Manufacturing & Service Operations Management

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Strategic Queueing Behavior and Its Impact on System Performance in Service Systems with the Congestion-Based Staffing Policy

Pengfei Guo, Zhe George Zhang,

To cite this article:
Pengfei Guo, Zhe George Zhang, (2013) Strategic Queueing Behavior and Its Impact on System Performance in Service Systems with the Congestion-Based Staffing Policy. Manufacturing & Service Operations Management 15(1):118-131. http://dx.doi.org/10.1287/msom.1120.0406

M&SOM MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 15, No. 1, Winter 2013, pp. 118–131
ISSN 1523-4614 (print) | ISSN 1526-5498 (online)

informs

http://dx.doi.org/10.1287/msom.1120.0406

Strategic Queueing Behavior and Its Impact on System Performance in Service Systems with the Congestion-Based Staffing Policy

Pengfei Guo
Faculty of Business, Hong Kong Polytechnic University, Hung Hom, Hong Kong, pengfei.guo@polyu.edu.hk

Zhe George Zhang
Department of Decision Sciences, Western Washington University, Bellingham, Washington 98225; and
Beedie School of Business, Simon Fraser University, Burnaby, British Columbia V5A 1S6, Canada, george.zhang@wwu.edu

We study strategic customer behavior in a multiserver stochastic service system with a congestion-based staffing (CBS) policy. With the CBS policy, the number of working servers is dynamically adjusted according to the queue length. Besides lining up for free service, customers have the option of paying a fee and getting faster service. Customers' equilibrium behavior is studied under two information scenarios: In the no information scenario, customers only know the long-term statistics, such as the expected waiting time; in the partial information scenario, customers observe the number of working servers and understand the staffing policy upon their arrival. Unlike a queueing system with a constant staffing level, a positive externality is associated with customers' joining the CBS system. Both avoid-the-crowd and follow-the-crowd customer behaviors are possible, and multiple equilibria could exist. We develop the stationary performance measures of the system by considering the customers' strategic behavior. Numerical analysis shows that information can either hurt or improve the performance of the system, depending on the staffing and pricing policy. Another important conclusion is that the system performance is more robust to setting a relatively high than a relatively low price.

Key words: congestion-based staffing; pricing; delay information; strategic customer
History: Received: September 27, 2010; accepted: May 15, 2012. Published online in *Articles in Advance* October 5, 2012.

1. Introduction

In recent years, travelers have experienced extensive lineups at the U.S. and Canada border-crossing stations, especially during weekends and holidays. Sometimes, major crossings such as the Peace Arch or the Pacific Highway Port of Entry between Washington State and the Province of British Columbia can delay customers for more than three hours (see http://www.borderlineups.com). The congestion problem is becoming worse for many reasons, such as high security levels, volatile exchange rates, increased trade, popular sport games, and fast population growth.

In such a system, the number of servers (open inspection booths) is dynamically adjusted according to the queue length. If the queue reaches an upper threshold, additional servers are opened; if the queue drops down to a lower threshold and some servers become idle, these idle servers are closed, and the queue starts to build up again. By choosing an appropriate upper threshold, the actual queue length can be effectively controlled in a desired range with a high probability. Such a staffing policy is called *congestion-based staffing* (CBS) as defined by Zhang (2009).

Besides the benefit of controlling the queue length in a desired range, the CBS policy allows system managers to balance customers' waiting costs and system operating costs. Compared with a fixed staffing level policy, the CBS policy can significantly improve server utilization. The CBS has also been found advantageous in other service systems, such as supermarket checkouts and airport security checkpoints, where employees have multiple skills and can freely switch between queueing and nonqueueing jobs.

Nevertheless, the arrival rate in a border-crossing system is highly time varying. The system can be heavily loaded during busy times such as weekends and holidays, even with all inspection booths opened under the CBS policy (http://www.cascadegatewaydata.com). Because of the infrastructure constraint, the maximum number of inspection booths is fixed. Consequently, the huge congestion cannot be relieved by simply increasing the staffing level in peak hours. On the other hand, currently, major border-crossing stations have the fast pass lanes (also called "Nexus lanes") for Nexus permit holders, which are underutilized. How to make better use of the almost empty Nexus lane to reduce the congestion of the regular lanes is one of the main motivations

for developing our model in this paper. We consider a refined CBS system with a fast pass service option (probably directly converted from the Nexus lanes) at a price.

The fast lane option has been adopted by airline companies: In December 2008 United Airlines launched a "Premier Line" option, which allows ordinary customers to purchase a fast pass service comprising priority check-in, security clearance, and boarding at a cost of $25 (Airline Industry Information 2008). Adding a price lane may require advanced high-tech facilities or equipment such as radio-frequency identification (RFID) chip readers, high-resolution surveillance cameras, or even X-ray detectors to perform more effective and speedy inspections. Since April 2011, U.S. Customs and Border Protection (CBP) has had "ready lanes" at many border-crossing stations that are used for vehicles equipped with RFID-enabled cards. The cards allow CBP officers to screen travelers faster, and it has been reported that this technique can save 10 to 15 seconds per car, which is approximately 20% of the average inspection time (Associated Press 2011). Such a faster service rate combined by a relatively low arrival rate (controlled by the permit requirement or a toll price as discussed later) can make the queue length very short or close to zero. The current implementation of these fast lanes at border crossings indicates the maintenance of the same level of security screening as in regular lanes. For a detailed discussion of the costs and risks of using different types of inspection equipment for security reasons, see Wein et al. (2006).

The goal of this study is to examine the customers' decentralized queueing behavior for either free service or fast service and its impact on system performance. We consider scenarios of *no* and *partial* information. Under the no information scenario, customers choose service based on the long-term statistics, such as expected waiting time. Under the partial information case, customers know server status (the number of working servers) on arrival and understand the CBS policy. In a border-crossing system, congestion information and server status can be provided by a large electronic board placed at the divergence of the price and free lanes. Unlike a queue with a fixed staffing level where a customer joins and increases the waiting cost of other customers (a *negative externality*), when a customer joins a CBS system, the action may benefit other customers (a *positive externality*). With a CBS policy, the system switches between a low (mode A) and a high (mode B) staffing level. A joining customer can either facilitate the system to switch from mode A to mode B, thus reducing the expected waiting time of those ahead of him or her, or keep the system in mode B for a longer period, reducing the expected waiting time of those after him or her. The positive externality makes the analysis of customer equilibrium behavior intriguing.

In the no information case, we observe that the expected waiting time can first increase, then decrease, and finally increase again with the system utilization level. In the range of the decreasing part, a customer's tendency to join a specific lane increases with other customers' tendencies. This is called *follow-the-crowd* (FTC) behavior (see Hassin and Haviv 2003), in contrast to *avoid-the-crowd* (ATC) behavior in regular queues with fixed staffing levels. FTC behavior is also observed in a partial information case with mode A information disclosed to customers. We show that in both information scenarios, except some extreme cases, there could exist as many as three equilibria for strategic customers. We discuss the stability of these equilibria.

In the partial information case, we observe a paradox for customers' behavior when the upper threshold of the staffing policy is small and the price is in a certain range: When all servers are open, all join the price system, and when some servers are closed, all customers join the free system. Such a behavior, caused by customers' self-interest choice, is socially undesirable and can result in frequent mode changeover, incurring a high setup cost. Therefore, in a service system with positive externality, managers should be more *cautious* about providing congestion information. Another important finding is that system performance is more robust when the price is set relatively high. First, when the price is reduced to slightly below optimal, the equilibrium arrival rate to the free system can decrease sharply, and the service cost increases dramatically. Second, when the price is set relatively high, the total cost becomes less sensitive to the upper threshold of the CBS policy, which facilitates the implementation of the policy in practice. Therefore, given the inherent uncertainty in border-crossing systems and the imperfections of modeling, systems should err on the side of overpricing, at least initially, until they gain a better understanding of customer reactions.

The remainder of this paper is organized as follows. Section 2 provides the literature review. Section 3 presents the notation, assumptions, and model formulation. Section 4 focuses on the equilibrium analysis of customer choice behavior. Section 5 discusses the practical implications of our study for border-crossing systems. Section 6 concludes the paper with a summary and a discussion on limitations and future research directions. All proofs are relegated to the online appendices (available at http://dx.doi.org/10.1287/msom.1120.0406).

2. Literature Review

Our work is related to studies in multiple fields. First, the CBS model belongs to the class of multiserver

vacation models that address the performance effects of changing service capacity based on the congestion level of the system. For a survey of this area, see Tian and Zhang (2006). Recently, Zhang (2009) presented a performance analysis of CBS queueing systems, assuming that there is no price lane option and the arrival rate to the free lane is constant. In contrast, we consider a system with a price lane option and different levels of congestion information. Customers in our model strategically choose between a free or price lane upon arrival. The arrival rates to the CBS queueing system are *equilibrium* rates, which are determined by price and information.

Our work is also related to studies on modeling port/border security and congestion. Wein et al. (2006) studied a multilayered port security system and numerically evaluated the best inspection strategy for the CBP given a certain level of security requirement. Bakshi and Gans (2010) formed a game-theoretic framework to model the strategic interaction between the CBP, trading firms, and terrorists. A particularly interesting conclusion is that the CBP may have the opportunity to use strategic delay to motivate firms to join an incentive program, named the Customs-Trade Partnership Against Terrorism. Zhang et al. (2011) modeled a two-tier security checking system with consideration of balancing both security and customer service goals. They determined the optimal further inspection proportion to achieve the balance of the two goals.

The third stream of related work is queueing models with strategic customer behavior. Naor (1969) first studied the use of an admission fee to induce the socially optimal arrival rate in a single-server system with an observable queue. Edelson and Hildebrand (1975) considered an unobservable queue case. More related work can be found in a survey book by Hassin and Haviv (2003) and a survey paper by Akşin et al. (2007). Guo and Hassin (2011) examined an $M/M/1$ system with an N-policy in which the server is triggered to work only after the queue length reaches threshold N. By assuming that customers react in a strategic way with knowledge of the N-policy, they found that socially optimal arrival rates may be larger than those determined by self-interested customers. Dimitrakopoulos and Burnetas (2011) studied an unobservable $M/M/1$ queue with the service rate switching between a low and a high value. They proved that there could exist at most three equilibria and showed that the socially optimal arrival rate lies between the extreme equilibrium values in cases with multiple equilibria. Our setting of the multiserver service system with the CBS policy is more complex than single-server models with the threshold policy. In addition, we consider the optimal pricing issue. Because of the coexistence of positive and negative externalities, many of our results, such as the monotonicity of the optimal CBS threshold and the impact of price and information, are significantly different from those in past studies.

A feature of our model is that a joining customer can bring positive externality to the system. There are some related studies in this subfield. Afimeimounga et al. (2005) considered a transportation system with two parallel routes. On one route, service slows as traffic increases, and on the other, service frequency increases with demand. They showed that multiple user equilibria may exist and discussed their stability. Burnetas and Economou (2007) studied a system with an exponential setup time for each busy period. By considering customers' strategic behavior under different information levels about queue length and/or server status, they observed that both ATC and FTC customer behaviors could occur. Haviv and Kerner (2007) treated an $M/G/1$ queue with the observable server's status. In the case of busy server, customers can see whether the queue is empty or not. They showed that when the queue is empty, both ATC and FTC behaviors are possible, depending on the service distribution. Veeraraghavan and Debo (2009) considered the situation in which the queue length indicates not only congestion level, but also service quality. They showed that when the service rates are negatively (positively) correlated with unknown service values, customers prefer to join longer (shorter) queues. Johari and Kumar (2008) studied an online gaming system in which users form a club. Because both negative and positive externalities exist, they found that the club size determined by self-interested users is always smaller than that chosen by the service manager. Economou and Manou (2011) investigated customers' balking strategies in a clearing system with two alternating service rates.

Our study is also related to the literature on the impact of delay announcements or congestion information in different queueing systems (see Hassin 1986; Whitt 1999; Armony and Maglaras 2004a, b; Guo and Zipkin 2007; Allon et al. 2011; Ibrahim and Whitt 2009a, b; Armony et al. 2009). A common feature of these studies is that a customer arrival increases the delay for other customers and thus brings only negative externality to the system.

The last area of related work is on staffing issues in multiple-server service systems, particularly call centers. References in this area and a high level overview of utilizing queueing models to set staffing requirements were provided by Whitt (2007).

3. System Description and Model Formulation

Customers arrive at a service system according to a Poisson process with rate Λ. Upon arrival, a customer

Table 1 List of Symbols

System parameters	
Λ	Total arrival rate
μ	Service rate
c	Total number of servers
n	Lower threshold of queue length; once the queue reduces to this level and the staffing level is high, $c-n$ idle servers are shut down and become inactive
N	Upper threshold of queue length; once the queue reaches this number and the staffing level is low, $c-n$ inactive servers are reopened
θ	Customers' delay-sensitivity parameter, indicating the importance of time
S	The switchover cost between CBS modes
K_0, a	Cost parameters of fast service
τ	Price for using the fast service
Derived parameters	
λ	Effective arrival rate to the free system
W	Expected waiting time of the free system
L	Expected number of customers in the free system

must choose between joining a multiserver queue for free service or paying a price τ for fast service. Because we focus on the analysis of the congested free system, we call customers choosing the free service "joining customers" and those choosing the fast service "balking customers." The system can then be considered a multiserver queue with balking customers. Throughout this paper, we assume that fast service customers do not have to wait. This assumption is reasonable if the delay in the price lane is much smaller than that in the free lane. In practical border-crossing stations with Nexus lanes, more than 95% of the time, the waiting time for fast lanes is minimal or almost zero (see http://www.wcog.org/Border.aspx). In other words, only the free lane has a congestion problem. We verify the robustness of this assumption by simulations in a later section. Next, we describe the CBS policy for the system and then consider the customer's decision as a queueing game. The basic symbols used in the paper are listed in Table 1.

3.1. Congestion-Based Staffing Policy

The free system is operated under a CBS policy with an A/B mode (see Zhang 2009). There are c homogeneous servers offering exponentially distributed service with rate μ. Whenever $c-n$ servers become idle (or the number of customers in the system is reduced to n), they are shut down, and the system enters mode A. These $c-n$ servers are reopened when the queue length exceeds an upper threshold N, and the system is switched to mode B. Note that any server can be shut down, depending on which becomes idle first. The upper and lower thresholds allow managers to control the queueing dynamics. Besides the cost consideration, there are other factors in determining the lower threshold n, such as the number of available agents from back-office jobs. Because of this complexity, we assume that it is exogenously given and focus on the decision on the upper threshold level. As shown by Zhang (2009), the average queue length and frequency of mode switching can be effectively controlled by this upper threshold N.

The system can be modeled with a two-dimensional state variable $\{\mathbf{L}, \mathbf{J}\}$, where $\mathbf{L}$ is the number of customers in the system, and $\mathbf{J}=0$ for mode A or 1 for mode B.

Because of the exponential interarrival and service times, we have a Markov process with the state space

$$\Omega = \{(k,0)\colon 0 \le k \le n\} \cup \{(k,j)\colon n < k \le N-1,\ j=0,1\} \cup \{(k,1)\colon k \ge N\}.$$

Figure 1 shows the transition diagram for the CBS system with a constant arrival rate λ. The fast service option and customer choice behavior guarantee the stability of the system.

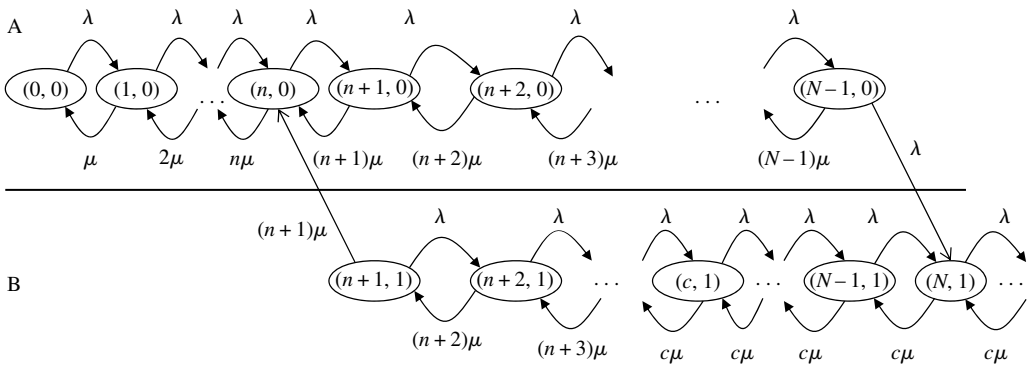
The stationary distribution is denoted by

$$p_k = P\{\mathbf{L}=k, \mathbf{J}=0\}, \quad 0 \le k \le N-1;$$
$$q_k = P\{\mathbf{L}=k, \mathbf{J}=1\}, \quad k \ge n+1.$$

3.2. Queueing Strategies

In the partial information case, customers are informed of the CBS policy and the number of working

Figure 1 State Transition Diagram with Partition (A and B Regions) of State Space



servers (mode A or B) on arrival. However, they do not know the real-time queue length or waiting time. In fact, this scenario is quite close to the practical situation of a border-crossing station. The waiting time information posted on the electronic board is most likely to be updated when the number of open inspection booths change rather than being continuously updated with the real-time queue length. Therefore, the conditional expected waiting time given the number of open booths is the most likely information disclosed to the customers.

We assume that the service reward for a customer is infinite; that is, every customer must go through the system, either by a free or a price lane. Customers are risk neutral and experience a delay cost θ per unit of time.

Customers are rational and individual optimizers. Upon arrival, a customer makes the joining decision by comparing the expected waiting cost of the free system with the price of the price system. It is assumed that customers' decisions are irrevocable and the system has reached the steady state. The customers are treated as indistinguishable players in a symmetric game. We consider a symmetric profile of customer strategies, which is denoted by a pair of joining probabilities $\boldsymbol{\alpha} = (\alpha_A, \alpha_B)$, where α_A is the probability of joining for a customer who observes mode A, and α_B is the probability of joining for mode B. Let $U(\boldsymbol{\beta}, \boldsymbol{\alpha})$ denote the payoff for a tagged customer who follows strategy $\boldsymbol{\beta}$ while all others follow strategy $\boldsymbol{\alpha}$. A strategy profile $\boldsymbol{\alpha}^e$ is a symmetric Nash equilibrium strategy if

$$U(\boldsymbol{\alpha}^e, \boldsymbol{\alpha}^e) = \max\{U(\boldsymbol{\beta}, \boldsymbol{\alpha}^e), \boldsymbol{\beta} \in [0,1] \times [0,1]\};$$

that is, a symmetric equilibrium is a symmetric strategy profile such that it is a best response against itself.

In the no information case, the servers' status is not available and a customer's strategy is represented by a single joining probability. The foregoing definition of equilibrium can still be adopted by positing a constraint that the two elements of the strategy vector are equal.

It is known that there is at most one equilibrium in the ATC situation, whereas multiple equilibria can exist in the FTC situation (see, e.g., Hassin and Haviv 2003). In the case of multiple equilibria existing, we adopt two concepts related to stability to rule out some of them.

The first concept refers to the stability subject to small disruptions in the neighborhood of an equilibrium, which is called local stability.

Definition 1. An equilibrium strategy $\boldsymbol{\alpha}$ is said to be (locally) stable if there exists an $\varepsilon > 0$ such that for all strategy $\boldsymbol{\beta}$ subject to $\|\boldsymbol{\beta} - \boldsymbol{\alpha}\| \le \varepsilon$, where $\|\mathbf{x}\| = \sqrt{x_1^2 + x_2^2}$ for a vector $\mathbf{x} = (x_1, x_2)$, the following condition holds:

$$U(\boldsymbol{\alpha}, \boldsymbol{\beta}) > U(\boldsymbol{\beta}, \boldsymbol{\beta}).$$

The second concept concerns the comparison of two equilibria, which is known as an *evolutionarily stable strategy* (ESS) (see Hassin and Haviv 2003, Haviv and Kerner 2007, Burnetas and Economou 2007).

Definition 2. An equilibrium strategy $\boldsymbol{\alpha}$ is said to be an ESS if it satisfies the condition that $U(\boldsymbol{\alpha}, \boldsymbol{\beta}) > U(\boldsymbol{\beta}, \boldsymbol{\beta})$ for all $\boldsymbol{\beta} \neq \boldsymbol{\alpha}$, which is a best response against $\boldsymbol{\alpha}$.

The best response toward an equilibrium strategy need not be unique. Consequently, for a (symmetric) equilibrium strategy $\boldsymbol{\alpha}$, there could exist another best response $\boldsymbol{\beta} \neq \boldsymbol{\alpha}$. Therefore, starting with equilibrium $\boldsymbol{\alpha}$, some customers could deviate by adopting strategy $\boldsymbol{\beta}$. If $\boldsymbol{\beta}$ is *strictly* a better response against itself than $\boldsymbol{\alpha}$ is, eventually all customers could migrate to and be "trapped" in the new equilibrium $\boldsymbol{\beta}$. In this sense, the equilibrium $\boldsymbol{\alpha}$ is not stable. If no such $\boldsymbol{\beta}$ exists, $\boldsymbol{\alpha}$ is an ESS.

The ESS is useful for ruling out the mixed equilibria in the existence of pure and mixed equilibria: In a mixed equilibrium, customers are indifferent between joining and balking, whereas in the pure equilibrium "always join," customers prefer joining to balking. However ESS criteria fail in comparing two mixed equilibria because customers have the same payoff in both equilibria. Therefore, in comparing two mixed equilibria, we introduce a risk-dominance concept, similar to the definition for a symmetric two-player game in Harsanyi and Selten (1988).

Definition 3. Suppose that both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are equilibrium strategies. Define $\widehat{\boldsymbol{\alpha}\boldsymbol{\beta}}(\varepsilon) = (1-\varepsilon)\boldsymbol{\alpha} + \varepsilon\boldsymbol{\beta}$ to represent the situation in which a majority of players, $1-\varepsilon$, choose $\boldsymbol{\alpha}$, and the remaining smaller fraction of players, ε, choose $\boldsymbol{\beta}$. Similarly, define $\widehat{\boldsymbol{\beta}\boldsymbol{\alpha}}(\varepsilon) = (1-\varepsilon)\boldsymbol{\beta} + \varepsilon\boldsymbol{\alpha}$ to represent the situation in which $1-\varepsilon$ players choose $\boldsymbol{\beta}$, whereas ε players choose $\boldsymbol{\alpha}$. Equilibrium $\boldsymbol{\alpha}$ risk-dominates equilibrium $\boldsymbol{\beta}$ if there exists $0 < \epsilon < 1$ such that, for all $0 \le \varepsilon \le \epsilon$,

$$U(\boldsymbol{\alpha}, \widehat{\boldsymbol{\alpha}\boldsymbol{\beta}}(\varepsilon)) > U(\boldsymbol{\beta}, \widehat{\boldsymbol{\beta}\boldsymbol{\alpha}}(\varepsilon)). \tag{1}$$

In practice, customers are likely to have some noisy information on other customers' strategies. When customers do not have complete information about which strategy the opponent will pick, they assign probabilities for each strategy. One equilibrium risk-dominates the other if the expected payoff from adopting the former strategy exceeds the expected payoff from adopting the latter.

3.3. Social Welfare

In regular queues with only negative joining externality, customers' decentralized queueing behavior often results in overcongestion of the system. Therefore, a price can be used to adjust the arrival stream to achieve social optimality. This method is broadly used in transportation systems.

In our model, the maximization of social welfare is equivalent to the minimization of the total expected cost, denoted as TC, which is the sum of the expected customer waiting cost and service provider's operating cost.

The total expected cost is the sum of four terms:

$$TC = E(\text{waiting cost}) + E(\text{mode switching cost}) + E(\text{fast service cost}) + E(\text{CBS staffing cost}).$$

The first term, the customer expected waiting cost, can be expressed as θL, where L is the expected number of customers in the free system. If the switchover cost between CBS modes is denoted by S, the second term can be obtained by the product of S and the switching frequency. To calculate the latter, we consider the queue length dynamics as a regenerative process alternating between the two modes. The changeover cycle is defined as the time interval between the two consecutive regeneration points, denoted by T. Define T_B as the expected first passage time from state $(N,1)$ to state $(n,0)$, and T_A as the expected first passage time from state $(n,0)$ to state $(N,1)$ as shown in Figure 1. Then,

$$E(T) = T_B + T_A.$$

Recursive formulas for computing T_B in an $M/M/c$ queue have been presented in the literature (see Omahen and Marathe 1978). To calculate T_A, one can consider a continuous-time Markov chain (CTMC) process in mode A with absorbing state $(N,1)$ (see Proposition EC.2 of Zhang 2009). Therefore, the switchover cost can be expressed as $S/E[T]$. The third term is the fast service cost. To maintain the required security level and service quality for the fast service lane, some extra staff and equipment such as X-ray detectors/scanners are needed for opening additional high-tech inspection booths when traffic demand is high. Thus, the fast service cost (denoted by C) is assumed to be increasing in the arrival rate of fast service customers. In this study, we use a power cost function as follows:

$$C(\Lambda-\lambda) = K_0((\Lambda-\lambda)/(c\mu))^a, \quad K_0 > 0,\ a > 0.$$

If $a > 1$, the above function is convex, which implies an increasing marginal cost for serving fast pass customers. Note that we rescale the arrival rate to the toll lane $(\Lambda-\lambda)$ by dividing it with $c\mu$. The fourth term is the CBS staffing cost. Although CBS inspectors are full-time multitask employees, putting more inspectors to work in inspection booths implies that there is a need to hire more employees to do other jobs. We assume that this cost is proportional to the average number of busy servers.

For a given price τ and a threshold N, based on the renewal reward theorem, the average cost function can be written as

$$TC = \theta L + \frac{S}{E[T]} + P(A)C(\Lambda-\lambda_A) + (1-P(A))C(\Lambda-\lambda_B) + [P(A)n + (1-P(A))c]C_o,$$

where $P(A) = \sum_{i=0}^{N-1} p_i$ is the probability that the system is in mode A; λ_A (λ_B) is the equilibrium arrival rate in mode A (B), depending on τ and the information scenario; and C_o is the unit staffing cost. There is a main trade-off between the first two terms because a large N increases customers' expected waiting time but also reduces the frequency of mode switching.

4. Equilibrium Analysis

In this section, we study customers' equilibrium choice behaviors under two information scenarios.

4.1. Partial Information Scenario

If all customers follow the mixed strategy (α_A, α_B), then $\lambda_A = \Lambda\alpha_A$, and $\lambda_B = \Lambda\alpha_B$. For convenience, instead of obtaining equilibrium joining probabilities α_A^e and α_B^e, we directly derive the equilibrium arrival rates λ_A^e and λ_B^e.

According to the PASTA (Poisson Arrivals See Time Averages) property, the expected waiting times seen by a customer upon arrival in modes A and B equal the conditional expected waiting times in mode A and mode B, respectively.

Let W_A (W_B) be the conditional expected waiting time in mode A (B). The transition diagram should be similar to Figure 1 except that λ is replaced with λ_A in mode A and replaced with λ_B in mode B. Based on this, the balance equations can be obtained (see Online Appendix A for details). The conditional distribution for queue length i on mode A or B can be written as

$$p_i(A) = \frac{p_i}{\sum_{k=0}^{N-1} p_k}, \quad q_i(B) = \frac{q_i}{\sum_{k=n+1}^{\infty} q_k}.$$

The closed-form expressions of $p_i(A)$ and $q_i(B)$ are listed in Lemma 5 in Online Appendix A.

In mode B, the conditional expected waiting time can be expressed as

$$W_B = \sum_{i=n}^{\infty} q_i(B)\left(\frac{\max[0, i-c+1]}{c\mu} + \frac{1}{\mu}\right).$$

Note that W_B is independent of λ_A, and we can write it as $W_B(\lambda_B)$.

With all servers being busy, the queue with a larger arrival rate is stochastically longer. It follows that $W_B(\lambda_B)$ is increasing in λ_B. This monotonicity allows us to examine how a particular customer or tagged customer responds to other customers' decisions. When other customers have a higher tendency to join the queue in mode B, λ_B is larger and so is W_B. This effect will reduce the joining tendency of a tagged customer. Therefore, the ATC behavior occurs in mode B. As the shortest expected waiting time is $W_B(0)$, the necessary condition for a customer to join the queue is $\tau > \theta W_B(0)$.

Proposition 1. *In mode B, there exists a unique equilibrium arrival rate* λ_B^e.

1. *If* $\tau \geq \theta W_B(\Lambda)$, *joining is the dominating strategy, and* $\lambda_B^e = \Lambda$.

2. *If* $\theta W_B(0) < \tau < \theta W_B(\Lambda)$, *customers adopt a mixed joining strategy, and the equilibrium arrival rate* λ_B^e *solves*

$$\tau = \theta W_B(\lambda_B).$$

3. *If* $\tau \leq \theta W_B(0)$, *balking is the dominating strategy, and* $\lambda_B^e = 0$.

For W_A, we use a recursive algorithm as in Proposition 4 of Ata and Shneorson (2006) to calculate the state-dependent expected waiting time for each state, which depends not only on the number of customers waiting before the arriving customer, but also the time to change mode. To capture this characteristic, for each mode A state encountered by a customer, we define a CTMC with absorption states. Denote the state of the CTMC by (i, j), where i represents the number of customers in front of a tagged customer and j the total queue length. If a customer arrives in mode A with i, then the initial state will be $(i, i+1)$. During the waiting time of this tagged customer, the system may change to mode B, and thus we denote the mode B state by (i, B). Following a first-step analysis, we obtain a recursion equation for the transient states (i, j) with $n \leq i \leq N-2$, $i+1 \leq j \leq N-1$:

$$W(i,j) = \frac{1}{\lambda_A + n\mu} + \frac{\lambda_A}{\lambda_A + n\mu} W(i, j+1) + \frac{n\mu}{\lambda_A + n\mu} W(i-1, j-1).$$

For the absorption states and transient mode B states, we have

$$W(n-1, j) = 1/\mu \quad \text{for } j \leq N-2$$

and

$$W(i, B) = \begin{cases} 1/\mu & \text{for } n \leq i \leq c-1, \\ \dfrac{i-c+1}{c\mu} + \dfrac{1}{\mu} & \text{for } i \geq c. \end{cases}$$

See Online Appendix B for details of calculating $W(i, j)$.

The expected waiting time for an incoming customer who observes i customers in the system is $W(i, i+1)$. Therefore, the conditional expected waiting times on mode A are

$$W_A = \sum_{i=n}^{N-1} p_i(A) W(i, i+1). \tag{2}$$

From Lemma 5 in Online Appendix A, $p_i(A)$ is a function of λ_A only. In addition, $W(i, i+1)$ is independent of λ_B. Therefore, W_A is a function of λ_A only, and we can also write it as $W_A(\lambda_A)$. Clearly, it is bounded and approaches $1/\mu$ if the arrival rate is closer to zero. Intuitively, W_A need not be monotonic with λ_A because although a large λ_A builds up the queue length, it also reduces the time to reach mode B. Hence, it is possible that $W_A(\lambda_A)$ reaches its local maximum when λ_A is moderate, which implies that multiple solutions to the equation $\tau = \theta W_A(\lambda_A)$ may exist.

For the extreme case with $N \leq c$, it is possible that the number of equilibrium solutions is not bounded or even countable. Consider a simple example with $c = 2$, $n = 1$, and $N = 2$: there are two servers, and a server is turned off if it becomes idle and turned on again when there are two customers in the system. In this case, with partial information, mode A means that at least one server is idle. Hence, whenever a customer arrives, it is served immediately and $W_A = 1/\mu$, which is independent of λ_A. Therefore, if $\tau = \theta/\mu$, a customer is always indifferent between joining and balking, and any $\alpha_A \in [0, 1]$ is an equilibrium joining probability in mode A. In this case, there exists a continuum of equilibria.

For more general cases with $N > c$, we observe at most three equilibria. Because of complexity of the performance measures, we seek the analytical results for two systems, one with two servers and a CBS policy of $(n = 1, N = 3)$, and the other with four servers and a CBS policy of $(n = 1, N = 4)$. As illustrated in those examples, W_A can be a strictly unimodal function of λ_A, implying that both very light and very heavy traffic flows could be beneficial for a customer arriving in mode A.

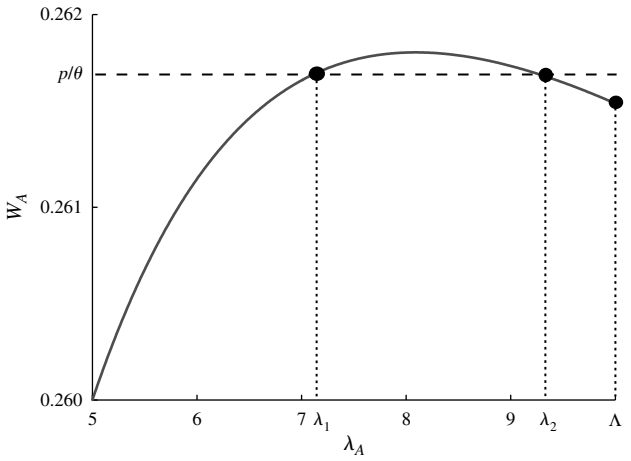
Lemma 1. *For a system with* $c = 2$, $n = 1$, *and* $N = 3$, *under partial information,* $W_A(\lambda_A)$ *is a strictly unimodal function of* λ_A *and reaches a maximum value* $(11\sqrt{5}+25)/((8\sqrt{5}+20)\mu)$ *at*

$$\lambda_A^* = \frac{\sqrt{5}+1}{2}\mu.$$

In addition, $W_A(0) = 1/\mu$, *and* $\lim_{\lambda_A \to \infty} W_A(\lambda_A) = 5/(4\mu)$.

Therefore, when $(5/4)(\theta/\mu) < \tau < ((11\sqrt{5}+25)/(8\sqrt{5}+20))(\theta/\mu)$, there exist two solutions, λ_1 and λ_2,

Figure 2 **Equilibrium Analysis Under Mode A Information Assuming $c = 2$, $n = 1$, $N = 3$, and $\mu = 5$**



which satisfy $0 < \lambda_1 < \lambda_A^* < \lambda_2$, to the following equation

$$\tau = \theta W_A(\lambda_A). \tag{3}$$

When $\Lambda > \lambda_2$, there exist three equilibria $\lambda_A^e = \lambda_1, \lambda_2$, and Λ, as shown in Figure 2. The figure shows that three equilibrium points exist because the line of $\tau/\theta = 0.2617$ crosses $W_A(\lambda_A)$ twice. The full list of equilibria is included in Proposition 2 in Online Appendix D.

The decreasing pattern of W_A in the range of $\lambda_A \geq \lambda_A^*$ implies that the more customers join the system, the smaller the expected waiting time for a tagged customer arriving in mode A. Hence, the FTC customer behavior exists here, which explains the existence of multiple equilibria.

Remark 1. Between the two mixed-strategy equilibrium arrival rates λ_1 and λ_2, λ_2 is *unstable* because a slight increase in the arrival rate will decrease W_A and attract more arrivals, converging to Λ. Both λ_1 and Λ are locally stable because a slight disturbance will not cause the equilibrium to drift away. Now we examine λ_1 and Λ. Let $\boldsymbol{\beta}$ be a strategy with joining probability 1 in mode A (corresponding equilibrium rate Λ) and $\boldsymbol{\alpha}$ be a mixed strategy with joining probability λ_1/Λ in mode A (corresponding equilibrium rate λ_1). Assume that $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ have the same joining probability in mode B. Clearly, $\boldsymbol{\beta}$ is a best response against $\boldsymbol{\alpha}$ because a customer is indifferent between joining and balking in mode A when all others choose $\boldsymbol{\alpha}$. However, when all others choose joining in mode A, it is better for an incoming customer to choose joining in mode A as well because of the condition $\theta W_A(\Lambda) < \tau$. Consequently, $\boldsymbol{\alpha}$ is not ESS according to Definition 2.

Remark 2. In this example, note that as $\lambda_B \to 0$, the conditional distribution in mode B becomes a uniform distribution on states $(2, 1)$ and $(3, 1)$. Hence, $W_B(0) = (1/2)(3/(2\mu)) + (1/2)(2/\mu) = (7/(4\mu)) > W_A(\lambda_A^*)$. Therefore, W_A is always smaller than W_B. If $\theta W_A(\Lambda) \leq \tau \leq \theta W_B(0)$, the strategic behavior is interesting under partial information: the equilibrium arrival rate in mode B is 0, and in mode A is Λ. This means that all customers join the free system in mode A. However, once the system reaches mode B, none join the free system because all join the price system. This behavior is paradoxical: when the system runs at full capacity, nobody joins, but when the system runs in a partial capacity, everybody joins. Consequently, the queue will quickly build up in mode A to reach N; once it reaches N, it will quickly be reduced because of all c working servers and no arrivals. Such a behavior results in a high frequency of mode switching and a periodically large rate of arrival in the price system, which can significantly hurt system performance. Similar results can be obtained in the four-server case.

Lemma 2. *When $c = 4$, $n = 1$, and $N = 4$, under partial information, $W_A(\lambda_A)$ is a strictly unimodal function of λ_A.*

4.2. No Information Scenario

In the no information case, there exists an equilibrium arrival rate λ. The stationary distribution of the system states can be obtained from the balance equations in Online Appendix A by using $\lambda_A = \lambda_B = \lambda$.

In equilibrium, the customer choice behavior yields the following equation:

$$\tau = \theta W(\lambda), \tag{4}$$

where $W(\lambda)$ is the expected waiting time for the free service system.

Note that $W(\lambda)$ can be expressed as a weighted average of W_A and W_B, i.e., $W(\lambda) = P(A)W_A(\lambda) + (1 - P(A))W_B(\lambda)$. As observed in the previous section, W_B is monotone increasing in λ, although $W_A(\lambda)$ need not be monotonic in λ. Therefore, $W(\lambda)$ may not be monotonic in λ.

In a CBS system, $W(\lambda)$ is close to $1/\mu$ if the arrival rate is close to zero, and it goes to infinity if the arrival rate is close to the maximum service rate $c\mu$. Hence, if $\tau \geq \theta/\mu$, there exists a solution to the above equation. However, the uniqueness of the equilibrium depends on the monotonicity of the function $W(\lambda)$ with λ.

We seek analytical results through probing the foregoing two examples. The two-server system shows that although W_A is unimodal, W, as the weighted average of W_A and W_B, is still increasing in λ. The four-server system shows a complex shape of $W(\lambda)$, with a local maximum and minimum.

Lemma 3. *For a system with $c = 2$, $n = 1$, and $N = 3$, under no information, $W(\lambda)$ is strictly increasing in λ.*

In this case, there exists a unique equilibrium owning to the monotonicity of $W(\lambda)$. It is interesting that

only ATC exists here. Although a larger arrival rate may decrease the average delay in mode A, it also increases the average delay in mode B. Therefore, under no information, the aggregate effect of the larger arrival rate in this example is still increasing delay.

Lemma 4. *When $c = 4$, $n = 1$, and $N = 4$, under no information, $W(\lambda)$ has a local maximum $r_1 \in (0, \mu)$ and a local minimum $r_2 \in (\mu, 2\mu)$.*

In this example, W first increases, then decreases, and finally increases again with the arrival rate. This behavior can be explained as follows. In a light traffic case, the system is mainly operated in mode A and behaves like an $M/M/n$ queue; thus, the expected waiting time increases with the arrival rate. In a moderate traffic case, the system switches between mode A and mode B, and a larger arrival rate facilitates more servers to open or keeps the system in mode B for a longer time. Therefore, the expected waiting time can decrease with the arrival rate in a certain range. In a heavy traffic case, the system is mainly operated in mode B, and it behaves like an $M/M/c$ queue. Hence, the expected waiting time increases again with the arrival rate. The difference in behavior between the two examples can be explained by the difference in the ratio of the number of turned-off servers to the total number of servers and N. When $(c - n)/c$ is larger, the benefit associated with reopening inactive servers tends to be larger; when N is larger, it takes a longer time to switch to mode B, and the benefit associated with a joining customer in mode B tends to be larger.

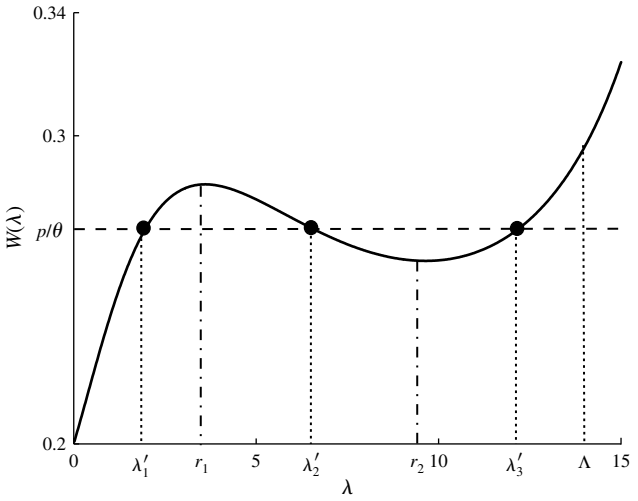
Given the shape of the expected waiting time function, if $W(r_2) < \tau/\theta < W(r_1)$, there exist three solutions, λ_1', λ_2', and λ_3', that satisfy $0 < \lambda_1' < r_1 < \lambda_2' < r_2 < \lambda_3'$, to the following equation:

$$\tau = \theta W(\lambda). \tag{5}$$

As illustrated in Figure 3, there are at most three equilibria. See Proposition 3 in Online Appendix D for a complete list of equilibria in this case. In this example, under no information we observe the ATC behavior in light traffic, FTC behavior in moderate traffic, and ATC behavior again in heavy traffic.

Remark 3. In the existence of mixed and pure equilibria, the pure equilibrium arrival rate Λ is always an ESS. However, there is a situation in which all three equilibria are mixed-strategy equilibrium, $\lambda^e = \lambda_1', \lambda_2'$, and λ_3', as illustrated in Figure 3. Equilibrium λ_2' is locally unstable because a slight increase in it can reduce the expected waiting time W, which, in turn, will attract more arrivals. Both equilibria λ_1' and λ_3' are ESSs because both are mixed equilibria in which customers are indifferent between joining and balking and hence obtain the same payoff in the two equilibria. Therefore we need to consider other criteria to select one of them. Let $\boldsymbol{\alpha}$ be the mixed strategy with a joining probability λ_3'/Λ, and let $\boldsymbol{\beta}$ be the strategy with a joining probability λ_1'/Λ. For customers intending to choose $\boldsymbol{\beta}$, there is a risk that a small stream of other customers will choose $\boldsymbol{\alpha}$; thus, the effective arrival rate is a little larger than λ_1', generating an expected waiting cost higher than price τ. For customers intending to choose $\boldsymbol{\alpha}$, there is a risk that a small stream of other customers will choose $\boldsymbol{\beta}$; thus the effective arrival rate will be a little smaller than λ_3', generating an expected waiting cost lower than price τ. Consequently, the equilibrium with strategy $\boldsymbol{\alpha}$ generates a larger expected payoff for customers. It follows from Definition 3 that λ_3' risk-dominates λ_1'.

Figure 3 **Equilibrium Analysis Under No Information Assuming $c = 4$, $n = 1$, $N = 4$, and $\mu = 5$**

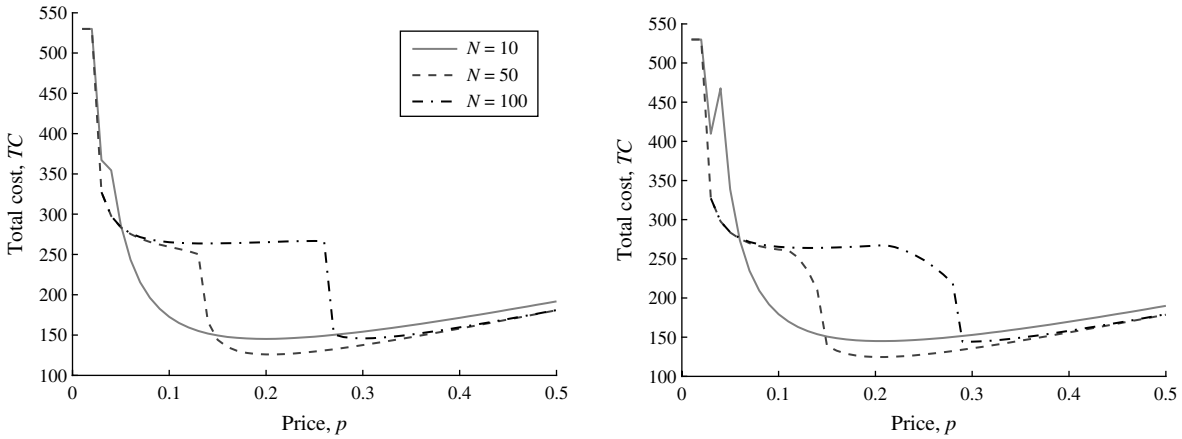


Remark 4. In Figure 3, as long as the τ/θ line is above the local minimum $W(r_2)$, the equilibrium arrival rate is λ_3'. As τ/θ decreases to slightly below $W(r_2)$, the equilibrium arrival rate jumps down from λ_3' to λ_1', with the expected waiting time roughly unchanged or slightly reduced. This sudden decrease in the equilibrium arrival rate could result in a sharp increase in operations cost as the main stream of customers switch to the price system.

5. Implications for Border-Crossing Systems

Based on the Pacific Truck Crossing Station, we consider a free system with $c = 5$ servers, $n = 3$, and $\mu = 50$ vehicles per hour (these values are from the statistics of real operations; see http://www.borderlineups.com). Because congestion is not an issue for normally loaded traffic situations, we focus on a critically loaded system with $\Lambda = c\mu$. This setting

Figure 4 Cost Function Along with Price τ



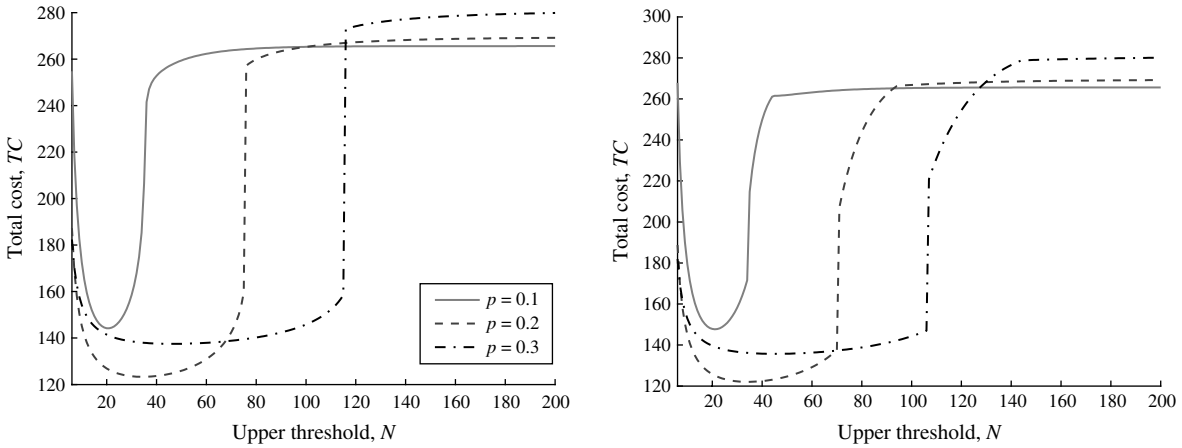
Note. The left graph assumes no information, and the right graph assumes partial information.

is close to a realistic border-crossing station during rush hours and justifies the use of the price lane. For the cost parameters, without loss of generality, we consider a benchmark case with $S = 50$, $\theta = 1$, $K_0 = 500$, $C_o = 10$, and a linear (with $a = 1$) price lane service cost function. Note that at present we do not have a real data set for these cost parameters because the toll has not yet been implemented. However, when the toll system is put in practical use, these parameters can be estimated. We conduct extensive numerical studies on the systems with different combinations of (τ, N) policies and information scenarios.

5.1. Total Cost

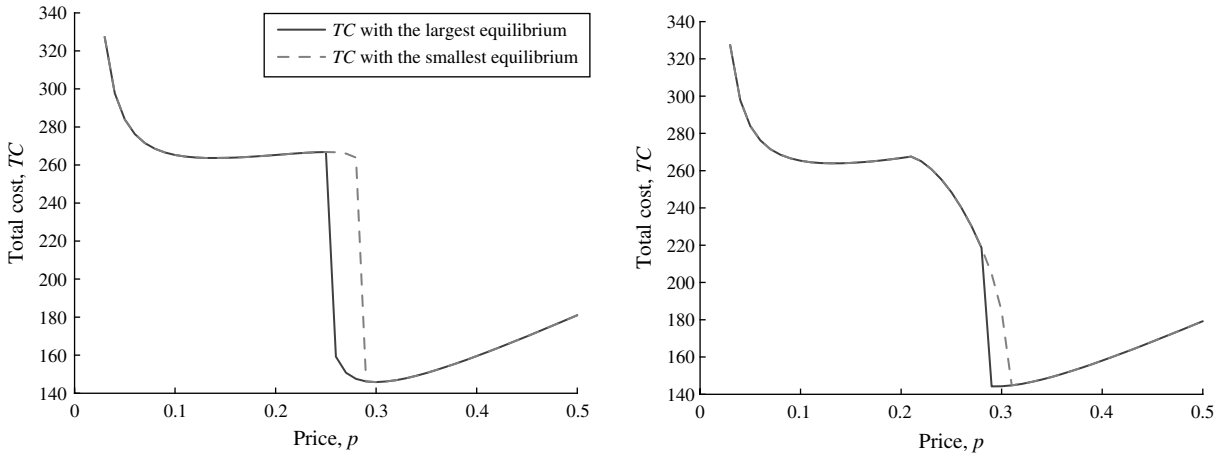
We now examine the impact of strategic customer behavior on the total cost. To calculate the total cost, we assume that the largest equilibrium arrival rate is chosen in the case of multiple equilibria, as suggested by the two selection criteria (ESS and risk dominance). We first plot the function of TC with respect to the two decision variables, τ and N, in Figures 4 and 5. When $N = 10$, there exists a local minimum of TC. It is worth noting that in the partial information case, there exists a sharp peak of TC when τ is small. This can be explained by customers' strategic behavior, as described in Remark 2 of §4.1. When $N = 50$ or 100, TC roughly consists of two convex curves of τ. The first curve is formed when the price is low, resulting in the system behaving like an $M/M/n$ queue because of the small arrival rate; the second curve represents the two-mode status when the price is high enough. Figure 4 also shows the insensitivity of TC with respect to the higher price, as the TC curve is quite flat to the right of the optimal price. However, TC increases sharply when the price is reduced from the optimal value. The sharp increase of TC is due to the dramatic decrease of the equilibrium arrival rate, as explained in Remark 4 of §4.2. Therefore, managers should be very cautious about setting the optimal price in a CBS system. When the system parameters

Figure 5 Cost Function Along with Upper Threshold N



Note. The left graph assumes no information, and the right graph assumes partial information.

Figure 6 Comparison of Cost Functions with Different Equilibrium Arrival Rates



Note. The left graph assumes no information, and the right graph assumes partial information.

are uncertain, it is safer to set a relatively *high* price; otherwise, strategic customer behavior could result in a *too thin* arrival stream into the free system, which could significantly increase the total cost. The numerical study also shows that the total cost under partial information is larger than the one under no information.

As illustrated in Figure 5, a higher price also ensures that the total cost is less sensitive to N. This is another benefit of setting a relatively high price.

Next, we relax the assumption that the largest equilibrium arrival rate is chosen in calculating the total cost. Instead, we assume that the smallest one is chosen. Using the same example as before with $N = 10$ and $N = 50$, we find that the total cost function remains the same regardless of the equilibrium choice. For the $N = 100$ case, there exists a slight difference in the TC curve, as shown in Figure 6. The sharp decrease or downward jump occurs earlier with the largest equilibrium chosen than with the smallest equilibrium chosen. Except for this small difference, the whole pattern of the TC curve remains almost the same. Therefore, the main insight gained in the foregoing analysis is still valid, regardless of the equilibrium choice.

5.2. A Comparison with Simulation Results

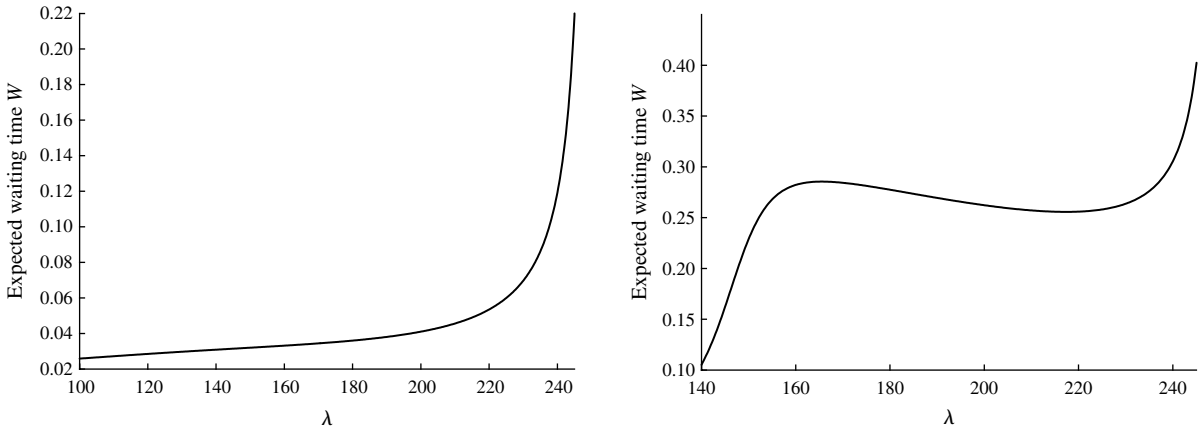
One of our key assumptions is that customers do not need to wait in the price system. In a real situation such as a border-crossing station with a fast pass lane, this assumption is not perfectly satisfied, although most of the time there is no lineup. We now examine the impact of nonzero waiting in the price system by using an Arena simulation model. We consider $c = 5$, $n = 3$, $\mu = 50$, that the price lane is a single-server queue with exponentially distributed service times, and $\mu_t = 120$, which is more than two times the free service rate. In the simulation model, the customers make their lane selection decision based on a cost comparison between the two queues in which the expected waiting time for each queue is approximated by the updated average waiting time of past customers when they make their decisions. We simulate the critically loaded system with $\Lambda = 250$ by changing the price from low to high values.

Table 2 lists the long-run average waiting time of the free lane and the average queue lengths for both the free and price lanes. To gain a better understanding on the two systems' congestion, we list the expected number of customers and the expected waiting time in price system, denoted as L_t and W_t, respectively. Figure 7 shows the shapes of the expected waiting time function for $N = 10$ and $N = 100$ cases. Figure 8 shows those generated through simulation. Clearly, their shapes are very similar except that those generated by simulation are slightly higher than the corresponding ones calculated by the model. This

Table 2 Simulation Outcomes Assuming $c = 5$, $n = 3$, $\mu = 50$, $\mu_t = 120$, $\Lambda = 250$, and $N = 100$

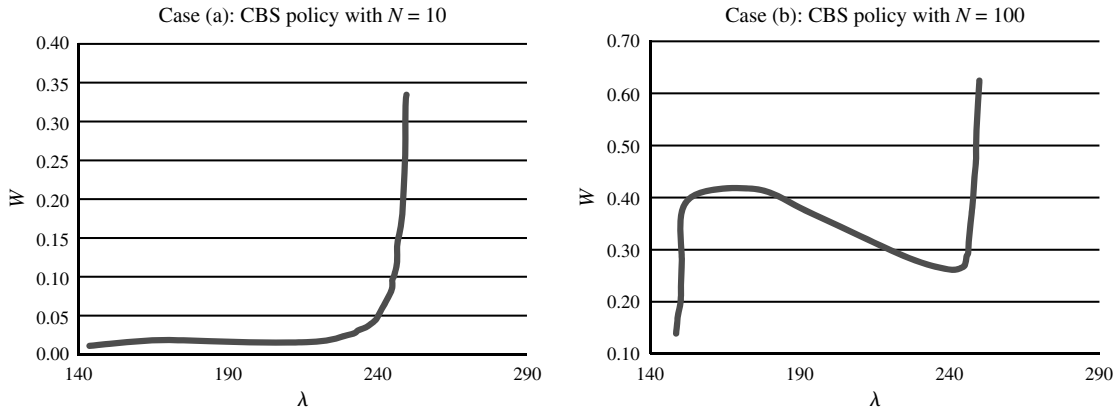
τ	λ	W	L	95% CI for L	W_t	L_t
0.03	148.46	0.1391	20.70	(20.39, 21.00)	0.0706	7.19
0.04	149.17	0.1764	26.37	(25.25, 27.48)	0.0845	8.60
0.05	150.01	0.2027	30.35	(29.62, 31.07)	0.0656	6.59
0.07	150.34	0.2692	40.41	(39.62, 40.84)	0.0677	6.75
0.10	152.21	0.3945	59.84	(58.98, 60.70)	0.0662	6.42
0.13	174.70	0.4166	72.34	(69.38, 75.30)	0.0627	4.71
0.14	193.88	0.3694	70.52	(66.40, 74.63)	0.0502	2.80
0.15	225.76	0.2853	64.17	(60.56, 67.77)	0.0369	0.92
0.16	239.88	0.2619	62.62	(60.89, 64.35)	0.0173	0.17
0.17	244.31	0.2667	65.33	(62.41, 68.24)	0.0117	0.07
0.18	245.23	0.2757	67.71	(63.22, 72.20)	0.0115	0.06
0.21	246.40	0.3233	79.76	(73.23, 86.29)	0.0114	0.04
0.24	247.21	0.3672	90.84	(83.29, 98.40)	0.0095	0.03
0.27	247.78	0.4006	99.21	(88.55, 109.87)	0.0094	0.02
0.30	248.07	0.4305	106.80	(92.43, 121.17)	0.0084	0.02
0.35	248.71	0.4714	117.22	(99.45, 134.98)	0.0074	0.01

Figure 7 Shapes of Expected Waiting Times Under No Information Assuming $c = 5$, $n = 3$, and $\mu = 50$



Note. The left graph assumes $N = 10$, and the right graph assumes $N = 100$.

Figure 8 Free Lane Expected Waiting Time with Nonzero Waiting in the Price Lane



Note. Case (a) assumes $N = 10$; case (b) assumes $N = 100$.

inflated expected waiting time is due to the non-Poisson arrival process to the free lane resulting from the price lane queueing effects. Note that the main purpose of developing the simulation model is to show the pattern of the system performance for a system with a nonzero-wait toll lane rather than to check the accuracy of approximations. Extensive simulations reveal the similar patterns for all cases considered. Therefore, the simulation model verifies that the main performance characteristics from our zero-wait-time price lane model hold for more realistic two-queue situations. The pattern of $W(\lambda)$ first increases, then decreases, and finally increases again with λ for a large N case, and is monotonically increasing with λ for a small N case. This even strengthens some insights obtained from our analytical model. For example, we showed before, as illustrated in Figure 6, that when the price is slightly lower than the optimal level, the arrival rate into the free system is reduced, whereas it actually results in a sharp increase of the total cost due to the CBS policy. This striking conclusion persists here. In Table 2, for a capacity-limited price system with $N = 100$, a slight drop in price can significantly congest both the free and price systems. For example, when τ decreases from 0.16 to 0.13, W of the free system increases from 0.2619 to 0.4166, and the expected number of customers in the price system increases from 0.17 to 4.17. Another observation from Table 2 is that a higher price ($\tau \geq 0.15$) induces more customers to choose the free lane so that the zero-waiting-time assumption is satisfied because of customers' strategic behavior.

6. Conclusion

We have studied strategic customer behavior in joining a free service queue or paying for fast pass service in a multiserver service system with the CBS policy, motivated by border-crossing waiting lines. We consider two information scenarios: (i) the *no information* case in which customers only use the long-term expected waiting time, and (ii) the *partial information* case in which customers know the CBS policy with the servers' status but not the real-time queue length.

Under each information scenario, we discuss the equilibrium arrival rates by considering the customers' choice behavior and then investigate its impact on system performance.

We find that for a CBS system with the price service option, multiple equilibrium arrival rates may exist. We address the equilibrium selection issue in terms of stability and risk dominance. Based on the equilibrium analysis, we examine the impact of information on system performance. One important finding is that information can significantly degrade the system performance, especially in a CBS system with a small upper threshold for reopening inactive servers. This is because customers make their lane selection decisions in a decentralized manner, ignoring both the positive and negative externalities of their actions. Information, in some sense, encourages this socially undesirable behavior, which may degrade the system performance.

We show that because of customers' strategic behavior, system performance is very sensitive to price when it is relatively low: A slight price drop can induce a sharp increase in the number of customers choosing the price lane, which significantly increases the total system cost. On the other hand, the total system cost becomes less sensitive to the price when it is set to be relatively high. Furthermore, the total system cost also becomes less sensitive to N when the price is set high. Therefore, to avoid dramatic performance oscillation, the system manager should start with a relatively high price and gradually adjust it downward as the system matures.

A critical assumption of this study is the zero waiting time in the fast lane. Although the motivating example approximately satisfies this assumption and the simulation model verifies the robustness of the results, our model does not fit situations in which the fast lane queue is comparable with (or not significantly shorter than) the free lane queue. Thus, our study has the limitation of not being able to apply to situations in which two queues must be considered for performance evaluation. However, given that the results are shown to be robust to this key assumption, it is still expected that they are applicable to many other contexts in which the model might be a better fit (e.g., fast pass lanes for a bridge or tunnel). Furthermore, our analysis can be viewed as the first step in studying the more complex two-queue system. Another limitation is that we only considered the no information and partial information scenarios. There exist some real systems in which the real-time queue length or waiting time is available and can be disclosed to customers. Our results may not be valid in such situations. These limitations can be seen as fruitful avenues for future research under the same theme.

Electronic Companion

An electronic companion to this paper is available as part of the online version at http://dx.doi.org/10.1287/msom.1120.0406.

Acknowledgments

The authors thank editor-in-chief Stephen C. Graves, the associate editor, and three anonymous reviewers for many helpful comments and suggestions that led to significant improvement of this paper. The first author's research was supported in part by the Hong Kong General Research Foundation [Grant PolyU 543210]. The second author is thankful for support from the Natural Sciences and Engineering Research Council of Canada [Grant RGPIN197319].

References

Afimeimounga H, Solomon W, Ziedins I (2005) The Downs-Thomson paradox: Existence, uniqueness and stability of user equilibrium. *Queueing System* 49:321–334.

Airline Industry Information (2008) United Airlines launches Premier Line Service. (December 9), http://www.highbeam.com/doc/1G1-190254352.html.

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

Allon G, Bassamboo A, Gurvich I (2011) "We will be right with you": Managing customer expectations with vague promises and cheap talk. *Oper. Res.* 59:1382–1394.

Armony M, Maglaras C (2004a) Contact centers with a call-back option and real-time delay information. *Oper. Res.* 52:527–545.

Armony M, Maglaras C (2004b) On customer contact centers with a call-back option: Customer decisions, routing rules and system design. *Oper. Res.* 52:271–292.

Armony M, Shimkin N, Whitt W (2009) The impact of delay announcements in many-server queues with abandonment. *Oper. Res.* 57:66–81.

Associated Press (2011) "Ready lane" opening at Peace Arch border crossing. (April 13), http://www.komonews.com/news/local/119763214.html.

Ata B, Shneorson S (2006) Dynamic control of an $M/M/1$ service system with adjustable arrival and service rates. *Management Sci.* 52:1778–1791.

Bakshi N, Gans N (2010) Securing the containerized supply chain: Analysis of government incentives for private investment. *Management Sci.* 56:219–233.

Burnetas A, Economou A (2007) Equilibrium customer strategies in a single server Markovian queue with setup times. *Queueing Systems* 56:213–228.

Dimitrakopoulos Y, Burnetas A (2011) Customer equilibrium and optimal strategies in an $M/M/1$ queue with dynamic service control. Working paper, University of Athens, Athens, Greece.

Economou A, Manou A (2011) Equilibrium balking strategies for a clearing queueing system in alternating environment. *Ann. Oper. Res.*, ePub ahead of print November 16, http://rd.springer.com/article/10.1007/s10479-011-1025-x.

Edelson N, Hildebrand K (1975) Congestion tolls for Poisson queueing processes. *Econometrica* 43:81–92.

Guo P, Hassin R (2011) Strategic behavior and social optimization in Markovian vacation queues. *Oper. Res.* 59:986–997.

Guo P, Zipkin P (2007) Analysis and comparison of queues with different levels of delay information. *Management Sci.* 53:962–970.

Harsanyi J, Selten R (1988) *A General Theory of Equilibrium Selection in Games* (MIT Press, Cambridge, MA).

Hassin R (1986) Consumer information in markets with random product quality: The case of queues and balking. *Econometrica* 54:1185–1195.

Hassin R, Haviv M (2003) *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems* (Kluwer, Boston).

Haviv M, Kerner Y (2007) On balking from an empty queue. *Queueing System* 55:239–249.

Ibrahim R, Whitt W (2009a) Real-time delay estimation based on delay history. *Manufacturing Service Oper. Management* 11: 397–415.

Ibrahim R, Whitt W (2009b) Real-time delay estimation in overloaded multiserver queues with abandonments. *Management Sci.* 55:1729–1742.

Johari R, Kumar S (2008) Congestible services and network effects. Working paper, Graduate School of Business, Stanford University, Stanford, CA.

Naor P (1969) The regulation of queue size by levying tolls. *Econometrica* 37:15–24.

Omahen K, Marathe V (1978) Analysis and applications of the delay cycle for the $M/M/c$ queueing system. *J. Assoc. Comput. Machinery* 25:283–303.

Tian N, Zhang ZG (2006) *Vacation Queueing Models: Theory and Applications* (Springer, New York).

Veeraraghavan S, Debo L (2009) Joining longer queues: Information externalities in queue choice. *Manufacturing Service Operations Management* 11:543–562.

Wein L, Wilkins A, Baveja M, Flynn S (2006) Preventing the importation of illicit nuclear materials in shipping containers. *Risk Anal.* 26:1377–1393.

Whitt W (1999) Improving service by informing customers about anticipated delays. *Management Sci.* 45:192–207.

Whitt W (2007) What you should know about queueing models to set staffing requirements in service systems. *Naval Res. Logist.* 54:476–484.

Zhang ZG (2009) Performance analysis of a queue with congestion-based staffing policy. *Management Sci.* 55:240–251.

Zhang ZG, Luh H, Wang C (2011) Modeling security-check queue. *Management Sci.* 57:1979–1995.