https://doi.org/10.1016/j.orl.2017.11.009



# Equilibrium Queueing Strategies of Two Types of Customers in a Two-Server Queue

Yanli Tang Pengfei Guo Yulan Wang*

*Faculty of Business, The Hong Kong Polytechnic University, Hong Kong,*
*yanli.tang@connect.polyu.hk, pengfei.guo@polyu.edu.hk, yulan.wang@polyu.edu.hk*

We consider a single queue with two identical servers and two types of customers. The high-type customer is more delay-sensitive but brings less workload to the system than the low-type customer. We obtain the equilibrium queueing strategy for each type of customers.

**Keywords**: Two types of customers Strategic queueing behavior Equilibrium analysis Two-server queue

## 1. Introduction

It is commonly observed that a queueing system contains different types of customers. Consider a checkout queue in a retailer store such as a supermarket. Customers can be categorized into two types based on the workload they bring to the server and their delay sensitivity. The high-type customers are often highly delay-sensitive but only checkout a few items, bringing little workload to the queue, whereas the low-type customers are often highly delay-insensitive but checkout many items, bringing heavy workload to the queue. Supermarkets serve these two types of customers either via two separate queues or via a single queue. For example, some supermarkets differentiate the customers who buy less than, say, 5 items from the others and serve these two groups of customers separately via a "fast line" and a "regular" line. In contrast, other supermarkets do not differentiate customers and serve them via a single multi-server queue, regardless of whether they checkout many or a few. Under the former separating case, the checkout system contains two dedicated queues. Customers' strategic queueing behavior can thus be found for each individual queue from the rich literature on queueing strategy; see Hassin and Haviv [1] for a survey. In contrast, under the latter pooling case, the system has multiple servers

*Corresponding author.

with a mixture of two types of customers with different delay sensitivities and bringing different workload requirements. The customer queueing strategy for this type of system has not been analyzed according to our best knowledge. This motivates our study.

For the sake of analytical tractability, we make the following assumptions in our model. Customers arrive to the system according to a Poisson process. There is a single queue with two identical servers, i.e., the servers have the same service speed. Customers bring random amounts of workload to a server, which is assumed to be exponentially distributed. And customers are classified into two types according to their delay sensitivity and workload amount. The high-type customer is highly delay-sensitive and brings a stochastically smaller amount of workload to the server whereas the low-type customer is lowly delay-sensitive and brings a stochastically larger amount of workload to the server. The service reward of two types of customers are also different.

Our paper is related to the study on customers' strategic queueing behaviors, dated back to Naor [2]. Naor [2] considers a fully observable queueing system with one server where the customers choose whether to balk or to join based on the service rewards and the waiting time. A large body of research along this line has been conducted henceafter; see Hassin and Haviv [1] and Hassin [3] for the comprehensive literature reviews in this field. In particular, our paper is related to those papers studying the queueing system with different types of customers and multiple servers. Kulkarni [4] studies an M/G/1/1 system with two players who have different waiting costs. Ni et al. [5] investigate the service provider's service speed and price decisions when the M/M/1 system contains two types of customers. However, in these papers, only one server is considered while we consider two servers. Moreover, the service time in our work is type-dependent. Thus, the unconditional service time does not follow an exponential distribution, which makes the analysis here much more challenging.

Our analysis shows that if the longest waiting time that the low-type customer is willing to bear is more than that of the high-type customer, all the low-type customers join the queue as long as some of the high-type customers join. Otherwise, the high-type customers all join the queue once some low-type customers join.

This paper is organized as follows. Section 2 describes the model and the notations. Section 3 provides the expression of the expected waiting time and related properties. In Section 4, we analyze the customers' strategic queueing behavior.

## 2. Model Setup

Consider two types of customers: a high type (labelled $H$) and a low type (labelled $L$). They arrive to the system according to a Poisson process with rate $\Lambda$. The fraction of the high-type customers in the arriving customers is $\gamma$. Let $\lambda_H$ and $\lambda_L$ denote the potential

arrival rates of the high- and low-type customers, respectively. Then,

$$\lambda_H = \gamma\Lambda \text{ and } \lambda_L = (1-\gamma)\Lambda.$$

The two types of customers differ from each other in the following three aspects. First, the two types of customers are heterogeneous in their delay sensitivity, denoted by $\theta_i$, $i = H, L$. The high-type customer is more delay sensitive than the low-type customer, i.e., $\theta_H > \theta_L$. Second, the two types of customers may checkout different kinds of products. Denote the reward received by a type-$i$ customer once the checkout is complete by $R_i$, $i = H, L$. We do not posit the constraint on the relative magnitude between $R_H$ and $R_L$. Third, the two types of customers bring to the system different workload sizes. The high-type customer normally checkouts less items than the low-type customer and thus normally brings less workload to the system. We assume that the workload size for the low-type customers follows an exponential distribution with mean $1/\mu_L = 1/\mu$ and that for the high-type customers follows an exponential distribution with mean $1/\mu_H = 1/(\beta\mu)$, where $\beta > 1$. By this assumption, the workload brought by each low-type customer is stochastically larger than that of the high-type customer.

The system has one single queue with two identical servers. The service policy is first-come first served. That is, the customer waiting at the front of the queue is served by the first available server.

Customers care about the expected waiting time in the queue. Hereafter, we refer to the waiting time in the queue as the waiting time for simplicity. The relaxation of this assumption to consider the *sojourn* time (the sum of waiting time in the queue and the service time) will not affect the qualitative results of our paper. We also assume that the queue length is unobservable when a customer makes her queueing decision and a customer knows her own type. This assumption is natural for the supermarket case. Note that customers usually have a shopping list in their mind and have a target shop before they go. Given the arrival rates $\lambda_H$ and $\lambda_L$, a tagged type-$i$ customer obtains the following utility if she joins the queue:

$$U_i = R_i - \theta_i W(\lambda_H, \lambda_L), i = H, L,$$

where $W(\lambda_H, \lambda_L)$ is the expected waiting time. If a customer balks, she receives an utility of 0. The expected waiting time of each customer is affected by the equilibrium joining strategies of both types of customers.

## 3. Expected Waiting Time

We follow Kotiah and Slater [6] to calculate the expected waiting time in this two-server Poisson queue with two types of customers. Given the arrival rates of the high-

and low-type customers, $\lambda_H$ and $\lambda_L$, the probability that an incoming customer is type $i$, $i = H, L$ can be derived as

$$\alpha_i = \frac{\lambda_i}{\lambda_H + \lambda_L}.$$

Define

$$\lambda := \lambda_H + \lambda_L;\ \rho := \frac{1}{2}\left(\frac{\lambda_H}{\mu_H} + \frac{\lambda_L}{\mu_L}\right).$$

Then following Kotiah and Slater [6], the expected waiting time if all the incoming customers join the queue can be derived as

$$W(\lambda_H, \lambda_L) = \frac{\left(1 - \frac{1}{2\rho}\right)\left(\frac{\alpha_H}{2\mu_H^2} + \frac{\alpha_L}{2\mu_L^2}\right)\lambda + \frac{1}{2\rho}\left(\frac{P_{H,0;1}}{\mu_H} + \frac{P_{L,0;1}}{\mu_L}\right)}{1 - \rho}, \tag{1}$$

where $P_{i,0;1}$ represents the probability that there is only *one* type-$i$ customer in the system and she is served by server 1, $i = H, L$. It can be shown that

$$P_{H,0;1} = \frac{(1-\rho)\lambda\alpha_H D_L}{\alpha_L(\lambda + 2\mu_L)D_H + \alpha_H(\lambda + 2\mu_H)D_L} \text{ and } P_{L,0;1} = \frac{(1-\rho)\lambda\alpha_L D_H}{\alpha_L(\lambda + 2\mu_L)D_H + \alpha_H(\lambda + 2\mu_H)D_L},$$

where

$$\begin{aligned} D_i &= 2\mu_i y + 3\mu_i x_0 + \lambda x_0^2, \quad y = \frac{2\mu_H\mu_L\rho}{\lambda^2} \\ x_0 &= 1 - \frac{\lambda + \mu_H + \mu_L - \sqrt{(\lambda + \mu_H + \mu_L)^2 - 4\lambda(\alpha_H\mu_H + \alpha_L\mu_L)}}{2\lambda}. \end{aligned}$$

We can further derive the following result. The proof is relegated to the appendix.

**Lemma 1.** $\frac{\partial W(\lambda_H,\lambda_L)}{\partial \lambda_L} > \frac{\partial W(\lambda_H,\lambda_L)}{\partial \lambda_H}$ *for any* $\beta > 1$.

Lemma 1 implies that the marginal impact of the low-type customer on the system waiting time is larger than that of the high-type customer. This is rather intuitive as the low-type customer brings more workload to the system and needs a longer service time than the high-type customer.

## 4. Equilibrium Queueing Strategy Analysis

In this section, we derive the equilibrium queueing strategies of two types of customers in the two-server queue. Let

$$v_i = \frac{R_i}{\theta_i}, i = H, L.$$

Recall that $R_i$ is the service reward of the type-$i$ customer from joining the queue and receiving the service from the server while $\theta_i$ is her delay sensitivity parameter, $i = H, L$. Thus, $v_i$ represents the longest time that a type-$i$ customer is willing to wait for getting the service. For ease of exposition, we say that the low-type customer has a relatively higher (respectively, lower) *patience level* than the high-type customer when $v_L > v_H$ (respectively, $v_L < v_H$). Below we analyze the equilibrium joining strategies of the high- and low-type customers for the cases $v_L > v_H$ and $v_L < v_H$, respectively.

*4.1. Customer Equilibrium Queueing Strategy When $v_L > v_H$*

When $v_L > v_H$, as long as high type customers join at a non-zero rate, then all of the low-type customers join the queue. This is because the low-type customer has a higher patience level than the high-type one and both types of customers queue in the same line. Therefore, we can derive the customers' equilibrium queueing strategies by focusing on analyzing the high-type customers' equilibrium joining behavior. Let the superscript "$e$" denote the results in equilibrium.

**Scenario 1: All the high-type customers balk, i.e., $\lambda_H^e = 0$.**

Under this scenario, equation (1) can be rewritten as

$$W(0, \lambda_L) = \frac{\rho_L^2}{(4 - \rho_L^2)\mu},$$

where

$$\rho_L := \frac{(1 - \gamma)\Lambda}{\mu}.$$

This result is consistent with that in the literature regarding the expected waiting time in an $M/M/2$ system; see, e.g., Gross et al. [7, pp. 67-69]. Regarding the low-type customers, if $U_L = R_L - \theta_L W(0, \lambda_L) \geq 0$, or equivalently, $W(0, \lambda_L) \leq v_L$, then all of the low-type customers join the queue in equilibrium. In addition, since no high-type customer joins the queue, it implies that $U_H = R_H - \theta_H W(0, \lambda_L) < 0$. Thus, when $v_H < W(0, \lambda_L) \leq v_L$, in equilibrium, all the low-type customers join the queue while all the high-type customers balk.

Next, consider the case that $U_L = R_L - \theta_L W(0, \lambda_L) < 0$. That is, if all the low-type customers join the queue, then their utility is negative. Thus, in equilibrium, some low-type customers join the queue while others balk. The equilibrium arrival rate of the low-type customers, denoted by $\lambda_L^e$, must satisfy

$$R_L - \theta_L W(0, \lambda_L^e) = 0, \tag{2}$$

where

$$W(0, \lambda_L^e) = \frac{(\rho_L^e)^2}{(4 - (\rho_L^e)^2)\mu}; \quad \rho_L^e := \frac{\lambda_L^e}{\mu}.$$

Because $W(0, \lambda_L^e)$ is increasing in $\lambda_L^e$, it can be shown that equation (2) has a unique solution as follows:

$$\lambda_L^e = 2\mu \sqrt{\frac{v_L \mu}{1 + v_L \mu}}.$$

**Scenario 2: A fraction of the high-type customers joins the queue, i.e., $\lambda_H^e > 0$.**

Under this scenario, since some high-type customers join the queue, as aforementioned, all the low-type customers shall join the queue. That is, $\lambda_L^e = \lambda_L$. Moreover, we have $U_H = R_H - \theta_H W(0, \lambda_L) > 0$; otherwise, it is impossible for a high-type customer to join the queue. When $U_H = R_H - \theta_H W(\lambda_H, \lambda_L) > 0$, or equivalently, when $W(\lambda_H, \lambda_L) \leq v_H$, then all of the high-type customers join the queue in equilibrium. However, if $W(\lambda_H, \lambda_L) > v_H$, then some high-type customers balk the system. Under such scenario (i.e., when $W(\lambda_H, \lambda_L) > v_H$ and $R_H - \theta_H W(0, \lambda_L) > 0$ ), the equilibrium arrival rate is given by the solution to $U_H = R_H - \theta_H W(\lambda_H^e, \lambda_L) = 0$, which can be simplified to

$$W(\lambda_H^e, \lambda_L) = v_H.$$

In this queueing system, as arrival rate $\lambda_H^e$ increases, the queue becomes stochastically longer and hence $W(\lambda_H^e, \lambda_L)$ is increasing in $\lambda_H^e$. It can be shown that there exists a unique $\lambda_H^e = \{\widehat{\lambda} : W(\widehat{\lambda}, \lambda_L) = v_H\}$.

Based on the above analysis, we can derive the following proposition.

**Proposition 1.** *When the low-type customer has a higher patience level than the high-type customer, i.e., $v_L > v_H$, the customers' equilibrium joining-balking behaviors are as follows.*

(a) *When $v_H < v_L < \frac{\rho_L^2}{(4-\rho_L^2)\mu}$, in equilibrium all the high-type customers balk, i.e., $\lambda_H^e = 0$. A fraction of the low-type customers join the queue and the corresponding equilibrium arrival rate $\lambda_L^e = 2\mu \sqrt{\frac{v_L \mu}{v_L \mu + 1}}$.*

(b) *When $v_H < \frac{\rho_L^2}{(4-\rho_L^2)\mu} \leq v_L$, in equilibrium all the low-type customers join the queue while all the high-type customers balk, i.e., $\lambda_H^e = 0$, $\lambda_L^e = \lambda_L$.*

(c) *When $v_L > v_H \geq \frac{\rho_L^2}{(4-\rho_L^2)\mu}$ and $W(\lambda_H, \lambda_L) > v_H$, in equilibrium all the low-type customers join the queue, i.e., $\lambda_L^e = \lambda_L$. A fraction of the high-type customers join the queue and the corresponding equilibrium arrival rate satisfies $W(\lambda_H^e, \lambda_L) = v_H$.*

(d) *When $v_L > v_H \geq \frac{\rho_L^2}{(4-\rho_L^2)\mu}$ and $W(\lambda_H, \lambda_L) \leq v_H$, all customers of both types join the queue in equilibrium, i.e., $\lambda_H^e = \lambda_H$, $\lambda_L^e = \lambda_L$.*

*4.2. Customer Equilibrium Queueing Strategy When $v_H > v_L$*

In this section, we consider the situation in which the high-type customer has a higher patience level than the low-type customer. We can also derive the equilibrium queueing strategies of the two types of customers by conducting the similar analysis as that of §4.1. Define

$$\rho_H = \frac{\gamma\Lambda}{\beta\mu}.$$

The following Proposition 2 summarizes the results.

**Proposition 2.** *When the high-type customer has a higher patience level than the low-type customer, i.e., $v_H > v_L$, the customers' equilibrium joining-balking behaviors are as follows.*

(i) *When $v_L < v_H < \frac{\rho_H^2}{(4-\rho_H^2)\beta\mu}$, in equilibrium all the low-type customers balk, i.e., $\lambda_L^e = 0$. A fraction of the high-type customers join the queue and the corresponding equilibrium arrival rate $\lambda_H^e = 2\beta\mu\sqrt{\frac{v_H\beta\mu}{v_H\beta\mu+1}}$.*

(ii) *When $v_L < \frac{\rho_H^2}{(4-\rho_H^2)\beta\mu} \le v_H$, in equilibrium all the high-type customers join the queue while all the low-type customers balk, i.e., $\lambda_H^e = \lambda_H$, $\lambda_L^e = 0$.*

(iii) *When $v_H > v_L \ge \frac{\rho_H^2}{(4-\rho_H^2)\beta\mu}$ and $W(\lambda_H, \lambda_L) > v_L$, in equilibrium all the high-type customers join the queue, i.e., $\lambda_H^e = \lambda_H$. A fraction of the low-type customers join the queue and the corresponding equilibrium arrival rate satisfies $W(\lambda_H, \lambda_L^e) = v_L$.*

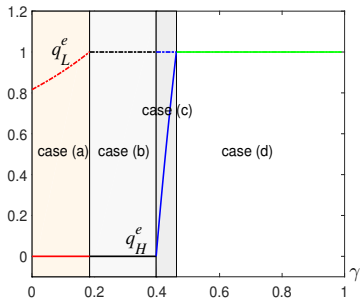
(iv) *When $v_H > v_L \ge \frac{\rho_H^2}{(4-\rho_H^2)\beta\mu}$ and $W(\lambda_H, \lambda_L) \le v_L$, all customers of both types join the queue in equilibrium, i.e., $\lambda_H^e = \lambda_H$, $\lambda_L^e = \lambda_L$.*

Propositions 1 and 2 show that when both types of customers are extremely patient (in terms of the longest waiting times each type can tolerate), they both adopt the "all join" strategy. When the high-type customer is more (respectively, less) patient than the low-type customer, the low-type (respectively, high-type) customers may adopt the "all balk" strategy whereas the high-type (respectively, low-type) customers adopt either the "all join" strategy or a " mixed strategy", joining the queue with a probability.

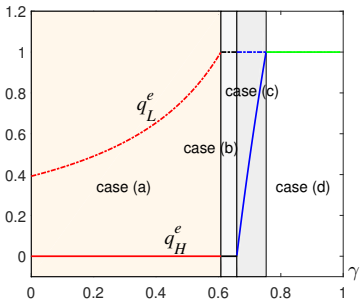
*4.3. Comparison*

In this subsection, we illustrate the above queueing equilibrium analysis stated in §4.1 and §4.2 via four numerical examples. We also let

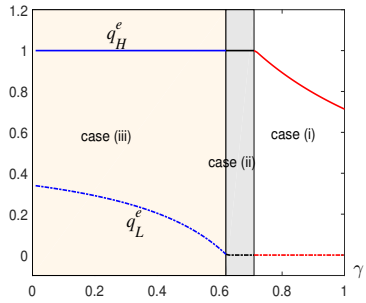
$$q_i^e = \frac{\lambda_i^e}{\lambda_i}, i = H, L.$$



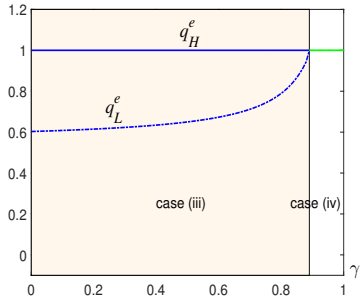

(a) $v_L = 0.2 > v_H = 0.1, \beta = 4, \mu = 1, \Lambda = 1$




(b) $v_L = 0.04 > v_H = 0.03, \beta = 4, \mu = 1, \Lambda = 1$




(c) $v_H = 0.04 > v_L = 0.03, \beta = 1.5, \mu = 1, \Lambda = 1$




(d) $v_H = 0.2 > v_L = 0.1, \beta = 1.5, \mu = 1, \Lambda = 1$

Figure 1: Customers' Equilibrium Joining Strategies

Then, $q_i^e$ represents the equilibrium joining probability of the type-$i$ customer, $i = H, L$.

Figure 1 depicts the equilibrium joining probabilities for both types of customers under four scenarios. From Figure 1(a), we can see that when the low-type customer is more patient than the high-type customer ($v_L > v_H$), a slight increase of $\gamma$ (the fraction of the high-type customer in the arrival population) may lead to a big change in the high-type customers' equilibrium joining probability. This may be caused by the fact that the marginal contribution of a high-type customer towards the system waiting time is lower than that of the low-type customer (as stated in Lemma 1). Consequently, the increase of the high-type customer has less impact on the waiting time. Moreover, the joining probabilities of both types of customers increase as there are more high-type customers in the arrival population (i.e., a larger $\gamma$). In contrast, when the high-type customer is more patient than the low-type customer ($v_H > v_L$), the equilibrium joining probability of the low-type customers can either increase or decrease with $\gamma$, as illustrated in Figures 1(c) and 1(d). Figure 1(c) shows the case that the equilibrium joining probability of the low-type customers decreases with $\gamma$ whereas Figure 1(d) shows the one that the equilibrium joining probability of the low-type customers increases with $\gamma$. Moreover, Figure 1 implies that under all the scenarios, the type that has a higher patience level is more likely to join the queue and "squeeze out" the other type.


## Acknowledgments

We are grateful to the area editor (Professor John Hasenbein) and an anonymous associate editor for very helpful comments and suggestions. The second author acknowledges the financial support by the Research Grants Council of Hong Kong (RGC GRF Reference Number: PolyU 15526716). The third author acknowledges the financial support provided by the Hong Kong Polytechnic University (Grant Number: G-YBQR).


## Appendix: Proofs

Proof of Lemma 1. By the definition of the differential, we have

$$\Delta W_L = \frac{\partial W(\lambda_H, \lambda_L)}{\partial \lambda_L} \Delta\lambda_L + o(\Delta\lambda_L),$$

$$\Delta W_H = \frac{\partial W(\lambda_H, \lambda_L)}{\partial \lambda_L} \Delta\lambda_H + o(\Delta\lambda_H).$$

Now, let $\Delta\lambda_H = \Delta\lambda_L = \Delta\lambda \to 0$. Because every low-type customer checkouts $\beta > 1$ items, we have $\Delta W_L = \frac{\beta}{2\mu}\Delta\lambda > \Delta W_H = \frac{1}{2\mu}\Delta\lambda$. Hence, we obtain $\frac{\partial W(\lambda_H,\lambda_L)}{\partial \lambda_L} > \frac{\partial W(\lambda_H,\lambda_L)}{\partial \lambda_H}$.

## References


[1] R. Hassin, M. Haviv, To queue or not to queue: Equilibrium behavior in queueing systems, Springer Science & Business Media, 2003.

[2] P. Naor, The Regulation of Queue Size by Levying Tolls, Econometric 37 (1) (1969) 15–24.

[3] R. Hassin, Rational queueing, CRC press, 2016.

[4] V. G. Kulkarni, A game theoretic model for two types of customers competing for service, Oper. Res. Lett. 2 (3) (1983) 119–122.

[5] G. Ni, Y. Xu, Y. Dong, Price and speed decisions in customer-intensive services with two classes of customers, Eur. J. Oper. Res. 228 (2) (2013) 427–436.

[6] T . C . T . Kotiah; N . B . Slater, On Two-Server Poisson Queues with Two Types of Customers, Oper. Res. 21 (2) (1973) 597–603.

[7] D. Gross, J. F. Shortle, J. M. Thompson, C. M. Harris, Fundamentals of Queueing Theory, John Wiley & Sons, 2008.