2} + \tfrac{1}{\min_j \mu_j^n} n^{-1/2}\|\boldsymbol{Z}^n(t_0)\|))|\boldsymbol{\Xi}^n(0)] \\ < &+\infty, \end{aligned} \tag{139}$$

where the last inequality follows from (145) in Lemma (4). We have thus proved (134), and thus (42) in Proposition 5.

It remains to show that $V(\cdot)$ is a Lyapunov function with drift size parameter -1, drift term parameter t_0, and exception parameter κ for $\boldsymbol{\Xi}$, or equivalently, to prove condition (40) for $\gamma = 1$. Because $V(\boldsymbol{\Xi}^n(t))$ and $n^{-1/2}\|\Delta\boldsymbol{\tau}^n(t_0)\|$ only differs by almost a constant, proving (40) is equivalent to proving the same condition for $\|n^{1/2}\Delta\boldsymbol{\tau}^n(t)\|$ for some positive constant γ. To that end, we choose t_0 as (136) and get

$$\begin{aligned} &\sup_{\|n^{1/2}\Delta\boldsymbol{\tau}^n(0)\|>\kappa} \{\mathbb{E}[\|n^{1/2}\Delta\boldsymbol{\tau}^n(t_0)\| \mid \|n^{1/2}\Delta\boldsymbol{\tau}^n(0)\|\} \\ \le &\sup_{\|n^{1/2}\Delta\boldsymbol{\tau}^n(0)\|>\kappa} \{\mathbb{E}[\tfrac{\kappa}{2} + \tfrac{1}{\min_j \mu_j^n} n^{-1/2}\|\boldsymbol{Z}^n(t_0)\| \mid \|n^{1/2}\Delta\boldsymbol{\tau}^n(0)\|]\} - \kappa \\ \le &\, c' - \tfrac{\kappa}{2} \end{aligned} \tag{140}$$

for some constant $c' > 0$. In (140), the first inequality follows from inequality (137) and that $\|n^{1/2}\Delta\boldsymbol{\tau}^n(0)\| > \kappa$, and the second inequality follows from (143) in Lemma 4 that $n^{-1/2}\|\boldsymbol{Z}^n(t_0)\|$ is uniformly upper bounded. By choosing a sufficiently large κ, we can have $c' - \frac{\kappa}{2} < -1$, which proves that $V(\cdot)$ is a Lyapunov function with drift size parameter -1. ■

Appendix N: Proof of Theorem 6

Proof. By Proposition 5, $V(\cdot)$ is a Lyapunov function with parameter -1, t_0, and κ. Moreover, the second inequality in (42) 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 Gamarnik and Zeevi (2006) 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 n,

$$\Pr_{\boldsymbol{\pi}^n}(\|\boldsymbol{Q}^n(0)\|_T > s) \le (1 - u_0/2)^{-1} L_1(u_0, t_0, n) \exp(-u_0(s - \kappa)). \tag{141}$$

By the above inequality and the inequality in (42), we have

$$\Pr_{\boldsymbol{\pi}^n}(\|\boldsymbol{Q}^n(0)\|_T > s) \le H_1 \exp(-h_1 s), \tag{142}$$

for properly selected constants H_1 and h_1. Inequality (142) implies uniform tightness of the sequence of distributions $(\boldsymbol{\pi}^n)$. The rest of the proof follows exactly as in Theorem 8 of Gamarnik and Zeevi (2006). ■

Appendix O: Lemma 4 and its Proof

The following Lemma was used in both Lemma 3 and Proposition 5.

Lemma 4 *There exists a constant $u_0 > 0$, such that the following inequalities hold for all fixed $t_0 \geq 0$,*

$$\limsup_{n\to\infty} \sup_{\|\boldsymbol{\Xi}^n(0)-\boldsymbol{\vartheta}\|>\kappa} n^{-1/2}\mathbb{E}[\|\boldsymbol{Z}^n\|_{t_0}|\boldsymbol{\Xi}^n(0)] < \infty, \tag{143}$$

$$\limsup_{n\to\infty} \sup_{\boldsymbol{\Xi}^n(0)\in\Omega} \mathbb{E}[\exp(n^{-1/2}u_0\|\boldsymbol{Z}^n\|_{t_0})|\boldsymbol{\Xi}^n(0)] < \infty, \tag{144}$$

$$\limsup_{n\to\infty} \sup_{\boldsymbol{\Xi}^n(0)\in\Omega} \mathbb{E}[\|\boldsymbol{Z}^n\|_{t_0}^2 \exp(n^{-1/2}u_0\|\boldsymbol{Z}^n\|_{t_0})|\boldsymbol{\Xi}^n(0)] < \infty, \tag{145}$$

where $\boldsymbol{\Xi}^n(0)$ gives the initial state of the Markovian process, and $\boldsymbol{Z}^n(t)$ is a J-dimensional centered process defined in (88)*.*

Proof. Using the argument provided at the beginning of the proof for Lemma A.1 in Gamarnik and Zeevi (2006), inequality (144) implies (143) and (145). To prove (144), define $A_j^n(t) := \int_0^t p_j(\boldsymbol{X}^n(s)\circ(n\boldsymbol{\mu}^n)^{-1})ds$. Let $S_j^*(t)$ denote the cumulative number of customers that have completed service at the j^{th} service provider up to time t,

By change of the time variables, we can derive the following bound for $n^{-1/2}\|Z_j^n\|_{t_0}$,

$$\begin{aligned}
&n^{-1/2}\|Z_j^n\|_{t_0}\\
\leq\ &\|n^{-1/2}(N(nt)-nt)\|_{A_j^n(t_0)} + \|n^{-1/2}(n\mu_j^n t - S_j^n(t))\|_{W_j^n(t_0)}\\
=\ &\|n^{-1/2}(N(t)-t)\|_{nA_j^n(t_0)} + \|n^{-1/2}(t - S_j^n(\tfrac{t}{n\mu_j^n}))\|_{n\mu_j^n W_j^n(t_0)}\\
\leq\ &\|n^{-1/2}(N(t)-t)\|_{nt_0} + \|n^{-1/2}(t - S_j^n(\tfrac{t}{n\mu_j^n}))\|_{2n\mu_j t_0}\\
\leq\ &n^{-1/2}\|N(t)-(t+B_j(t))\|_{nt_0} + n^{-1/2}\|B_j\|_{2n\mu_j t_0}\\
&+n^{-1/2}\|S_j^n(t)-(t+B_j'(t))\|_{2n\mu_j t_0} + n^{-1/2}\|B_j'(t)\|_{2n\mu_j t_0}
\end{aligned} \tag{146}$$

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, $\boldsymbol{B} = (B_j)$ and $\boldsymbol{B}' = (B_j')$ denote two independent J-dimensional standard Brownian motions.

We next derive the tail bounds for each term at the RHS of (146). 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,

$$\begin{aligned}
\Pr(\|B_j\|_{nt_0} > \tfrac{1}{4}a\sqrt{n}) &= c_1\exp(-c_2a^2)\\
\Pr(\|B_j'\|_{nt_0} > \tfrac{1}{4}a\sqrt{n}) &= c_1\exp(-c_2a^2).
\end{aligned} \tag{147}$$

Using the functional strong approximation theorem (FSAT) (Theorem 5.14 and Remark 5.17 in Chen and Yao (2001)), we may upper bound the tail probability of the other two terms in (146) with constants $c_3, c_4 > 0$ as follows:

$$\begin{aligned}
\Pr(n^{-1/2}\|N(t)-(t+B_j(t))\|_{nt_0} \geq \tfrac{1}{4}a) &\leq n^{c_3}\exp(-c_4an^{-1/2})\\
\Pr(n^{-1/2}\|S_j^n(t)-(t+B_j'(t))\|_{2n\mu_j t_0} \geq \tfrac{1}{4}a) &\leq n^{c_3}\exp(-c_4an^{-1/2})
\end{aligned} \tag{148}$$

(146), (147), and (148) together imply that

$$\Pr(n^{-1/2}\|Z_j^n\|_{t_0} \geq a) \leq 2c_1\exp(-c_2a^2) + 2n^{c_3}\exp(-c_4a\sqrt{n}). \tag{149}$$

We can then upper bound the expectation $\mathbb{E}[\exp(n^{-1/2}u_0\|\boldsymbol{Z}^n\|_{t_0})|\boldsymbol{\Xi}^n(0)]$ for all sufficiently large n and initial state $\boldsymbol{\Xi}^n(0)$ using the tail probability bounds,

$$\begin{aligned}
&\mathbb{E}[\exp(n^{-1/2}u_0\|\boldsymbol{Z}^n\|_{t_0})|\boldsymbol{\Xi}^n(0)]\\
\leq\ &2 + \int_2^\infty \Pr\left(\exp(n^{-1/2}u_0\|\boldsymbol{Z}^n\|_{t_0}) > a\right)da\\
=\ &2 + \int_2^\infty \Pr\left(\exp(n^{-1/2}\|\boldsymbol{Z}^n\|_{t_0}) > \tfrac{\log x}{u_0}\right)dx\\
\leq\ &2 + \int_2^\infty 2c_1\exp(-c_2\tfrac{\log^2 x}{u_0^2})dx + \int_2^\infty 2n^{c_3}\exp(-c_4\tfrac{\log x}{u_0}n^{-1/2})dx\\
<\ &2M,
\end{aligned} \tag{150}$$

where the second inequality follows from (149) 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 (144). ■

Appendix P: Proof of Proposition 7

Proof. It suffices to prove the η version of (131), that is, $\boldsymbol{Q}^{\kappa,\eta} \Rightarrow \boldsymbol{Y}^{\kappa,\eta}$ over a compact domain $\Omega(\kappa)$, where $\boldsymbol{Q}^{\kappa,\eta}$, $\boldsymbol{Y}^{\kappa,\eta}$ follow an analogous definition as $\boldsymbol{Q}^{\kappa}$ and $\boldsymbol{Y}^{\kappa}$ in (35), and $\Omega(\kappa)$ follows the definition in (34). In the rest of the proof, we omit the superscript κ for brevity.

The dynamics of queue j implies the following equation,

$$\begin{aligned} Q_j^{\eta,n}(t) &= Q_j^{\eta,n}(0) + n^{-1/2}N\left(\eta n\int_0^t p_j(\boldsymbol{\tau}^{\eta,n}(s))ds + (1-\eta)n\int_0^t p_j(\boldsymbol{\tau}^{n,*}(s))ds\right) - n^{-1/2}S_j^{\eta,n}(t) \\ &\quad + n^{-1/2}L_j^{\eta,n}(t) - n^{-1/2}U_j^{\eta,n}(t) \\ &= \underbrace{Q_j^{\eta,n}(0)}_{\text{(A.1)}} + \underbrace{\int_0^t \eta\left(n^{1/2}\left(p_j(\boldsymbol{\tau}^{\eta,n}(s)) - \mu_j^n\right) - \sum_i \frac{R^*_{ji}(Q_i^{\eta,n}(s)-\vartheta_i)}{\mu_i^n}\right)ds}_{\text{(A.2)}} \\ &\quad + \underbrace{\int_0^t (1-\eta)n^{1/2}\left(p_j(\boldsymbol{\tau}^{n,*}) - \mu_j^n\right)ds}_{\text{(A.3)}} + \underbrace{n^{-1/2}Z_j^{\eta,n}(t)}_{\text{(A.4)}} \\ &\quad + \int_0^t \sum_i \tfrac{\eta R^*_{ji}(Q_i^{\eta,n}(s)-\vartheta_i)}{\mu_i^n}ds + \tfrac{1}{\sqrt{n}}L_j^{\eta,n}(t) - \tfrac{1}{\sqrt{n}}U_j^{\eta,n}(t), \end{aligned} \tag{151}$$

where the centered process $\boldsymbol{Z}^{\eta,n} := (Z_j^{\eta,n})$ has the expression

$$\begin{aligned} Z_j^{\eta,n}(t) := N&\left(\eta(\textstyle\int_0^t np_j(\boldsymbol{\tau}^{\eta,n}(s)) + (1-\eta)(\int_0^t np_j(\boldsymbol{\tau}^{n,*}(s)))\right) \\ &- \left(\eta\textstyle\int_0^t np_j(\boldsymbol{\tau}^{\eta,n}(s)) + (1-\eta)\int_0^t np_j(\boldsymbol{\tau}^{n,*}(s))ds\right) + \left(n\mu_j^n t - S_j^{\eta,n}(t)\right) \end{aligned} \tag{152}$$

We next analyze the terms labeled as (A.1)-(A.3) in (151).

(A.1) Our assumption of the initial value implies that (A.1) $\Rightarrow \boldsymbol{Y}^{\eta}(0)$.

(A.2) Since $j \in \mathcal{J}^{++}$, we have $\tau_j^{*,n} > 0$ for all sufficiently large n. Then by complementary slackness (27), we have $n(\mu_j^n - p_j(\boldsymbol{\tau}^{n,*})) = 0$. Then using Taylor expansion, we have

$$n^{1/2}\left(p_j(\boldsymbol{\tau}^{\eta,n}(s) - \mu_j^n\right) = n^{1/2}\left(p_j(\boldsymbol{\tau}^{\eta,n}(s) - p_j(\boldsymbol{\tau}^{n,*})\right) \to \sum_i \frac{R^*_{ji}(Q_i^{\kappa,n}(s)-\vartheta_i)}{\mu_i^n} \tag{153}$$

Thus (A.2) $\to 0$.

(A.3) Following the logic in the last bullet, (A.3)$= n^{1/2}(\mu_j^n - p_j(\boldsymbol{\tau}^{n,*})) = 0$.

(A.4) By functional central limit theorem (Chen and Yao, 2001), $n^{-1/2}\boldsymbol{Z}^{\eta,n} \Rightarrow \boldsymbol{\Sigma}^{1/2}\boldsymbol{B}(t)$ with $\boldsymbol{\Sigma}^{1/2}$ a diagonal matrix with $\Sigma_{jj}^{1/2} = \sqrt{(1+\omega_j^2)\mu_j}$.

It then follows that $\boldsymbol{Q}^{\eta,n} \Rightarrow \boldsymbol{Y}^{\eta}$. ■

Appendix Q: Proof of Theorem 7

Proof. Let $SW(\boldsymbol{\tau})$ denote the expected utility for a unitary customer to join DCPQ at state $\boldsymbol{\tau}$, with its expression given by (20). In the n^{th} DPQS, let $SW^{\eta,n}$ denote the expected utility for a unitary customer to joining the n^{th} η-informed DPQS at its steady state $\boldsymbol{Q}^{\eta,n}(\infty)$. If this customer observes the expected waiting times, her expected utility will be $\mathbb{E}[SW\left(\boldsymbol{\tau}^* + n^{-1/2}\boldsymbol{Q}^{\eta,n}(\infty)\circ\boldsymbol{\mu}^{-1}\right)]$; if the customer does not observe the expected waiting times, then she will use the equilibrium

waiting times $\boldsymbol{\tau}^{n,*}$ as her believe. In that case, the expected utility of the customer joining any queue j is the same as that of joining a queue with the equilibrium waiting time $\tau_j^{n,*}$ due to linearity of the utility function. Therefore, the expected utility of an uninformed customer equals $SW(\boldsymbol{\tau}^{n,*})$, the expected utility of an informed customer arriving at state $\boldsymbol{\tau}^{n,*}$. Therefore,

$$\begin{aligned}
&n(SW^{\eta,n} - SW(\boldsymbol{\tau}^{n,*}))\\
&= n\left(\eta\mathbb{E}[SW\left(\boldsymbol{\tau}^* + n^{-1/2}\boldsymbol{Q}^{\eta,n}(\infty)\circ\boldsymbol{\mu}^{-1}\right)] + (1-\eta)SW(\boldsymbol{\tau}^{n,*}) - SW(\boldsymbol{\tau}^{n,*})\right)\\
&= n\eta\left(\mathbb{E}[SW\left(\boldsymbol{\tau}^* + n^{-1/2}\boldsymbol{Q}^{\eta,n}(\infty)\circ\boldsymbol{\mu}^{-1}\right)] - SW(\boldsymbol{\tau}^* + n^{-1/2}\boldsymbol{\vartheta}\circ\boldsymbol{\mu}^{-1})\right)\\
&\to n\eta\left(\mathbb{E}[SW\left(\boldsymbol{\tau}^* + n^{-1/2}\boldsymbol{Y}^{\eta}(\infty)\circ\boldsymbol{\mu}^{-1}\right)] - SW(\mathbb{E}[\boldsymbol{\tau}^* + n^{-1/2}\boldsymbol{Y}^{\eta}(\infty)\circ\boldsymbol{\mu}^{-1}])\right)\\
&= \eta\int_{\boldsymbol{z}\in\mathbb{R}^J}\tfrac{1}{2}(\boldsymbol{z}\circ\boldsymbol{\mu}^{-1})^T\nabla^2 SW(\boldsymbol{\tau}^*)(\boldsymbol{z}\circ\boldsymbol{\mu}^{-1})f(\boldsymbol{z},0,\eta^{-1}\Sigma_\infty)\,d\boldsymbol{z} + o(1)\\
&\to \tfrac{1}{2}\left(\mathrm{Diag}\,(\boldsymbol{\mu})^{-1}\right)\nabla^2 SW(\boldsymbol{\tau}^*)\mathrm{Diag}\,(\boldsymbol{\mu})^{-1}\right)\cdot\Sigma_\infty := C,
\end{aligned} \tag{154}$$

where the last equality follows Taylor expansion, the $\cdot$ on the right-hand-side limit denotes the matrix inner product. The right-hand-side is a constant, showing that $SW^{\eta,n}$ actually stays invariant with $\eta\in(0,1]$. ∎

Appendix R: Proof of Corollary 4

The proof is mostly similar to that of Theorem 5, but differs in two places: (1) the derivative of $Q_j^n(t)$ includes an extra term $-dX_j(t)$, which represents 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 (47). We will prove the theorem by highlighting the parts due to the above differences.

We next prove that by restricting the process to stay inside the bounded domain $\Omega(\kappa)$, the bounded process $\{\boldsymbol{Q}^{\kappa,n}(t)|0\le t\le T\}$ weakly converges to $\{\boldsymbol{Y}^{\kappa}(t)|0\le t\le T\}$. The rest of the proof, including Lemma 3, follows the same routine as in the proof for Theorem 5 and we will not repeat them.

We first express $Q_j^{\kappa,n}(t)$ in a similar way to (35) as follows:

$$\begin{aligned}
Q_j^{\kappa,n}(t) &= Q_j^{\kappa,n}(0) + n^{-1/2}N\left(n\textstyle\int_0^t p_j(\boldsymbol{\tau}^{\kappa,n}(s))ds\right) - n^{-1/2}N\left(\textstyle\int_0^t dX_j^{\kappa,n}(s)ds\right) - n^{-1/2}S_j^{\kappa,n}(t)\\
&\quad n^{-1/2}L_j^{\kappa,n}(t) - n^{-1/2}U_j^{\kappa,n}(t)\\
&= \underbrace{Q_j^{\kappa,n}(0)}_{\text{(A.1)}} +\\
&\quad \underbrace{\int_0^t\left(n^{1/2}\left(p_j(\boldsymbol{\tau}^{\kappa,n}(s)) - \mu_j^n - n^{-1}dX_j^{\kappa,n}(s)\right) - ((\sum_i \frac{R_{ji}^*}{\exp(\tau_i^n d)\mu_i^n} - d)(Q_i^{\kappa,n}(s)-\vartheta_i) - \theta_j)\right)ds}_{\text{(A.2)}}\\
&\quad + \underbrace{n^{-1/2}Z_j^{\kappa,n}(t)}_{\text{(A.3)}} + \textstyle\int_0^t\left((\sum_i \frac{R_{ji}^*}{\exp(\tau_i^n d)\mu_i^n} - d)(Q_i^{\kappa,n}(s)-\vartheta_i) - \theta_j\right)ds + \frac{1}{\sqrt{n}}L_j^{\kappa,n}(t) - \frac{1}{\sqrt{n}}U_j^{\kappa,n}(t),
\end{aligned} \tag{155}$$

where the additional superscript κ represents that the corresponding process has a domain $\Omega(\kappa)$. Note that the centered process $\boldsymbol{Z}^{\kappa,n} := (Z_j^{\kappa,n})$ has included an extra term for the reneging customers, which has the expression

$$\begin{aligned}
Z_j^{\kappa,n}(t) := &\left(N(\textstyle\int_0^t np_j(\boldsymbol{\tau}^{\kappa,n}(s))) - \textstyle\int_0^t np_j(\boldsymbol{\tau}^{\kappa,n}(s))ds\right)\\
&+ \left(n\mu_j^n t - S_j^{\kappa,n}(t)\right) - \left(N(\textstyle\int_0^t dX_j^{\kappa,n}(s)ds) - \textstyle\int_0^t dX_j^{\kappa,n}(s)ds\right)
\end{aligned} \tag{156}$$

We next analyze the terms labeled as (A.1)-(A.3) in (155).

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

2. Since τ_j has to be computed using (47), the expression for $\Delta\boldsymbol{\tau}^{\kappa,n}$ will be

$$\begin{aligned}\Delta\boldsymbol{\tau}^{\kappa,n}(s) &:= \boldsymbol{\tau}^{\kappa,n}(s) - \boldsymbol{\tau}^* \\ &= \tfrac{1}{d}\log(1+(n^{1/2}\boldsymbol{Q}^{\kappa,n}(s)+\boldsymbol{X}^*)\circ(n\boldsymbol{\mu}^n)^{-1}) - \tfrac{1}{d}\log(1+\boldsymbol{X}^{n,*}\circ(n\boldsymbol{\mu}^n)^{-1}).\end{aligned} \tag{157}$$

It is not difficult to show that $n^{\frac{1}{2}}\|\Delta\boldsymbol{\tau}^{\kappa,n}\|_t$ is uniformly bounded and thus it suffices to expand the Taylor series of $n^{1/2}\left(p_j(\boldsymbol{\tau}^{n,*}+\Delta\boldsymbol{\tau}^{\kappa,n}(s))-p_j(\boldsymbol{\tau}^{n,*})\right)$ till its first-order term. Some basic algebra leads to

$$n^{1/2}\left(p_j(\boldsymbol{\tau}^{n,*}+\Delta\boldsymbol{\tau}^{\kappa,n}(s))-p_j(\boldsymbol{\tau}^{n,*})\right) \to \sum_i \frac{Q_i^{\kappa,n}(s)-\vartheta_i}{\exp(\tau_j^* d)\mu_i^n}R_{ji}^* \tag{158}$$

Thus, by our definition of θ_j and ϑ_j, we have

$$\begin{aligned}&n^{1/2}\left(p_j(\boldsymbol{\tau}^n(s))-\mu_j^n-n^{-1}d\boldsymbol{X}^{\kappa,n}(s)\right) \\ =\ &n^{1/2}\left(p_j(\boldsymbol{\tau}^{n,*}+\Delta\boldsymbol{\tau}^{\kappa,n})-p_j(\boldsymbol{\tau}^{n,*})\right)+n^{1/2}\left(p_j(\boldsymbol{\tau}^{n,*})-\mu_j^n-n^{-1}d\boldsymbol{X}^{n,*}\right) \\ &+n^{1/2}\left(n^{-1}d\boldsymbol{X}^{n,*}-n^{-1}d\boldsymbol{X}^*\right)-n^{1/2}\left(n^{-1}d\boldsymbol{X}^{\kappa,n}(s)-n^{-1}d\boldsymbol{X}^*\right) \\ \to\ &\textstyle\sum_i \frac{Q_i^{\kappa,n}(s)-\vartheta_i}{\exp(\tau_j^* d)\mu_i^n}R_{ji}^*-\theta_j+d\vartheta_j-dQ_j^{\kappa,n}(s)\end{aligned} \tag{159}$$

The above convergence leads to that (A.2)$\to 0$ 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 5 that the sum of the first two terms converges to $\boldsymbol{\Sigma}^{1/2}\boldsymbol{B}(t)$ with $\boldsymbol{\Sigma}^{1/2}$ a diagonal matrix and $\Sigma_{jj}^{1/2}=\sqrt{(1+\omega_j)^2\mu_j}$, respectively. Since $\frac{1}{n}\int_0^t dX_j^{\kappa,n}(s)ds \to \frac{1}{n}dX_j^*t=(\exp(\tau_j^*d)-1)\mu_j t$ uniformly on any compact set $t\in[0,T]$, and $\frac{1}{n}\int_0^t dX_j^{\kappa,n}(s)ds$ is a non-decreasing process in t, we may invoke the random time-change theorem and FCLT to prove that

$$n^{-1/2}\left(N(\int_0^t dX_j^{\kappa,n}(s)ds)-\int_0^t dX_j^{\kappa,n}(s)ds\right)\Rightarrow B_j^D(t). \tag{160}$$

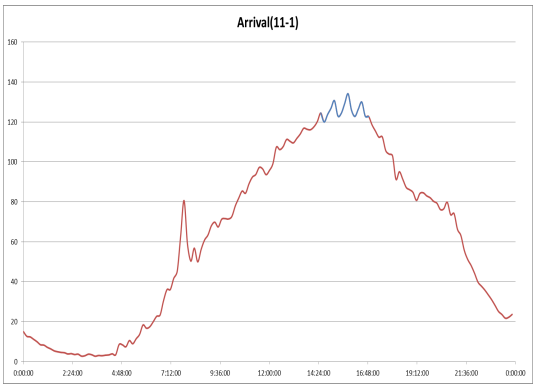
where $B_j^D(t)$ is a Brownian motion whose covariance matrix is a diagonal matrix and the j^{th} entry of its diagonal is given by $(\exp(\tau_j^*d)-1)\mu_j$. Since $n^{-1/2}\boldsymbol{Z}^{\kappa,n}(t)$ is the sum of three independent Brownian processes, we deduce that

$$n^{-1/2}\boldsymbol{Z}^{\kappa,n}(t)\Rightarrow\boldsymbol{\Sigma}^{R,1/2}\boldsymbol{B}(t). \tag{161}$$

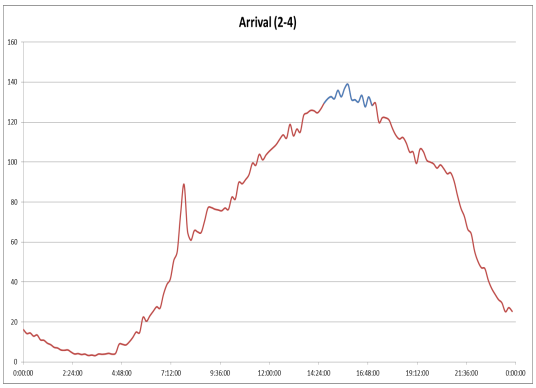
■

Appendix S: Supplementary Materials for Case Study

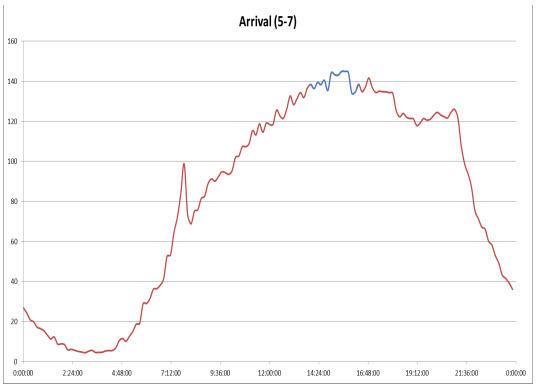
Figure 9 in plots The average total arrival rates $a_{pe}(t)+a_{pa}(t)$ for the two ports of entry on Tuesday/Wednesday/Thursday (T/W/T) in each season. It also highlights the selection of the peak hours.



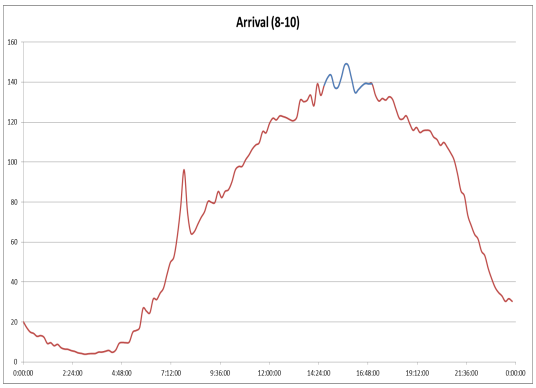
(a)Nov. 2017 - Jan. 2018, 14:30-17:00



(b)Feb. 2018 - Apr. 2018, 14:30-17:00



(c)May. 2018 - Jul. 2018, 14:00-16:30



(d)Aug. 2018 - Nov. 2018, 14:40-17:10

Figure 9 Plots of average total arrival rates on T/W/T in each season, with the peak hours marked in blue.