Equilibrium Joining Strategies and Optimal Control of a Make-to-Stock Queue

Qingying Li
Glorious Sun School of Business and Management, Donghua University,
1882 West Yan'an Road, Shanghai, P.R. China 200051
liqingying@dhu.edu.cn

Pengfei Guo*
Department of Logistics and Maritime Studies, Faculty of Business,
The Hong Kong Polytechnic University,
Hung Hom, Kowloon, Hong Kong
pengfei.guo@polyu.edu.hk

Chung-Lun Li
Department of Logistics and Maritime Studies, Faculty of Business,
The Hong Kong Polytechnic University,
Hung Hom, Kowloon, Hong Kong
chung-lun.li@polyu.edu.hk

Jing-Sheng Song
Fuqua School of Business, Duke University,
Durham, NC 27708, U.S.A.
jssong@duke.edu

*Corresponding author

Abstract

We consider a make-to-stock, finite-capacity production system with setup cost and delay-sensitive customers. To balance the setup and inventory related costs, the production manager adopts a two-critical-number control policy, where the production starts when the number of waiting customers reaches a certain level and shuts down when a certain quantity of inventory has accumulated. Once the production is set up, the unit production time follows an exponential distribution. Potential customers arrive according to a Poisson process. Customers are strategic, i.e., they make decisions on whether to stay for the product or to leave without purchase based on their utility values, which depend on the production manager's control decisions. We formulate the problem as a Stackelberg game between the production manager and the customers, where the former is the game leader. We first derive the equilibrium customer purchasing strategy and system performance. We then formulate the expected cost rate function for the production system and present a search algorithm for obtaining the optimal values of the two control variables. We further analyze the characteristics of the optimal solution numerically and compare them with the situation where the customers are non-strategic.

Key words: Make-to-stock queue; vacation queue; strategic customers; equilibrium analysis

History: Received: September 2015; Accepted: March 2016 by Michael Pinedo, after 1 revision.

1 Introduction

We consider a make-to-stock, finite-capacity production system with setup cost. Once the production is set up, the unit production time follows an exponential distribution. Potential customers arrive according to a Poisson process. To balance the setup and inventory related costs, the production manager adopts a two-critical-number control policy, where the production starts when the number of waiting customers reaches a certain level and shuts down when a certain quantity of inventory has accumulated. Unlike what is typically covered in the literature, here we consider strategic, delay-sensitive customers. That is, if a customer encounters a stock out upon arrival, he/she decides whether to wait for the product or to leave without purchase according his/her overall valuation of the purchase, which depends on the production manager's control decisions.

This model setting is suitable for the situation where customers order products online from a producer of high-end labor-intensive products (e.g., designer handbags, fashion clothing, or specialist bakery products), or from a producer of expensive equipment (e.g., aircraft or farming equipment). The key feature here is that the product concerned is not a commodity; it is expensive and takes time to set up and to produce. Because of high costs and relatively low demand rate, the production manager may not want to hold too much inventory. On the other hand, there are some, although only a few, competitors selling close substitutes. If the service is unsatisfactory, a customer may choose not to purchase the product from this particular producer. To stay competitive, the manager of the current system would need to determine the optimal values of the control variables to strike the right balance between production/inventory costs and customer service.

We formulate the problem as a Stackelberg game between the production manager and the customers, with the former being the game leader. We model the production system as an $M/M/1$ make-to-stock queue with two control variables, and we consider strategic customers who are delay sensitive. Upon arrival, the customers estimate their expected waiting time and then make decentralized decisions on whether to stay or to leave without purchase (i.e., the demand might be lost). The customers' likelihood of staying depends on their utility, which in turn depends on the control variables that the production manager sets. We first derive the equilibrium customer purchasing strategy and system performance. From this, we formulate the expected cost rate function for the production system. We then present a search algorithm for obtaining the optimal values of the two control variables. We also conduct a numerical study to analyze the characteristics of the optimal solution, and we compare the optimal solution with the solution where the customers are

non-strategic.

By considering strategic customers who make decentralized decisions on joining the queue for purchase or balking (leaving without purchase), we gain new insights for controlling the production system through two thresholds. Three conclusions are notable. First, we show that reducing the threshold that triggers production has a greater marginal effect on customers' expected waiting time than increasing the inventory threshold that stops production. Second, when compared to the traditional model with non-strategic customers, our model can yield different inventory control decisions and result in different total costs under various demand traffic situations. Third, we find that with strategic customers, the optimal control parameters exhibit different patterns for patient and impatient customers. When customers are patient, the system will serve all of them, and a slight increase in customer delay sensitivity could cause significant changes in the control parameters. However, when customers' delay sensitivity exceeds a certain level, it becomes optimal for the system not to serve all customers, and under this condition, a slight increase in customer delay sensitivity has less impact on the two control variables.

Our study also contributes to the literature of equilibrium analysis for vacation queues with strategic customers, which mainly focuses on systems with nonnegative queue lengths and one-critical-number policies (see, e.g., Guo and Hassin 2011). Our model is a generalized vacation queue by adopting a two-critical-number policy and allowing negative queue lengths.

In the following, we review the related literature. In a make-to-stock queue, replenishment is made at a finite production rate, and in the meantime the inventory level needs to be controlled. Thus, a production-inventory system with discrete random production and demand can be viewed as a make-to-stock queue. Hence, in the following discussion, we include some of the works on production-inventory systems.

We begin the discussion with optimal policies and performance evaluation of make-to-stock queues. Federgruen and Zheng (1993) consider a production-inventory system with compound Poisson demand, random processing times, and random vacation times to minimize the long-run average expected cost, and they prove that an (s, S) policy is optimal. Van Foreest and Wijngaard (2014) consider a production-inventory system with compound Poisson demand, constant production rate, backlogging, and fixed setup cost of production, and they obtain conditions under which the (s, S) policy is optimal. Other related works include Wu and Chao (2014), who study the optimal policy for a production-inventory system in which the production and demand are modeled by a two-dimensional Brownian motion process, and Shi *et al.* (2014), who study the op-

timization of two different cost metrics for a production-inventory system with compound Poisson demand, constant production rate, and lost sales. Research related to the performance measures of make-to-stock queues includes the development of the queueing systems' operating characteristics and the determination of the systems' optimal policy parameter values. Gavish and Graves (1980) consider an $M/D/1$ production-inventory system with production setup cost, inventory holding cost, and backlogging cost. They analyze a two-critical-number policy for this system and propose an efficient search procedure for finding the optimal policy parameter values. Gavish and Graves (1981) and Lee and Srinivasan (1989) extend Gavish and Graves's (1980) model to an $M/G/1$ production-inventory system, and they develop efficient procedures for finding the optimal policy parameter values. Graves and Keilson (1981) extend Gavish and Graves's (1980) model from a Poisson demand process to a compound Poisson demand process with exponentially distributed demand size. They use the compensation method to derive a closed-form expression for the cost function and apply search procedures to find the optimal policy. Srinivasan and Lee (1991) examine a random review production-inventory system with compound Poisson demand, general processing time, and backordering. They show that under the (s, S) policy, the average cycle cost is convex in S for a given value of $S - s$, and they develop a procedure for computing s and S.

The main difference between the abovementioned studies and ours is that these studies assume exogenously given demand processes, whereas our optimal control problem is conducted with delay-sensitive customers. There are some make-to-stock queue studies involving impatient customers. Li (1992) considers a single-item production control problem with customers staying for service if their utility, as determined by price, quality, and delivery time, is positive. He obtains a newsboy-formula to characterize the optimal threshold policy for the single firm setting, and then extends the analysis to a multiple-firm competition setting. So and Song (1998) study an $M/M/1$ queueing system, where demand is sensitive to delivery time and price, and the production capacity can be expanded at certain cost to satisfy the delivery time guarantee. They obtain joint optimal decisions on price, delivery time guarantee, and capacity expansion. Song (2009) investigates a make-to-stock queueing system with two independent Poisson demands with different priorities. Low-priority demand is fully backlogged, while high-priority demand is backlogged only if the existing number of backorders does not exceed a certain threshold. He analyzes the stability of this system and presents an optimized solution under a prioritized base-stock policy. Benjaafar *et al.* (2010) study an $M/M/1$ modeled production-inventory system, in which unmet demand can be rejected or backordered. Backordered demand may be canceled if a customer's waiting time exceeds his/her

patience time. Both rejected demand and canceled demand incur certain costs. They show that the optimal policy can be described by two thresholds: one for initiating the production and another for admitting an order. Benjaafar and Elhafsi (2012) consider a production-inventory system with two customer classes, where an unmet demand from a patient customer can be backordered if needed, but an unmet demand from an impatient customer is lost. They show that the optimal policy to minimize the cost can be described by two threshold functions. In these studies, when a customer makes a decision, he/she only considers the system status (e.g., price, existing backorders, waiting time experienced) but does not examine other customers' behavior. However, in our model, we assume that customers make strategic decisions by considering the system parameters and other customers' strategic behavior. Recently, Chen et al. (2015) consider a make-to-stock system with strategic customers and a base stock policy. Their work differs from ours in that it uses a base stock inventory policy, it assumes that customers are heterogeneous in delay sensitivity, and it focuses on the optimal decisions of the base stock level and stockout price.

Less attention has been paid to strategic customer behavior in make-to-stock queues or production-inventory systems, but the subject has been well studied in traditional queueing systems where inventory is not considered. Naor (1969) first shows that tolls can be levied to induce strategic customers to adopt a socially optimal behavior that might not be optimal for themselves. Since then, a number of studies on strategic customer behavior in queueing systems have been conducted. A comprehensive review of such works can be found in Hassin and Haviv (2003). Recently, Guo and Hassin (2011, 2012) study $M/M/1$ queues with strategic customers, where the former study considers identical customers and the latter considers heterogeneous customers. Both of these studies examine the no-information scenario and the full-information scenario. They consider an N-policy (i.e., the server starts working when the queue length reaches N and finishes all the work in the system before taking its next "vacation"), and they obtain and compare the equilibria and socially optimal strategies. Economou *et al.* (2011) study a problem with general service time and vacation time distributions. Guo *et al.* (2011) consider the case with partial information on service time. Debo *et al.* (2012) study an $M/M/1$ queue with impatient customers, where the product may be of high or low quality. Customers may be informed or uninformed of the quality, and they observe the queue length and decide whether to stay or not. Guo and Li (2013) conduct a complementary study to Guo and Hassin (2011) by considering two partial-information scenarios. Guo and Zhang (2013) investigate a multi-server queueing system with a congestion-based staffing policy, where the number of working servers is dynamically adjusted according to the queue length. Boudali

and Economou (2013) investigate the effects of catastrophes, where a catastrophe will force all customers to abandon the system, and customers' utility consists of the usual reward from receiving the service and a failure compensation for those who are forced to abandon the system. Manou *et al.* (2014) study the strategic joining decisions for the customers in a transportation station, where a transportation facility visits the station stochastically and serves customers according to its capacity. In this study, observable and unobservable queues are considered, and customers' symmetric Nash equilibrium strategies are determined. Guo and Hassin (2015) consider a service system where strategic customers can place duplicate orders, and the server has the intention of speeding up the service. Dimitrakopoulos and Burnetas (2016) consider a model where the service rate switches between a high level and a low level. Equilibrium analysis has also been conducted in other related queueing systems; see, for example, Sun *et al.* (2010), Burnetas (2013), Economou and Manou (2013), Wang and Zhang (2013), Debo and Veeraraghavan (2014), Xia (2014), and Ziani *et al.* (2015). It should be noted that inventory holding is not allowed in the abovementioned queueing systems.

The rest of the paper is organized as follows. In Section 2, we introduce the queueing model, derive the performance measures, and conduct equilibrium analysis, assuming the values of the control parameters are given. In Section 3, we derive the cost function of the production system and present an algorithm for determining the optimal control parameter values. In Section 4, we conduct a numerical study to analyze the characteristics of the optimal solution. Section 5 concludes the paper. The proofs of all lemmas and propositions are presented in the Online Appendix.

2 The Queueing Model and the Equilibria

As mentioned in Section 1, our problem can be regarded as a Stackelberg game between the production manager and the strategic customers. Specifically, the production manager first makes a decision on the two control parameters, and then the customers make their purchasing decisions. To solve this problem, we use a backward solution procedure: We first determine the customers purchasing decisions by assuming that the values of the control parameters are given. Given the customers purchasing decisions, we then analyze the firm's optimal decision on the control parameters.

In this section, we first provide a queueing model of the production system and derive the important performance measures such as customer expected waiting time, the average inventory

level and the average number of waiting customers. We then conduct the equilibrium analysis of customer purchasing decision and derive the equilibrium arrival rate.

2.1 Queueing model for the production system

The production system can be described as follows. We have a make-to-stock production system with a single production server, in which a two-critical-number policy is adopted. Under this policy, the production starts when N customer orders have accumulated and stops immediately when the inventory level reaches S, where $S \geq 0$. Note that $-N$ can be regarded the inventory level that triggers production, and S can be regarded as the order-up-to level. Thus, we have $-N \leq S-1$, or equivalently, $N \geq -S+1$. Note that N can be positive, zero, or negative. The customers that place orders follow a Poisson process with rate λ. We refer to λ as the *effective arrival rate*. Each customer places one order, and once an order is placed the customer will wait for the product, where the waiting time is either zero (if some inventory of the product is available) or positive (if the product is out of stock). During a production run, products are finished within an independent and identically distributed exponential processing time with a mean value of $1/\mu$, and the first-come first-served principle is adopted. Denote $\rho = \lambda/\mu$, which is the utilization level of the queueing model.

A special case where $S = 0$ has been studied by Guo and Hassin (2011). In this special case, production is activated when N customer orders have accumulated and is shut down when all the customers are served. Guo and Hassin (2011) model this case by a vacation queue with N-policy and obtain the equilibrium effective arrival rates. If $S > 0$, then the production stops only if S units of inventory have accumulated. We can still use a vacation queue to model such a production system by extending the analysis to allow negative queue lengths.

The queue length is interpreted differently depending on whether it is positive or negative. A positive queue length represents the number of customer orders that are waiting for the product, while a negative queue length represents the inventory level of the finished product. The queue length is bounded from below by $-S$ because the production stops immediately when there are S units of inventory. On the other hand, the queue length may be greater than N. Note that the production server must be shut down when there are S units of inventory, and it must be activated when the queue length is N or higher. However, both statuses are possible when the queue length is within $[-S+1, N-1]$. Figure 1 depicts the transition diagram of the system under this two-critical-number policy.

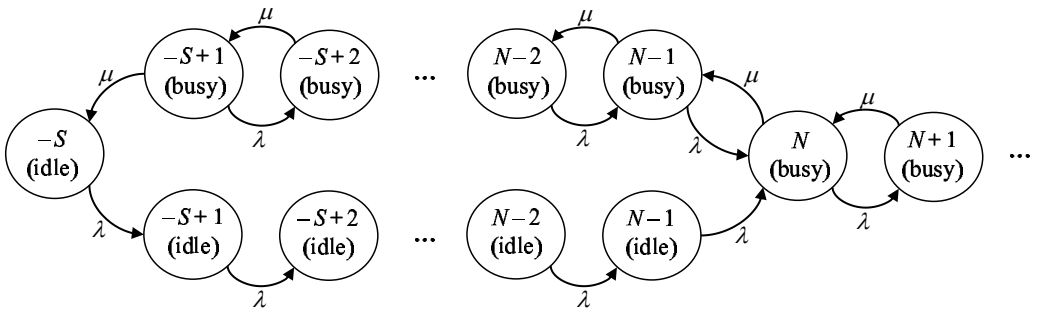


Figure 1: The transition diagram.

This model can be regarded as an $(N + S)$-vacation queue, and the steady-state probabilities can be readily obtained from the known literature result on such a queue (see, Jain *et al.* 2007, pp. 52–53). However, the important performance measures such as the expected waiting time, the average queue length, and the average inventory level have to be derived here, as shown in the following subsections.

2.1.1 Expected waiting time

Given an effective arrival rate λ and the values of the control parameters N and S, the following lemma provides the expected waiting time of a customer.

Lemma 1 *Given an effective arrival rate λ, where $0 < \lambda < \mu$, the expected waiting time of a customer is*

$$W(\lambda, N, S) = \begin{cases} \frac{N}{N+S}\left(\frac{N-1}{2\lambda} + \frac{1}{\mu-\lambda}\right) + \frac{\lambda[1-(\lambda/\mu)^S]}{(N+S)(\mu-\lambda)^2}, & \text{if } N \geq 2; \\ \frac{\mu[(\lambda/\mu)^{-N+1}-(\lambda/\mu)^{S+1}]}{(N+S)(\mu-\lambda)^2}, & \text{if } N \leq 1; \end{cases} \tag{1}$$

if the customer chooses to stay for the service.

The following propositions provide some properties of function $W(\lambda, N, S)$.

Proposition 1 *If $N \geq 2$, then $W(\lambda, N, S)$ is strictly convex in λ, for $0 < \lambda < \mu$. If $N \leq 1$, then $W(\lambda, N, S)$ is strictly increasing in λ, for $0 < \lambda < \mu$.*

Proposition 2 *(i) $W(\lambda, N, S)$ is strictly decreasing in S. (ii) $W(\lambda, N, S)$ is strictly increasing in N.*

Remark 1 *Proposition 1 enables us to derive the equilibrium effective arrival rate in the next subsection. Proposition 2 states the relationship between the expected waiting time and the threshold*

values S and N. When $S = 0$, the system becomes an $M/M/1$ queue with N-policy. Clearly, by Lemma 1,

$$W(\lambda, N, 0) = \frac{N-1}{2\lambda} + \frac{1}{\mu - \lambda} \tag{2}$$

for any $N \geq 2$ (it is easy to verify from Lemma 1 that this equation also holds when $N = 1$). Equation (2) is exactly the expected waiting time in a traditional N-policy queue (see, e.g., Guo and Hassin 2011). Proposition 2(i) implies that $W(\lambda, N, S)$ is smaller than $W(\lambda, N, 0)$ when $S > 0$. In other words, there is a reduction in expected waiting time when the inventory carrying feature is introduced to the traditional N-policy queueing system. Proposition 2(ii) is intuitive, because a larger N will make production occur less often and lead to a higher expected customer waiting time.

Proposition 2 shows that to reduce customer expected waiting time, the production manager may either reduce the production threshold N or increase the inventory threshold S. To see which control variable has a greater marginal effect on the expected waiting time, we define

$$\Delta(\lambda, N, S) = W(\lambda, N+1, S) - W(\lambda, N, S-1)$$

for any $S \geq 1$, $N \geq -S + 2$, and $0 < \lambda < \mu$. Note that $\Delta(\lambda, N, S)$ is equal to $W(\lambda, N+1, S) - W(\lambda, N, S)$ less $W(\lambda, N, S-1) - W(\lambda, N, S)$, where $W(\lambda, N+1, S) - W(\lambda, N, S)$ is the increase in expected waiting time if N is adjusted upward by one unit, and $W(\lambda, N, S-1) - W(\lambda, N, S)$ is the increase in expected waiting time if S is adjusted downward by one unit. Thus, $\Delta(\lambda, N, S)$ equals the difference in the marginal effect for the control variables. The following proposition states that $\Delta(\lambda, N, S)$ is always positive.

Proposition 3 *$\Delta(\lambda, N, S) > 0$ for any $S \geq 1$, $N \geq -S + 2$, and $0 < \lambda < \mu$.*

Proposition 3 implies that adjusting N has a greater marginal effect on the expected waiting time than adjusting S. This conclusion is insightful: In a quick-response industry such as the fashion industry, to reduce customer waiting time, managers should put the first priority on initiating production more aggressively (i.e., reducing N) instead of increasing inventory (i.e., increasing S).

2.1.2 Average inventory level, queue length, and length of production cycle

To determine the expected average inventory level and the expected queue length, we refer to the time period between two consecutive initiations of the production server as a production cycle. Let I be the expected inventory level (i.e., negative queue length) and L be the expected number of

waiting customers (i.e., positive queue length) at steady state. Let T be the expected duration of a production cycle. Let T_{busy} and T_{idle} be the expected duration that the server is busy and idle, respectively, in one production cycle. Then, $T = T_{\text{busy}} + T_{\text{idle}}$. We have the following lemma.

Lemma 2 *Given any $0 < \rho < 1$, $S \geq 0$, and $N \geq -S + 1$, (i) the expected average inventory level in a production cycle equals*

$$I = \begin{cases} \frac{1}{N+S}\left[\frac{S(S+1)}{2} + \frac{\rho^2(1-\rho^S)}{(1-\rho)^2} - \frac{S\rho}{1-\rho}\right], & \text{if } N \geq 2; \\ \frac{S-N+1}{2} + \frac{\rho^2(\rho^{-N}-\rho^S)}{(N+S)(1-\rho)^2} - \frac{\rho}{1-\rho}, & \text{if } N \leq 1; \end{cases}$$

and (ii) the expected average number of waiting customers in a production cycle equals

$$L = \begin{cases} \frac{1}{N+S}\left[\frac{N(N-1)}{2} + \frac{\rho^2(1-\rho^S)}{(1-\rho)^2} + \frac{N\rho}{1-\rho}\right], & \text{if } N \geq 2; \\ \frac{\rho^2(\rho^{-N}-\rho^S)}{(N+S)(1-\rho)^2}, & \text{if } N \leq 1. \end{cases}$$

Note that T_{idle} is the sum of $N + S$ interarrival times of a Poisson process with rate λ. We refer to the time period from the moment when the queue length is k till the moment that the queue length drops to $k - 1$ for the first time as a "1-busy period," where $k \geq -S + 1$ (note that a negative queue length represents a positive inventory level). Then, T_{busy} is the total time for the server to spend N 1-busy periods to clean up the waiting queue and S 1-busy periods to build up the required inventory. Thus, T_{busy} is the sum of $N + S$ stochastically identical 1-busy periods of an $M/M/1$ queue with arrival rate λ and service rate μ. Hence,

$$T_{\text{idle}} = \frac{N+S}{\lambda} = \frac{N+S}{\mu\rho}$$

and

$$T_{\text{busy}} = \frac{N+S}{\mu-\lambda} = \frac{N+S}{\mu(1-\rho)}, \tag{3}$$

which imply that

$$T = \frac{N+S}{\mu\rho} + \frac{N+S}{\mu(1-\rho)} = \frac{N+S}{\mu\rho(1-\rho)}. \tag{4}$$

Remark 2 *From (4), we observe that the expected cycle length is a function of the $N + S$. This implies that if the customer arrival rate is fixed, increasing N has the same impact on the production setup frequency as increasing S.*

Lemma 2 and equations (3)–(4) provide the expressions for I, L, T_{busy}, and T. These expressions will be used for deriving the cost function in Section 3.

2.2 The equilibrium effective arrival rate

The foregoing production model assumes that the demand process is Poisson with effective arrival rate λ. However, this effective arrival rate λ is determined by customers' decentralized decision of whether "to purchase or not to purchase." Since our production system can be modeled as a vacation queue, we can view the customers' purchasing decision as a queueing decision of whether "to queue or not to queue." When a customer makes a queueing decision, she needs to take the expected waiting time into consideration, which, in turn, is affected by how other customers make such a decision. Such a gaming behavior among customers results in an equilibrium, in which a customer's queueing decision is optimal when other customers' queueing decisions are given. In this subsection, we consider customer equilibrium queueing decision and derive the corresponding equilibrium effective arrival rate.

The setting for the queueing game is as follows. Potential customers arrive according to a Poisson process with rate Λ. Upon arrival, customers estimate the waiting time and then make decentralized decisions on whether to place an order or to leave without purchase, depending on their utility value. The queue length and server status are unobservable, a setting consistent with the applications described in Section 1 where customers order products online. Customers make a decision based on their perception of the system's long-run performance measure, i.e., the expected delay. Such a perception may derive from their past experience of patronizing the product or word-of-mouth effect.

Customers are identical and they have a linear reward-cost function. Let R be the reward from receiving the product and W be the expected waiting time if he/she chooses to place an order. The customer's utility of placing his/her order is given by

$$U = R - \theta W,$$

where θ is the delay-sensitivity parameter, representing how much a customer dislikes waiting. A customer will place an order if his/her utility is nonnegative, i.e., $U \geq 0$.

In a regular $M/M/1$ queueing system, as more customers decide to join, the system becomes more congested and hence the joining incentive for a tagged customer is less. Such a behavior is called avoid-the-crowd (ATC) in Hassin and Haviv (2003). What is interesting here is that when more customers decide to join, the expected waiting time for a tagged customer may drop. The reason is that as more customers decide to join, the production system is less likely to become idle, and hence the chance for an incoming customer to get the product immediately is higher. In other

words, as more customers decide to join, it can be beneficial for a tagged customer to join as well, exhibiting the follow-the-crowd (FTC) behavior. According to Hassin and Haviv (2003), such an FTC behavior is often associated with multiple equilibria, as is indeed the case demonstrated in the following analysis.

Since customers are assumed identical, we can consider a symmetric equilibrium in which every customer adopts the same strategy. A customer's queueing strategy can be represented by his/her joining probability, denoted as α, where $0 \leq \alpha \leq 1$. The effective arrival rate is then $\lambda = \alpha\Lambda$. There could exist pure- or mixed-strategy equilibria for such a queueing game. To find the equilibrium, we consider a tagged customer's best response, given other customers' strategy. There exist three possible cases.

First, if everybody else balks and the expected utility for a tagged customer is negative, that is,

$$R - \theta W(0, N, S) < 0, \tag{5}$$

where $W(0, N, S) \equiv \lim_{\lambda \to 0^+} W(\lambda, N, S)$, then the best response for the tagged customer is to balk also. In this case, "all balk" is a (pure-strategy) equilibrium, and zero is the corresponding equilibrium effective arrival rate. We can further examine the sufficient condition (5) under three production-activation policies.

(i) Consider the policy with $N \geq 2$, which means that the system cannot be activated from idleness if the number of waiting customers is less than 2. In this case, $W(0, N, S) = +\infty$, and hence condition (5) always holds. This is intuitive: when everybody else balks, the system cannot be activated, and hence it is optimal for the tagged customer to balk too.

(ii) Consider the policy with $N = 1$, which means that the server becomes idle if nobody is waiting in the queue but will get back to work when the tagged customer joins the empty queue. In this case, $W(0, 1, S) = 1/[(S + 1)\mu]$. Hence, condition (5) holds if and only if $R - \theta/[(S + 1)\mu] < 0$, or equivalently, $S < \frac{\theta}{R\mu} - 1$.

(iii) Consider the policy with $N \leq 0$, which means that as long as the inventory level drops below $-N$, the production system will start to work. In this case, $W(0, N, S) = 0$, and condition (5) never holds.

In summary, zero is an equilibrium effective arrival rate if (i) $N \geq 2$, or (ii) $N = 1$ and $S < \frac{\theta}{R\mu} - 1$.

Second, if everybody else joins and the expected utility for a tagged customer is nonnegative, that is,

$$R - \theta W(\Lambda, N, S) \geq 0, \tag{6}$$

then the best response for the tagged customer is to join as well. In this case, "all join" is a (pure-strategy) equilibrium, and the corresponding equilibrium effective arrival rate is Λ.

Third, if everybody else joins with a probability α, where $0 < \alpha < 1$, and the expected utility for a tagged customer is zero, that is,

$$R - \theta W(\alpha\Lambda, N, S) = 0, \tag{7}$$

then it is indifferent for the tagged customer to join or to balk. Hence, "joining with a probability α" is the best response for the tagged customer too. In this case, "joining with a probability α" is a mixed-strategy equilibrium, and the corresponding equilibrium effective arrival rate is $\lambda = \alpha\Lambda$.

Sufficient conditions (6) and (7) determine the positive equilibrium effective arrival rates. Condition (7) can be alternatively written as

$$R - \theta W(\lambda, N, S) = 0,$$

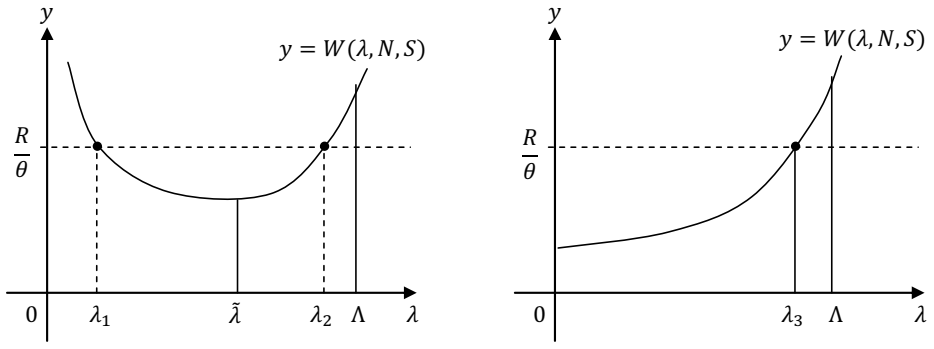
or equivalently,

$$W(\lambda, N, S) = \frac{R}{\theta}, \tag{8}$$

where $W(\lambda, N, S)$ is given by (1). To examine the details of these positive equilibria, we again consider two different cases: $N \geq 2$ and $N \leq 1$.

We first consider the case where $N \geq 2$. From Proposition 1, $W(\lambda, N, S)$ is strictly convex in λ when $0 < \lambda < \mu$. In addition, $\lim_{\lambda \to 0^+} W(\lambda, N, S) = \lim_{\lambda \to \mu^-} W(\lambda, N, S) = +\infty$. Hence, for any given N and S, function $W(\lambda, N, S)$ has a unique minimum $\tilde{\lambda}(N, S)$, or $\tilde{\lambda}$ for simplicity, between 0 and μ. Furthermore, equation (8) has at most two roots. Let $\lambda_1(N, S)$ and $\lambda_2(N, S)$, or λ_1 and λ_2 for simplicity, be the two roots of (8), if exist, where $0 < \lambda_1 < \lambda_2 < \mu$; see Figure 2(a) for illustration. By analyzing $\tilde{\lambda}$, λ_1, and λ_2, we can obtain the positive equilibrium effective arrival rate(s), as stated in the following proposition.

Proposition 4 *For any given $S \geq 0$ and $N \geq 2$, (i) if $W(\tilde{\lambda}, N, S) > R/\theta$, then there exist no positive equilibrium effective arrival rates; (ii) if $W(\tilde{\lambda}, N, S) = R/\theta$, then there exists one positive equilibrium effective arrival rate $\tilde{\lambda}$ if $\tilde{\lambda} \leq \Lambda$, and there exist no positive equilibrium effective arrival*



(a) $N \geq 2$ (b) $N \leq 1$

Figure 2: Expected waiting time versus effective arrival rate.

rates if $\tilde{\lambda} > \Lambda$*; and (iii) if* $W(\tilde{\lambda}, N, S) < R/\theta$*, then there exist two positive equilibrium effective arrival rates* λ_1 *and* $\min\{\lambda_2, \Lambda\}$ *if* $\lambda_1 \leq \Lambda$ *(which reduce to one equilibrium rate if* $\lambda_1 = \Lambda$*), and there exist no positive equilibrium effective arrival rates if* $\lambda_1 > \Lambda$*.*

Proposition 4 is similar to the equilibrium effective arrival rate results presented in Proposition 1 of Guo and Hassin (2011). Note that the vacation queue with N-policy analyzed by Guo and Hassin is a special case of our model with $S = 0$.

Next, we consider the case where $N \leq 1$. From Proposition 1, $W(\lambda, N, S)$ is strictly increasing in λ when $0 < \lambda < \mu$. Hence, equation (8) has at most one root between 0 and μ. Let λ_3 denote this root, if exists; see Figure 2(b) for illustration. We have the following proposition.

Proposition 5 *(i) If* $N = 1$*, then there exists one positive equilibrium effective arrival rate* $\min\{\lambda_3, \Lambda\}$ *if* $S > \frac{\theta}{R\mu} - 1$*, and there exists no positive equilibrium effective arrival rate if* $S \leq \frac{\theta}{R\mu} - 1$*. (ii) If* $N \leq 0$*, then there always exists one positive equilibrium effective arrival rate* $\min\{\lambda_3, \Lambda\}$*.*

Through the above equilibrium analysis, we can see that at most three equilibria could exist in the customer queueing game: zero effective arrival rate and two positive equilibrium effective arrival rates. We will not consider the zero equilibrium effective arrival rate because, in real life, production managers can use short-term price promotions to attract customers to avoid being stuck in such an equilibrium. Between the two positive equilibrium effective arrival rates, we consider the (locally) stable one. The stability here means that, when a small disruption happens to the equilibrium effective arrival rate, the equilibrium will not diverge away. Consider the two positive equilibria in the case where $N \geq 2$ and $\lambda_1 < \lambda_2 \leq \Lambda$ as depicted in Figure 2(a). The larger one λ_2 is stable, while the smaller λ_1 is unstable. To see that, note that the expected waiting time

$W(\lambda, N, S)$ is decreasing in λ at the point $\lambda = \lambda_1$; hence, a slight increase over λ_1 will reduce the expected waiting time and attract even more arrivals, making the equilibrium diverge away. However, this is not the case for λ_2. Therefore, in the following analysis, we will focus on the positive and stable equilibrium effective arrival rate.

For the case where $N \geq 2$ ($N = 1$), by Proposition 4 (Proposition 5(i)), the positive stable equilibrium effective arrival rate, if exists, can only be λ_2 (λ_3) or Λ, whichever is smaller. For the case where $N \leq 0$, by Proposition 5(ii), a positive stable equilibrium effective arrival rate always exists and is equal to λ_3 or Λ, whichever is smaller. This implies that the positive stable equilibrium effective arrival rate, if exists, must be unique. To use a uniform notation, we define

$$\lambda_+ = \begin{cases} \lambda_2, & \text{if } N \geq 2; \\ \lambda_3, & \text{if } N \leq 1. \end{cases}$$

With this definition, the positive stable equilibrium effective arrival rate, if exists, is equal to $\min\{\lambda_+, \Lambda\}$.

Note that the value of λ_+ depends on the control parameters as well as the system parameters, and it possesses the following property.

Proposition 6 *λ_+ is strictly decreasing in θ and N, and is strictly increasing in S.*

3 Determining the Optimal Control Values

In this section, we consider the system control decision with strategic customers. We aim to determine the optimal values of the control variables N and S such that the expected cost rate of the system is minimized. As is commonly assumed in the inventory management literature, the expected cost rate of the system consists of four cost components: setup and operating cost, inventory holding cost, backordering cost, and lost-sales penalty cost. We list the cost notations as follows.

K = fixed setup cost for each initiation of production;

c = system operating cost per unit time when the server is active;

h = inventory holding cost per unit of inventory per unit time;

θ = backordering cost per waiting customer per unit time;

p = lost-sales penalty per unit.

Note that the delay-sensitivity parameter θ in the customer utility model represents the customer's waiting cost per unit time. This parameter can also be viewed as a backordering cost in the system.

Using renewal theory (Ross 1996, p. 133), the expected total setup and operating cost per unit time can be expressed as $(K + cT_{\text{busy}})/T$. The expected inventory holding cost rate is hI, and the expected backordering cost rate for having customers waiting in the system is θL. Also, the expected lost-sales penalty cost rate is $p(\Lambda - \lambda)$, where $\Lambda - \lambda$ is the rate of customer loss. Denote $\hat{\rho} = \Lambda/\mu$, then $p(\Lambda - \lambda) = p\mu(\hat{\rho} - \rho)$. Thus, the expected cost of the system per unit time is

$$\Gamma = \frac{K + cT_{\text{busy}}}{T} + hI + \theta L + p\mu(\hat{\rho} - \rho). \tag{9}$$

The expected cost rate function Γ can be expressed in terms of ρ, N, and S, as shown in the following proposition.

Proposition 7 *Given any* $0 < \rho < 1$, $S \geq 0$, *and* $N \geq -S+1$, *the expected cost rate of the system equals*

$$\Gamma(\rho, N, S) = \begin{cases} \mu\rho\frac{K(1-\rho)}{N+S} + c\rho + \frac{h}{N+S}\left[\frac{S(S+1)}{2} + \frac{\rho^2(1-\rho^S)}{(1-\rho)^2} - \frac{S\rho}{1-\rho}\right] \\ \quad + \frac{\theta}{N+S}\left[\frac{N(N-1)}{2} + \frac{\rho^2(1-\rho^S)}{(1-\rho)^2} + \frac{N\rho}{1-\rho}\right] + p\mu(\hat{\rho} - \rho), & \text{if } N \geq 2; \\ \mu\rho\frac{K(1-\rho)}{N+S} + c\rho + h\left[\frac{S-N+1}{2} + \frac{\rho^2(\rho^{-N} - \rho^S)}{(N+S)(1-\rho)^2} - \frac{\rho}{1-\rho}\right] \\ \quad + \frac{\theta\rho^2(\rho^{-N} - \rho^S)}{(N+S)(1-\rho)^2} + p\mu(\hat{\rho} - \rho), & \text{if } N \leq 1. \end{cases}$$

The closed-form cost function presented in Proposition 7 and the convexity/monotonicity of function $W(\lambda, N, S)$ presented in Proposition 1 enable us to search for the optimal values of N and S easily. Denote

$$W_\lambda(\lambda, N, S) = \begin{cases} \frac{N}{N+S}\left[\frac{-N+1}{2\lambda^2} + \frac{1}{(\mu-\lambda)^2}\right] + \frac{1-(S+1)(\lambda/\mu)^S}{(N+S)(\mu-\lambda)^2} + \frac{2\lambda[1-(\lambda/\mu)^S]}{(N+S)(\mu-\lambda)^3}, & \text{if } N \geq 2; \\ \frac{2\mu[(\lambda/\mu)^{-N+1} - (\lambda/\mu)^{S+1}]}{(N+S)(\mu-\lambda)^3} + \frac{(-N+1)(\lambda/\mu)^{-N} - (S+1)(\lambda/\mu)^S}{(N+S)(\mu-\lambda)^2}, & \text{if } N \leq 1; \end{cases}$$

which is the first-order partial derivative of $W(\lambda, N, S)$ with respect to λ. For any given values of N and S, we denote $\lambda^e(N, S)$ as the positive stable equilibrium effective arrival rate, if exists; otherwise, we let $\lambda^e(N, S) = 0$ (note: it is easy to check that if no positive stable equilibrium effective arrival rate exists, then zero will be an equilibrium effective arrival rate). We can determine the value of $\lambda^e(N, S)$ via the following procedure.

Procedure P(N, S):

Step 1. If $N = 1$ and $S \leq \frac{\theta}{R\mu} - 1$, set $\lambda^e(N, S) := 0$ and stop.

If $N \leq 0$ or ($N = 1$ and $S > \frac{\theta}{R\mu} - 1$), set $\lambda_{\min} := 0$, go to Step 3.

If $N \geq 2$, go to Step 2.

Step 2. Search for $\lambda_{\min} \in [0, \mu)$ such that $W_\lambda(\lambda_{\min}, N, S) = 0$.

Step 3. If $W(\lambda_{\min}, N, S) > R/\theta$, set $\lambda^e(N, S) := 0$.

If $W(\Lambda, N, S) \leq R/\theta$, set $\lambda^e(N, S) := \Lambda$.

Otherwise, search for $\lambda^e(N, S) \in [\lambda_{\min}, \Lambda)$ such that $W(\lambda^e(N, S), N, S) = R/\theta$.

In procedure $\mathbf{P}(N, S)$, Step 2 searches for the minimum of $W(\lambda, N, S)$. This step is only applicable to the case where $N \geq 2$. Step 3 first considers the special cases where $W(\lambda_{\min}, N, S) > R/\theta$ or $W(\Lambda, N, S) \leq R/\theta$ (see Section 2.2 for a discussion of these two cases). If the conditions of these two special cases are not met, then it searches for the intersection of the curve $y = W(\lambda, N, S)$ and the horizontal line $y = R/\theta$. The searches in Step 2 and 3 can be done via some standard search technique such as binary search or Newton-Raphson method. The procedure sets $\lambda^e(N, S)$ to either λ_+, Λ, or zero.

Let N^* and S^* denote the optimal values of N and S, respectively. The following proposition provides us with upper bounds on N^* and S^*.

Proposition 8 *Let $\hat{\Gamma}$ be an upper bound on the optimal expected cost rate of the system. Let γ be any value such that $\gamma > 4$. Let*

$$\bar{N} = \left\lceil \max\left\{ \frac{4\hat{\Gamma}}{\theta}, \frac{8\gamma\hat{\Gamma}^2}{h\theta(\gamma - 4)} \right\} \right\rceil \quad \textit{and} \quad \bar{S} = \left\lfloor \frac{\gamma\hat{\Gamma}}{h} \right\rfloor.$$

Then, $\bar{N} \geq N^*$ and $\bar{S} \geq S^*$.

In Proposition 8, $\hat{\Gamma}$ can be obtained by arbitrarily selecting a pair of integers $(\hat{N}, \hat{S})$ such that $\hat{S} \geq 0$ and $\hat{N} \geq -\hat{S} + 1$, and letting $\hat{\Gamma} = \Gamma(\lambda^e(\hat{N}, \hat{S})/\mu, \hat{N}, \hat{S})$. The quantity γ can be any value greater than 4. A larger γ will lead to a smaller $\bar{N}$ but a larger $\bar{S}$, while a smaller γ will lead to a smaller $\bar{S}$ but a larger $\bar{N}$. The optimal threshold values N^* and S^* can be obtained via the following exhaustive search algorithm. In this algorithm, $\bar{N}$ and $\bar{S}$ are upper bounds on N and S, which can be obtained using the formulas given by Proposition 8. Variable Γ^* stores the best possible value of $\Gamma(\lambda^e(N, S)/\mu, N, S)$ obtained by the search.

Algorithm $\mathbf{A}_{\text{optimal}}$:

Step 1. Set $\Gamma^* := \hat{\Gamma}$, where $\hat{\Gamma}$ is any upper bound on the optimal expected cost rate.

Step 2. For $S := 0, 1, \ldots, \bar{S}$ and $N := -S + 1, -S + 2, \ldots, \bar{N}$:

(i) Calculate $\lambda^e(N, S)$ using procedure $\mathbf{P}(N, S)$. Set $\rho := \lambda^e(N, S)/\mu$.

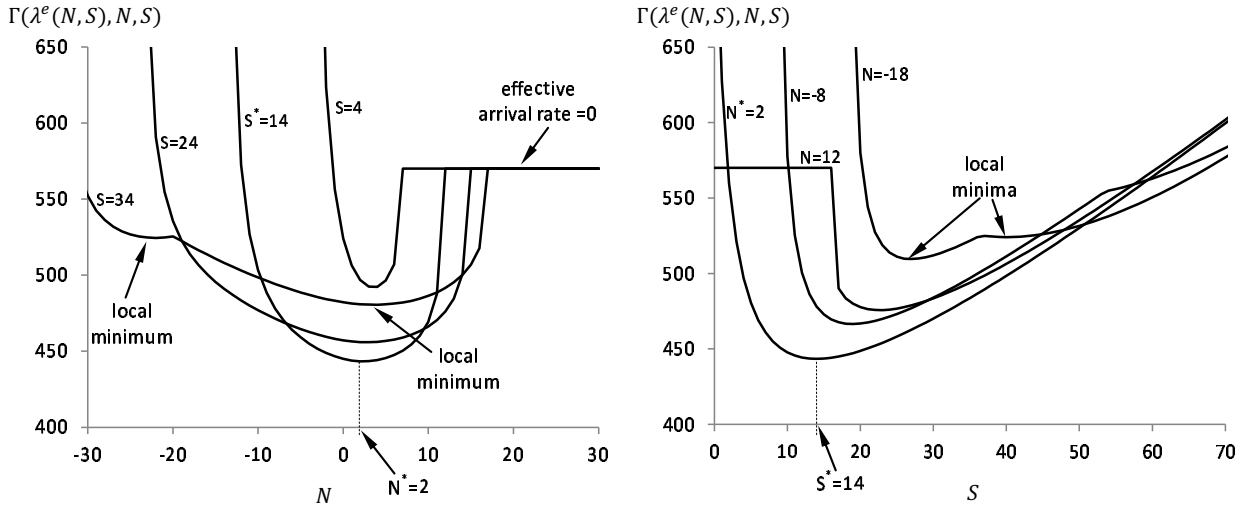
(ii) If $\Gamma(\rho, N, S) < \Gamma^*$, then set $N^* := N$, $S^* := S$, and $\Gamma^* := \Gamma(\rho, N, S)$.

Remark 3 *By Proposition 6, $\lambda_+(N,S)$ is strictly decreasing in N and strictly increasing in S. Hence, $\lambda^e(N,S) < \lambda^e(N-1,S)$ and $\lambda^e(N,S) > \lambda^e(N,S-1)$. Using this property, procedure $\mathbf{P}(N,S)$ can be implemented slightly more efficiently by restricting the scope of the search of $\lambda^e(N,S)$. Instead of searching for $\lambda^e(N,S)$ between 0 and μ, it is sufficient to conduct the search between $\lambda^e(N,S-1)$ and* $\min\{\mu, \lambda^e(N-1,S)\}$.

4 Numerical Study

In this section we conduct a numerical study to illustrate certain characteristics of our model. This numerical study is based on an example with the following parameter setting: $\mu = 10$ units/day, $R = \$20$/unit, $h = \$10$/unit/day, $\theta = \$40$/unit/day, $p = \$60$/unit, $c = \$200$/day, $K = \$400$ per setup, and $\Lambda = 9.5$ units/day (i.e., $\hat{\rho} = 0.95$). We have repeated our numerical study with many other parameter settings and find that our findings are representative.

We first consider the relationship between (N,S) and expected cost rate $\Gamma(\lambda^e(N,S)/\mu, N, S)$. Figure 3(a) depicts $\Gamma(\lambda^e(N,S)/\mu, N, S)$ versus N for four different values of S, namely $S = 4, 14, 24, 34$. Figure 3(b) depicts $\Gamma(\lambda^e(N,S)/\mu, N, S)$ versus S for four different values of N, namely $N = -18, -8, 2, 12$. We observe that both functions $\Gamma(\lambda^e(\cdot,S)/\mu, \cdot, S)$ and $\Gamma(\lambda^e(N,\cdot)/\mu, N, \cdot)$ may have more than one local minimum. This justifies the use of an exhaustive search of N^* and S^* in algorithm $\mathbf{A}_{\text{optimal}}$. In this example, $(N^*, S^*) = (2, 14)$. From Figure 3(a), we observe that the



(a) $\Gamma(\lambda^e(N,S),N,S)$ versus N

(b) $\Gamma(\lambda^e(N,S),N,S)$ versus S

Figure 3: Expected cost rate versus N and S.

bottom of the curve $\Gamma(\lambda^e(\cdot, S^*)/\mu, \cdot, S^*)$ is quite flat. Similarly, from Figure 3(b), we observe that the bottom of the curve $\Gamma(\lambda^e(N^*, \cdot)/\mu, N^*, \cdot)$ is quite flat. Thus, the optimal expected cost rate is not very sensitive to a change in N and S when (N, S) is close to the optimum. Hence, it does not incur a significant increase in cost if the system manager chooses to deviate the control parameter values from (N^*, S^*) slightly. However, the increase in cost may be substantial if the deviation is large.

Next, we consider the relationship between ρ and the expected waiting time. Understanding this relationship will help us understand the demand sensitivity with respect to a change of N and S. Note that the expected waiting time function in (1) can be expressed in terms of ρ:

$$\bar{W}(\rho, N, S) = \begin{cases} \frac{1}{\mu}\left[\frac{N}{N+S}\left(\frac{N-1}{2\rho} + \frac{1}{1-\rho}\right) + \frac{\rho(1-\rho^S)}{(N+S)(1-\rho)^2}\right], & \text{if } N \geq 2; \\ \frac{1}{\mu} \cdot \frac{\rho^{-N+1} - \rho^{S+1}}{(N+S)(1-\rho)^2}, & \text{if } N \leq 1. \end{cases}$$

Figure 4 depicts this function for $S = 10$ and three different values of N, namely $N = -1, 2, 5$. We observe that the expected waiting times have flat bottoms and increase sharply when ρ is close

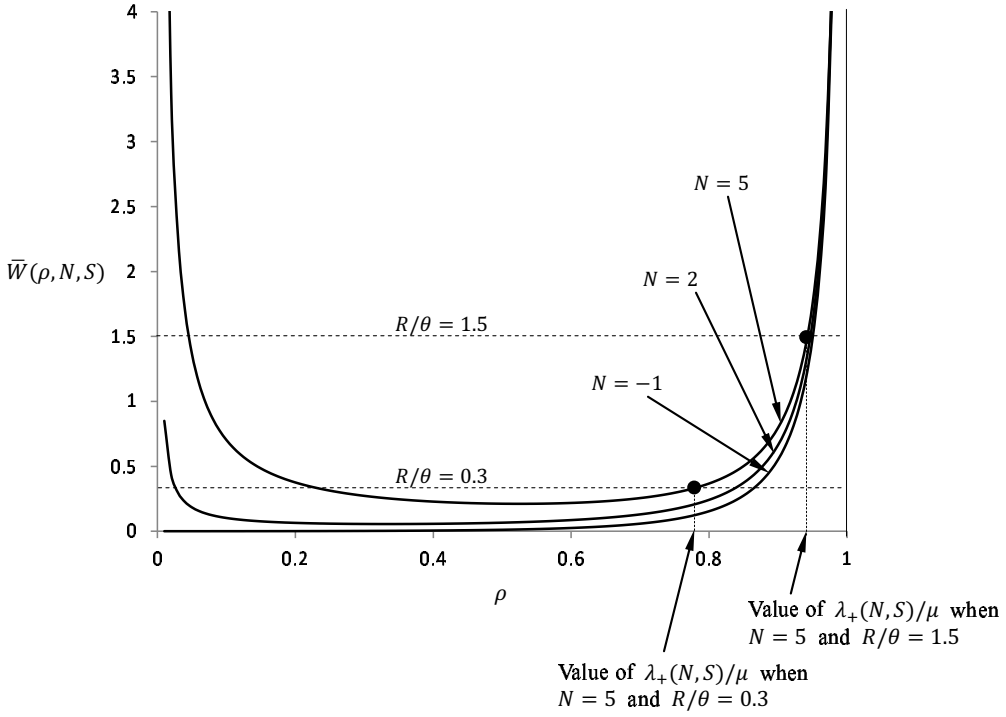
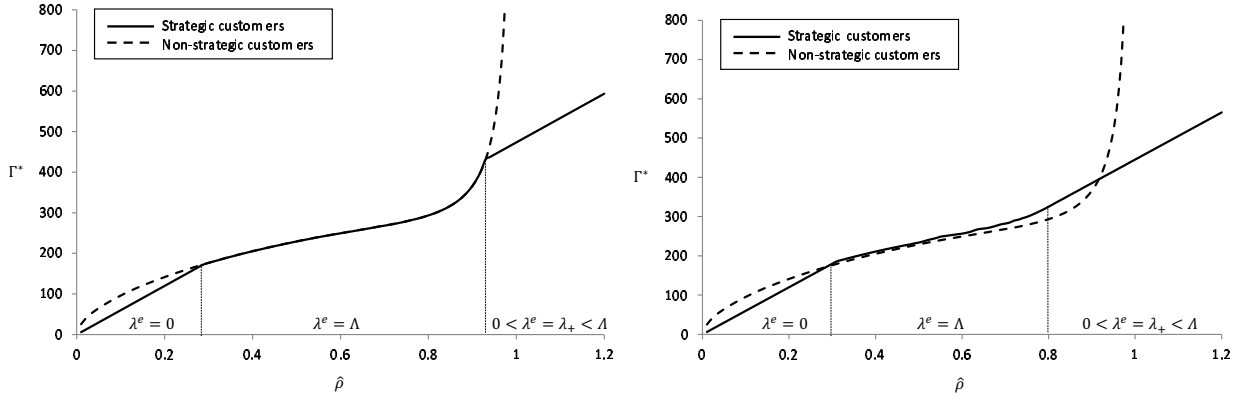


Figure 4: Expected waiting time versus ρ.

to 1. This pattern is generally true for other combinations of N and S. Note that the stable equilibrium $\lambda_+(N, S)$ is the larger root of the equation "$W(\lambda, N, S) = R/\theta$." If R/θ is large (e.g., when $R/\theta = 1.5$), then the solution $\lambda_+(N, S)$ is close to μ. As shown in Figure 4, changing N

brings little change in $\lambda_+(N,S)$, suggesting that $\lambda_+(N,S)$ is rather insensitive to a change in N when R/θ is large. If R/θ is small (e.g., when $R/\theta = 0.3$), then $\lambda_+(N,S)$ is attained near the bottom of the curve, and as shown in Figure 4, $\lambda_+(N,S)$ is rather sensitive to a change in N. Note that R/θ measures the maximal amount of time that a customer is willing to wait. Therefore, when customers' maximal willingness-to-wait time is long (short), the effective arrival rate is insensitive (sensitive) to a change of N. Similarly, we find that the impact of S on the effective arrival rate is also closely related to customers' maximal willingness-to-wait time. From Figure 4, we also observe that for $N \geq 2$, $\bar{W}(\rho,N,S)$ tends to infinity as ρ tends to zero. This is because the service does not start until N or more customer orders have accumulated. Thus, when the arrival rate is very low, an arriving customer has to wait for a long time before getting served.

We now consider the impact of demand traffic $\hat{\rho}$ on the optimal solution by varying $\hat{\rho}$ within the range $(0, 1.2]$. Figure 5(a) shows the optimal cost rate Γ^* versus $\hat{\rho}$. We observe that Γ^* increases as $\hat{\rho}$ increases. This is because as the demand traffic increases, the company needs to either serve more customers or incur more lost sales, and this leads to an increase of the overall cost of the company. However, this characteristic does not always hold. For example, if the operating cost c and the delay-sensitivity θ are very low and the setup cost K is very high, then a higher demand traffic may enable the production manager to choose a longer production run to avoid the expensive setups, which may lead to a lower overall cost rate.



(a) $R = 20$ (b) $R = 1$

Figure 5: Optimal cost rate versus $\hat{\rho}$.

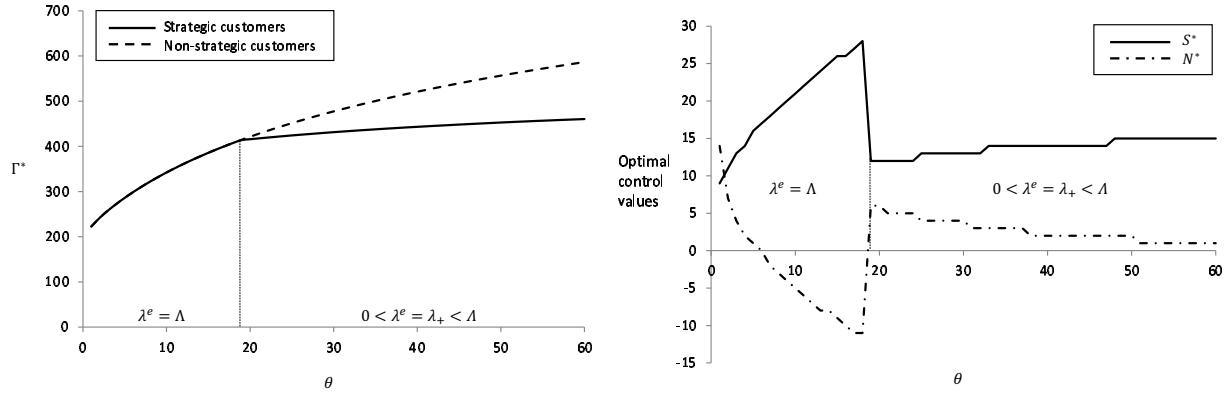
Figure 5(a) also depicts the optimal cost rates for the problem with non-strategic customers, i.e., when ρ is set equal to $\hat{\rho}$ for $0 < \hat{\rho} < 1$. This "non-strategic customer model" can be solved using the algorithm developed by Lee and Srinivasan (1989), where they consider a cost minimization

problem in an $M/G/1$ production inventory system with fixed demand arrival rate. The two control variables they considered are S and $N+S$, which they refer as r. Then, they show that the average system cost rate $C(r, S^*(r))$ is unimodal in r, where $S^*(r)$ is the optimal value of S for a given r, and thus propose a searching algorithm. In the example depicted in Figure 5(a), the case "$\lambda^e = 0$" occurs when $\hat{\rho} \in (0, 0.28]$. When $\hat{\rho}$ is within this interval, the demand traffic is light, and the system with strategic customers sets a low S value to discourage all customers from joining the queue. If the customers are non-strategic, they must join the queue, and the system will have a higher cost than that with strategic customers. The case "$\lambda^e = \Lambda$" occurs when $\hat{\rho} \in (0.28.0.92]$. When $\hat{\rho}$ is within this interval, the demand traffic is sufficiently heavy, and the system with strategic customers admits all customers. The case "$0 < \lambda^e = \lambda_+(N^*, S^*) < \Lambda$" occurs when $\hat{\rho} \in (0.92, 1.2]$. When $\hat{\rho}$ is within this interval, only a portion of the strategic customers will join the queue. Note that for the problem with non-strategic customers, we only consider the situation "$\hat{\rho} < 1$," which is a necessary condition for the existence of a steady state. When $\hat{\rho}$ is close to 1, the optimal cost rate of the non-strategic customer model is substantially higher than that of the strategic customer model. This is because when the demand traffic is heavy, the system with non-strategic customers becomes highly congested. On the other hand, if the customers are strategic, the production manager may set the N and S values in such a way that some of the strategic customers will choose to leave the system. This allows the system to maintain a reasonable cost rate. As shown in Figure 5(a), the situation "$0 < \lambda^e = \lambda_+ < \Lambda$" occurs only when $\hat{\rho}$ is large. This can be explained as follows: Recalling from the discussion of Figure 4, the expected waiting time curves generally have flat bottoms and increase sharply when ρ is close to 1. Hence, unless R/θ is very small, λ_+/μ is typically close to 1. Therefore, the case "$\lambda^e = \lambda_+ < \Lambda$" occurs when Λ/μ (i.e., $\hat{\rho}$) is close to 1 or greater than 1.

Figure 5(a) indicates that in this particular example, the optimal cost rate of the problem with strategic customers is no greater than that with non-strategic customers. However, this may not be always true in general. Consider a modified example with $R = 1$ instead of $R = 20$. Figure 5(b) depicts the optimal cost rates of this modified example, in which the optimal cost rate for the system with strategic customers is higher than that with non-strategic customers in the second interval and the beginning portion of the third interval. This is because in this example the customers' reward from receiving the product is very low. Thus, unless $\hat{\rho}$ is high, the production manager needs to choose a larger N^* and/or a larger S^* to reduce the expected waiting time to attract customers to join the queue. It is worth mentioning that even if all strategic customers will join the queue, it does not necessarily mean that the optimal cost rate will be the same as that in the system with

non-strategic customers.

Next, we consider the impact of the delay-sensitivity parameter θ on the optimal solution by varying θ within $(0, 60]$. Figure 6(a) shows the optimal cost rate Γ^* versus θ, while Figure 6(b) shows the corresponding N^* and S^* values. From Figure 6(a), we observe that Γ^* increases as θ



(a) Γ^* versus θ (b) N^* and S^* versus θ

Figure 6: The optimal solution versus θ.

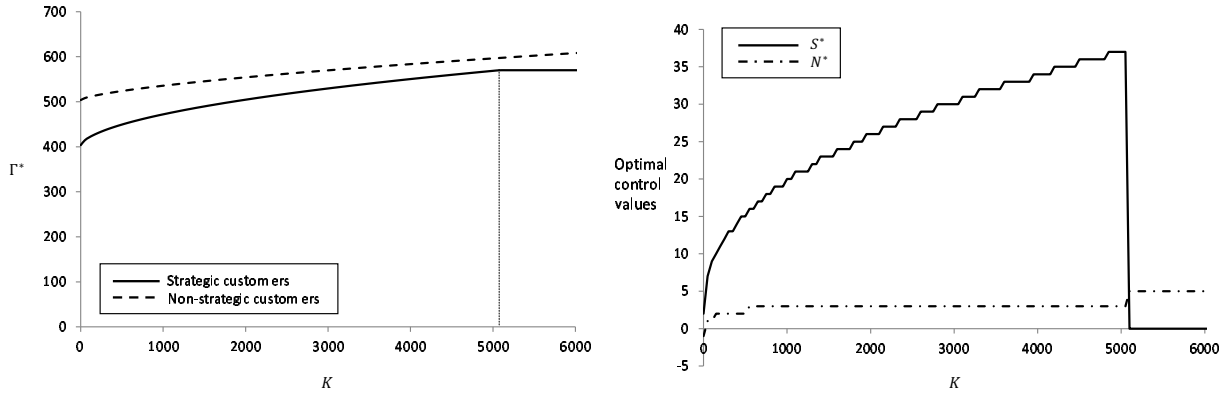
increases. This is because when customers are more sensitive to waiting, the production manager either needs to put in more resources (e.g., keep more inventory, set up the production runs more frequently, etc.) or faces more lost sales.

Figure 6(a) also depicts the optimal cost rates when the customers are non-strategic. Note that in this example, $\hat{\rho} = 0.95$, which, according to Figure 5, leads to a significant difference between the system with strategic customers and the system with non-strategic customers. Figure 6(a) indicates that the difference between these two systems is also affected by the customers' delay-sensitivity θ. (Note: θ is also the backordering cost of the system. Thus, in the system with non-strategic customers, the optimal cost rate is also affected by θ.) When the customers are strategic, the increase in Γ^* is less drastic as θ increases. This is because when customers are strategic and when customers' delay-sensitivity is sufficiently high (i.e., $\theta \geq 19$ in the current example), the production manager will try to discourage some customers from joining the queue by choosing a larger N^* and a smaller S^* (see the sharp increase in N^* and the sharp decrease in S^* in Figure 6(b) when θ reaches 19). In this case, the system is not as crowded as the case with non-strategic customers, and the overall cost rate is lower.

Another interesting observation from Figure 6(b) is that the increasing (decreasing) rate in S^* (N^*) differs significantly before and after the point where $\theta = 19$. When $\theta < 19$, customers are

relatively patient, implying a small backlogging cost. It is optimal to allow all customers to enter the system to avoid the penalty cost for lost sales. Thus, the tradeoff is mainly among setup cost, holding cost, and backlogging cost. As mentioned in Remark 2, the expected cycle length, and hence the setup frequency, is a function of $N+S$. So, the manager can keep $N+S$ relatively stable but increase S and decrease N to reduce the expected waiting time. Therefore, the main concern is the tradeoff between holding cost and backordering (waiting) cost. As θ increases, the weight for backordering becomes heavier and hence the inventory threshold S should be increased while the production threshold N should be decreased. However, when $\theta \geq 19$, customers are relatively sensitive to delay, and it is no longer optimal to keep all customers in the system. In other words, it becomes optimal to have some balking customers (i.e., lost sales) instead of keeping the queue long. In this case, the changes in S^* and N^* are less sensitive to a change in θ.

Next, we consider the impact of the setup cost K on the optimal solution by varying K within $[0, 6000]$. Figure 7(a) shows the optimal cost rate Γ^* versus K, while Figure 7(b) shows the corresponding N^* and S^* values. Figure 7(a) also depicts the optimal cost rates when the customers



(a) Γ^* versus K (b) N^* and S^* versus K

Figure 7: The optimal solution versus K.

are non-strategic. In this example, the optimal cost rate for the case with strategic customers is lower than that for the case with non-strategic customers.

For the strategic customer case, when $K < 5100$, the optimal cost rate increases as K increases. This is because as K increases, economy of scale becomes more important, and therefore the duration of a production run needs to be lengthened. This can be achieved by setting a larger N^* and/or a larger S^*, as depicted in Figure 7(b). When $K \geq 5100$, the optimal cost stays constant as K changes. This is because when the setup cost is too large, the optimal decision of the production

manager is to discourage all customers from joining the queue.

One interesting observation is that when $K < 5100$, the increase in S^* is much more significant than the increase in N^*. There are two reasons for this. First, in our example, $h = 10$ and $\theta = 40$. In other words, we penalize backordering more heavily than inventory holding. Second, the marginal effect of adjusting N has a greater effect on the expected waiting time than adjusting S; see Proposition 3. As K increases, the manager not only needs to increase N^* and S^* to achieve economy of scale, but also needs to take into consideration the strategic behavior of customers. Therefore, the production manager would rather adjust S^* significantly and adjust N^* mildly, such that the expected waiting time can be maintained at a low level to keep customers' interests in joining the queue.

5 Conclusions

In this paper we consider a make-to-stock queue with a two-critical-number policy and delay-sensitive customers. We derive the customers' expected waiting time and equilibrium effective arrival rate, as well as the system's expected cost rate. We develop a search algorithm for the optimal control decision variables and conduct a numerical study.

Our findings offer several interesting managerial insights. We find that customers' expected waiting time is convex in the effective arrival rate when $N \geq 2$, and it is increasing in the effective arrival rate when $N \leq 1$. In the former case, there might be two positive equilibria, and only the larger one is stable. In the latter case, there exists at most one positive equilibrium effective arrival rate, and it is stable. These results are generalizations of those in the (make-to-order) vacation queue system. We also show that a reduction in expected waiting time can be achieved by either reducing the critical number on the waiting customers or increasing the critical number on the cumulated inventory, and the former has a higher marginal impact than the latter.

From our numerical study, we find that when the demand traffic is very light or very heavy, there might exist lost sales, and the optimal cost rate can be significantly lower than that in the case with non-strategic customers. When the demand traffic is moderate, all customers may join the queue, but the cost rate may not necessarily be the same as that in the case with non-strategic customers. We also analyze how the customers' delay-sensitivity parameter θ affects the results. We observe that when the customers are relatively patient, the optimal control variables are rather sensitive to a change in θ. When the customers are impatient, the optimal control variables are

less sensitive to a change in θ. We also find that as the setup cost K increases, the values of the control variables increase, but the increase in S is more significant than the increase in N.

In this study, for simplicity and tractability, we have adopted a standard $M/M/1$ make-to-stock queue. Using an expanded Markov chain construction, a similar analysis can be generalized to an $M/E_k/1$ setting. However, such a generalization is too complex to be included in the current paper, and hence will be left to future research. Another possible extension of the current topic is to consider two competing production systems. Under a competition environment, customers strategically choose a company that offers them higher utility, and if so, the impact of the customers' strategic behavior on system performance might be very different from what have been observed in this paper.

Acknowledgments

The authors thank Professor Michael Pinedo (the department editor), the senior editor, and three anonymous referees for their helpful comments and suggestions. The first author was supported in part by the National Natural Science Foundation of China under grant no. 71501037. The second author was supported in part by the Hong Kong Research Grants Council under grant no. PolyU155045/15B. The third author was supported in part by The Hong Kong Polytechnic University under grant no. 1-BBZL. The fourth author was supported in part by the National Natural Science Foundation of China under grant no. 71390331.

References

Benjaafar, S., J.-P. Gayon, S. Tepe. 2010. Optimal control of a production-inventory system with customer impatience. *Operations Research Letters* **38**(4), 267–272.

Benjaafar, S., M. Elhafsi. 2012. A production-inventory system with both patient and impatient demand classes. *IEEE Transactions on Automation Science and Engineering* **9**(1), 148–159.

Boudali, O., A. Economou. 2013. The effect of catastrophes on the strategic customer behavior in queueing systems. *Naval Research Logistics* **60**(7), 571–587.

Burnetas, A. 2013. Customer equilibrium and optimal strategies in Markovian queues in series. *Annals of Operations Research* **208**, 515–529.

Chen, J., S. Huang, R. Hassin, N. Zhang. 2015. Two backorder compensation mechanisms in inventory systems with impatient customers. *Production and Operations Management* **24**(10), 1640–1656.

Debo, L.G., C. Parlour, U. Rajan. 2012. Signaling quality via queues. *Management Science* **58**(5), 876–891.

Debo, L., S. Veeraraghavan. 2014. Equilibrium in queues under unknown service times and service value. *Operations Research* **62**(1), 38–57.

Dimitrakopoulos, Y., A. Burnetas. 2016. Customer equilibrium and optimal strategies in an $M/M/1$ queue with dynamic service control. *European Journal of Operational Research*, forthcoming.

Economou, A., A. Gómez-Corral, S. Kanta. 2011. Optimal balking strategies in single-server queues with general service and vacation times. *Performance Evaluation* **68**(10), 967–982.

Economou, A., A. Manou. 2013. Equilibrium balking strategies for a clearing queueing system in alternating environment. *Annals of Operations Research* **208**, 489–514.

Federgruen, A., Y.-S. Zheng. 1993. Optimal control policies for stochastic inventory systems with endogenous supply. *Probability in the Engineering and Informational Sciences* **7**(2), 257–272.

Gavish, B., S.C. Graves. 1980. A one-product production/inventory problem under continuous review policy. *Operations Research* **28**(5), 1228–1236.

Gavish, B., S.C. Graves. 1981. Production/inventory systems with a stochastic production rate under a continuous review policy. *Computers & Operations Research* **8**(3), 169–183.

Graves, S.C., J. Keilson. 1981. The compensation method applied to a one-product production/inventory problem. *Mathematics of Operations Research* **6**(2), 246–262.

Guo, P., R. Hassin. 2011. Strategic behavior and social optimization in Markovian vacation queues. *Operations Research* **59**(4), 986–997.

Guo, P., R. Hassin. 2012. Strategic behavior and social optimization in Markovian vacation queues: The case of heterogeneous customers. *European Journal of Operational Research* **222**(2), 278–286.

Guo, P., R. Hassin. 2015. Equilibrium strategies for placing duplicate orders in a single server queue. *Operations Research Letters* **43**(3), 343–348.

Guo, P., Q. Li. 2013. Strategic behavior and social optimization in partially-observable Markovian vacation queues. *Operations Research Letters* **41**(3), 277–284.

Guo, P., W. Sun, Y. Wang. 2011. Equilibrium and optimal strategies to join a queue with partial information on service times. *European Journal of Operational Research* **214**(2), 284–297.

Guo, P., Z.G. Zhang. 2013. Strategic queueing behavior and its impact on system performance in service systems with congestion-based staffing policy. *Manufacturing & Service Operations Management* **15**(1), 118–131.

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

Jain, J.L., S.G. Mohanty, W. Böhm. 2007. *A Course on Queueing Models.* CRC Press, Boca Raton, FL.

Lee, H.-S., M.M. Srinivasan. 1989. The continuous review (s, S) policy for production/inventory systems with Poisson demands and arbitrary processing times. Technical Report 87-33, Department of Industrial and Operations Engineering, The University of Michigan.

Li, L. 1992. The role of inventory in delivery-time competition. *Management Science* **38**(2), 182–197.

Manou., A., A. Economou, F. Karaesmen. 2014. Strategic customers in a transportation station: When is it optimal to wait? *Operations Research* **62**(4), 910–925.

Naor, P. 1969. The regulation of queue size by levying tolls. *Econometrica* **37**(1), 15–24.

Ross, S.M. 1996. *Stochastic Processes,* 2nd Edition. Wiley, New York.

Shi, J., M.N. Katehakis, B. Melamed, Y. Xia. 2014. Production-inventory systems with lost sales and compound Poisson demands. *Operations Research* **62**(5), 1048–1063.

So, K.C., J.-S. Song. 1998. Price, delivery time guarantees and capacity selection. *European Journal of Operational Research* **111**(1), 28–49.

Song, D.-P. 2009. Stability and optimization of a production inventory system under prioritized base-stock control. *IMA Journal of Management Mathematics* **20**(1), 59–79.

Srinivasan, M.M., H.-S. Lee. 1991. Random review production/inventory systems with compound Poisson demands and arbitrary processing times. *Management Science* **37**(7), 813–833.

Sun, W., P. Guo, N. Tian. 2010. Equilibrium threshold strategies in observable queueing systems with setup/closedown time. *Central European Journal of Operations Research* **18**(3), 241–268.

Van Foreest, N.D., J. Wijngaard. 2014. On optimal policies for production-inventory systems with compound Poisson demand and setup costs. *Mathematics of Operations Research* **39**(2), 517–532.

Wang, J., F. Zhang. 2013. Strategic joining in $M/M/1$ retrial queues. *European Journal of Operational Research* **230**(1), 76–87.

Wu, J., X. Chao. 2014. Optimal control of a Brownian production/inventory system with average cost criterion. *Mathematics of Operations Research* **39**(1), 163–189.

Xia, L. 2014. Service rate control of closed Jackson networks from game theoretic perspective. *European Journal of Operational Research* **237**(2), 546–554.

Ziani, S., F. Rahmoune, M.S. Radjef. 2015. Customers' strategic behavior in batch arrivals $M^2/M/1$ queue. *European Journal of Operational Research* **247**(3), 895–903.