Article

On the Benefits of Providing Timely Information in Ticket Queues with Balking and Calling Times

Gabi Hanukov *, Michael Hassoun and Oren Musicant

Department of Industrial Engineering and Management, Ariel University, Kiriath Hamada, 3, Ariel 40700, Israel; michaelh@ariel.ac.il (M.H.); orenm@ariel.ac.il (O.M.)

* Correspondence: gabih@ariel.ac.il

Abstract: We study a phenomenon causing server time loss in ticket queues with balking and calling time. A customer who balks from the queue after printing a ticket leaves a virtual entity in the queue that requires server time to be cleared. The longer the queue, the larger the proportion of customers abandoning their place, and the larger the server time loss due to calling customers that left the queue. The solution is suggested by giving the customer the best possible estimate of her expected waiting time before printing a ticket, thus ensuring that, if she balks, no number in the queue is created that will waste server time. Although partially observable ticket queues have been studied in the literature, the addition of a calling time for absent customers creates a new type of problem that has been observed in real life but has not been formally addressed yet. We analyze this stochastic system, formulate its steady state probabilities, and calculate the system's performance measures. The analytical solution provided here is robust and can be applied to a wide range of customers' behavior functions. Finally, numerical analysis is performed that demonstrates the benefits of providing timely information to customers for different levels of traffic congestion.

Citation: Hanukov, G.; Hassoun, M.; Musicant, O. On the Benefits of Providing Timely Information in Ticket Queues with Balking and Calling Times. *Mathematics* **2021**, *9*, 2753. <https://doi.org/10.3390/math9212753>

Academic Editor: János Sztrik

Received: 1 October 2021

Accepted: 26 October 2021

Published: 29 October 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: queueing theory; ticket queue; value of information

1. Introduction

Waiting in a queue is one of the most ubiquitous human activities nowadays. We wait at the bank or at the pharmacy; we wait on the phone or in traffic at a crossing. As soon as we wait because others require the same service as us, we are “in the queue”.

More often than not, a customer is free to leave the queue (abandonment) when it is too long (traffic may be a good counter-example). This decision is usually taken based on the perception of the expected waiting time. However, actually standing in a first come first served (FCFS) waiting line and being able to visually assess the queue length and its progression is becoming the exception. An extreme case is found in call centers that are simply unobservable to the customer. With a low joining cost (picking up the phone) and no other information on the queue status than what is provided by the server, leaving the queue is an inherent part of call center operations (and an important performance measure). Gans et al. [1] provide an extensive review on the subject.

Ticket queues as studied here offer only partial observability. Nonetheless, they present strong advantages and are quickly becoming the queue management norm. While in regular queues there are numerous ways that servers can optimize their utilization by using their non-service time to perform other tasks, see, e.g., [2–15], customers, on their side, have little to no flexibility in the way they use their waiting time. Ticket queues allow the handling of large numbers of customers in relatively small places. Customers are not required to stand, and the responsibility of keeping their place in the line is transferred to the server. They can use their waiting time to run some errands or in other more enjoyable ways than standing in line, such as enjoying a cup of coffee at a nearby location [16]. Ticket queues also hide FCFS violations (sometimes justified), thus lowering the customer perception of social injustice. All these aspects reduce customer stress, improve their

overall satisfaction, and reduce their tendency to renege. Recent field studies show that customers prefer ticket queues over standing in line, especially in low traffic intensity settings (see [17]).

Besides their clear managerial advantages, ticket queues prompt new problems. Even when most customers are usually in the vicinity of the server, the expected waiting time can be difficult to assess. For the simplest ticket queues, a slip of paper is presented at the entrance of the waiting room with a number on it. The number is visible, and a customer can compare it to the one currently served as announced by a screen near the counter without actually taking the ticket. A customer, discouraged by what she estimates to be too long a wait, will balk without taking the ticket. In that simple case, a new customer has a good appraisal of the actual length of the queue by observing the number of people waiting, and the gap in ticket numbers is biased upward only by renegeing customers.

In more complex queues, different classes of customers are dispatched to multiple servers, creating several parallel queues, all waiting in the same room. Tickets are printed upon request. The ticket dispenser is connected through a central dispatch unit to the display, and rules other than FCFS can be implemented without the outrage it would otherwise create (see [18]), for example, by mixing customers that made an appointment beforehand with the regular flow of walk-in customers. When entering the room, a customer requires a specific service from the automatic ticket dispenser, which then prints a ticket corresponding to the relevant queue. Only then can she evaluate the expected waiting time. In this case, balking creates gaps in the ticket queue, which make the ticket number gap unreliable as an estimator of the waiting time. Additionally, the fact that several queues are waiting in a common room makes it almost impossible to evaluate the queue length based on the apparent level of congestion. As shown in [19], the combination of the abandonment with the observability limitations sets ticket queues apart from regular queues. In models tackling this combination of features in the literature, customers systematically overestimate the actual queue length, which increases abandonment.

The problem we present here, illustrated in Figure 1, has been observed first in local post offices and pharmacies, in which a basic flaw yields server losses at the worst time possible: when the system is congested. A customer enters the office and is prompted by a machine to take a ticket for one of several queues running independently of one another. All the queues share a common waiting room, making it difficult to estimate the actual length of one specific queue. Once the ticket is printed, the customer can compare her number to the one currently being served. At this point, some decide to leave immediately based on their estimation of the waiting time (balking). The longer the queue (apparent or real), the higher the probability for a customer to leave. If she decides to leave before being served, the customer's ticket is thrown away, but a virtual customer has been created in the queue. We remark that, given known customer patience and a reasonably accurate waiting time estimate, customers knowing that their odds of renegeing are high will not join the queue at all, and balking should become the preponderant abandonment mechanism.

On the server side, numbers are being advanced by the clerk and served one after another in an FCFS regime. The important point is that it takes some time after a number is called to determine that no one will answer the call and that the number represents a customer who has left. The absurdity in this system is that, at peak periods, when the queue is longer, a larger portion of customers leave. The solution lies in giving the customer the best possible estimate of her expected waiting time before printing a ticket, thus ensuring that, if she balks, no number in the queue is created that will waste server time. This will also decrease the chances of renegeing because of a longer than expected waiting time.

Balking, partially observable queues, and queues with calling times have all been studied in the literature. However, combining these features together reveals a problem that has yet to be addressed. In this work, we propose a robust methodology that is able to solve any reasonably well-behaved balking function.

Figure 1. The ticket queue with balking.

2. Literature Review

The present work studies a service system with two different classes of customers—i.e., real versus virtual. In this context, bodies of research dealing with queues with heterogeneous customers are particularly relevant to us. Guo and Hassin [20], for example, investigate the join-or-balk decisions of customers with heterogeneous delay sensitivity in a queueing system with vacations, i.e., in which idle servers can leave the system to carry out ancillary work. Based on empirical data from a call center, Yu et al. [21] analyze the abandonment behavior of customers and demonstrate how waiting costs are impacted by delayed announcements. In their setup, customer heterogeneity is modelled through the ratio between cost (waiting) and reward (being served), which is sampled from a folded normal distribution. In Hu et al. [22] a unique server processes two classes of customers prone to balking. One class is informed about delays, while the other is not. The authors show how the system throughput and customers' welfare are directly impacted by the relative importance of the two classes in the system.

Klimenok et al. [23] analyze queueing systems in which customer priorities are dynamic. When they enter the queue, customers are assigned a low priority and a timer is switched on. When the timer expires, they either leave the system unserved with a certain probability or are assigned a higher priority. For further reading on heterogeneous customers in queueing systems, see [24–29]. In all these works, all the classes describe actual customers, while, in our work, one type of customer is not actually present in the system.

Another stream of literature related to our work addresses queues with balking customers. Sun et al. [30] study customers' equilibrium and socially optimal balking strategies in single-server Markovian queues with multiple vacations and N-policy. The authors observe that customers' individual behavior in stable equilibrium always congests the system more than the socially optimal one. Morozov et al. [31] consider a single-server retrial model with multiple customer classes. A new customer, facing a busy server upon his arrival, may join the corresponding (class-dependent) orbiting queue with a class-dependent probability or leave the system forever (balking). An extensive simulation analysis allows insights on the model stability and performance. Zirem et al. [32] present a batch arrival queue with general retrial time, breakdowns, repairs, and reserved time. Upon arrival, when a batch of customers finds the server free, one of the customers from the batch begins service and the rest of them join into a so-called server orbit. Otherwise, the customers either balk or enter said orbit on an FCFS retrial basis. The authors conduct an analysis of their model through the use of the supplementary variables technique. Ke et al. [33] formalize an M/M/c balking retrial queue with vacation and study both single and multiple vacation policies. The optimization of the system is obtained through

various methods, such as Quasi-Newton, Nelder–Mead simplex, and simulated annealing. For further discussion on balking customers, see [34–39]. Contrary to all these research efforts, in our setup, the system is not aware of the customer’s balking.

Finally, unobservable queues have also been the subject of numerous studies in the literature. Haviv and Oz [40] consider an unobservable M/M/1 queue where customers are homogeneous with respect to both the service value and waiting time cost. The authors present a classification of the regulation schemes under which the resulting equilibrium joining rate coincides with the socially optimal one. Yu et al. [41] study the equilibrium threshold balking strategies for the unobservable single-server queues with server breakdowns and delayed repairs. Equilibrium mixed strategies are derived for both the partially observable and the unobservable queues. Lingenbrink and Lyer [42] demonstrate how, in a fixed-price service setup, a threshold-based (join/leave) partial information sharing policy can be optimal to the system. Kim and Kim [43] consider an unobservable queueing system with strategic customers. The authors show that customers who decide based on social welfare optimization arrive at a higher rate than those who decide based on profit maximization. Consequently, the admission fee required by the latter type of customers is higher than that required for the former. For further discussion on unobservable queues, see [44–49].

Xu et al. [50] consider balking customers with a threshold-type balking function while neglecting the clearance (calling) time for virtual customers. Through an approximation procedure, Ding et al. [51] show that communicating the current queue length to joining customers allows the reduction of the percentage of reneging customers. Numerous other papers on strategic queues assume that full information is provided to customers, who then decide to either balk or join the queue and wait for service (see e.g., [52–55]). By striving for a more realistic modelling of customers’ behavior, Kuzu et al. [56] show that ticket queues are more efficient than formerly predicted in the literature. For further research on abandonments in ticket queues, see [57].

In the present work, we address the same problem for different levels of workload, with a special interest in overloaded cases where the stability of the queue is obtained only due to customers leaving the system. We study the value of providing timely information to customers and thus preventing the creation of tickets for customers who decide to leave. The damages shown by our study are, in some cases, considerable and fully justify the efforts by researchers to reach accurate models for abandonment in overloaded, partially observable queues and by practitioners to limit the waste related to calling absent customers as much as possible. We demonstrate the aforementioned phenomenon on a simple model according to which customers arrive in a ticket queue, receive a ticket on which their number in line is provided, and then decide to either stay in line or balk. This case is hereafter referred to as the “post office model”, operating under the late information policy (LIP). The proposed solution is to inform customers of their number in line prior to printing a ticket, which is hereafter referred to as the early information policy (EIP). Our main objective is to study a realistic representation of the problem at hand, measure the damages caused by clearing customers who have left the system, and try to correlate these damages with the system characteristics.

The outline of the paper is as follows: Section 2 presents the analysis of the LIP model, including the exact model formulation and calculation of steady state probabilities and performance measures. In Section 3, the EIP model is derived. Section 4 provides a numerical comparison between the LIP and EIP models.

3. The Late Information Policy

3.1. Mathematical Modelling

A single server is assigned to customers who follow a Poisson arrival process with the rate λ . The customer queue is unobservable, and the server calls and serves customers following the order that the tickets are issued upon their arrival in an FCFS regime. Upon arrival, a customer draws a number from a ticket machine, observes the displayed running

number of the current customer being served, and, based on the difference between these two numbers, decides to either join the queue or balk. The difference between the two numbers is called the queue length. Since a customer is informed of the current queue length only after her ticket is issued, a balking customer leaves a trace in the system, one that will be dispatched to the server and that we call a virtual customer. When a ticket number is called, the server either serves the corresponding customer if this one did not balk (real customer) or spends a certain amount of time waiting for a customer before acknowledging that the ticket number represents a customer who balked (virtual customer). Both the service and calling times are assumed to follow an exponential distribution. The calling rate for virtual customers and the service rate for real customers are denoted μ and σ , respectively ($\mu > \sigma$).

Each arriving customer who sees q customers in the system acts as follows: (i) she enters the system if the number of customers in the system is less than or equal to the pre-specified value q_1 ; (ii) balks if the number of customers is larger than q_2 ($q_1 < q_2$); and (iii) in the in-between case, the customer enters the system with the probability decreasing based on the queue length. We denote by $\psi(q)$ the probability for an arriving customer to balk when the queue length is q . Accordingly, $\psi(q) = 0$ for $q \leq q_1$, and $\psi(q) = 1$ for $q \geq q_2$.

At any moment in time, we denote by r the number of contiguous real customers at the head of the queue (including the customer who gets service) and by v the number of contiguous virtual customers immediately following those r real customers. In order to analyze the system in a steady state, we define the system's state by a $(k+2)$ -dimension vector $(r, v, c_1, c_2, \dots, c_k)$. The value of c_i , $i = 1, 2, \dots, k-1$ indicates the number of virtual customers standing between the i th and $(i+1)$ th real customers (after the first $r+v$ customers), and c_k denotes the number of virtual customers standing after the k th real customer; k is, therefore, the number of real customers standing after the aforementioned $r+v$ customers. To illustrate the above definition, let us consider a few examples in which R and V stand for one real/virtual customer. In our notation, state $(3, 2, 2, 0, 0, 2)$ stands for the queue composition RRRVVRVRRRVV, state $(0, 2, 1, 3, 1)$ stands for VVRVRVVVRV, and state $(1, 0, 2, 0)$ stands for RRVVR. Consequently, the steady-state probabilities are defined as $p_{r,v,c_1,c_2,\dots,c_k}$.

Each level is specified by the value of its state's last component, either v or c_k , which indicates a number of virtual customers standing at the end of the line. Thus, the number of levels is unbounded. Within each level, the system's states (each representing the system's phase) arranged in order are described in Feature 1.

Feature 1. *The order of phases within a specific level.*

First, the phases are partitioned into groups, each of which is characterized by the state's length (i.e., the number of components in the state), which is given by $k+2$, $k = 0, 1, 2, \dots, q_2 - 1$, and arranged in increasing order of k , i.e., $(r, v), (r, v, c_1), (r, v, c_1, c_2), \dots, (r, v, c_1, c_2, \dots, c_k, \dots, c_{q_2-1})$.

Within each group, the states are arranged in lexicographic order of each state's component's values. Then, each group of length k is itself partitioned into sub-groups specified by the value of c_k , each of which is partitioned again into sub-groups specified by the value of c_{k-1} , and so on.

We now show in Proposition 1 that, for any given value of q_2 , each level consists of 2^{q_2} phases.

Proposition 1. *For any given value of q_2 , each level consists of 2^{q_2} phases.*

Proof. We denote an accepting state any state in which a real customer can enter the system, i.e., any state in which the number of customers is less than q_2 . At each accepting state, increasing the number of real customers by one transfers the system to another phase. Thus, the number of phases is equal to number of accepting states plus the first phase. Let $AS(q)$ be the number of accepting states with q customers, $q = 0, 1, 2, \dots, q_2 - 1$. Since, at each state with q customers, each customer can be either a real or a virtual customer, it

follows that $AS(q) = 2^q$. Thus, for any given value of q_2 , the number of accepting states is $\sum_{q=0}^{q_2-1} AS(q) = \sum_{q=0}^{q_2-1} 2^q = \frac{2^{q_2}-1}{2-1} = 2^{q_2} - 1$. Adding the first phase completes the proof. $\square$

The transition rates between the system's states are listed below:

- (i) From $(r, 0)$, $r = 0, 1, 2, \dots, q_1, \dots, q_2 - 1$ to $(r + 1, 0)$ with rate $\lambda(1 - \psi(r))$.
- (ii) From (r, v) , $r = 0, 1, 2, \dots, q_1, \dots, q_2$, $v = 0, 1, 2, \dots$ to $(r, v + 1)$ with rate $\lambda\psi(r + v)$.
- (iii) From (r, v) , $r = 1, 2, 3, \dots, q_2$, $v = 0, 1, 2, \dots$ to $(r - 1, v)$ with rate σ .
- (iv) From $(0, v)$, $v = 1, 2, 3, \dots$ to $(0, v - 1)$ with rate μ .
- (v) From (r, v) , $r = 0, 1, 2, \dots, q_1, \dots, q_2$, $v = 1, 2, 3, \dots$ to $(r, v, 0)$ with rate $\lambda(1 - \psi(r + v))$.
- (vi) From $(r, v, c_1, c_2, \dots, c_k)$, $k = 1, 2, 3, \dots, q_2 - 1$, $r \geq 1$, $v \geq 1$, $r + v + k + \sum_{i=1}^{k-1} c_i \leq q_2$ to $(r - 1, v, c_1, c_2, \dots, c_k)$ with rate σ .
- (vii) From $(0, v, c_1, c_2, \dots, c_k)$, $k = 1, 2, 3, \dots, q_2 - 1$, $v \geq 2$, $v + k + \sum_{i=1}^{k-1} c_i \leq q_2$ to $(0, v - 1, c_1, c_2, \dots, c_k)$ with rate μ .
- (viii) From $(0, 1, c_1, c_2, \dots, c_k)$, $k = 1, 2, 3, \dots, q_2 - 1$, $1 + k + \sum_{i=1}^{k-1} c_i \leq q_2$ to $(1, v = c_1, c_2, \dots, c_{k-1} = c_k)$ with rate μ .
- (ix) From $(r, v, c_1, c_2, \dots, c_k)$, $k = 1, 2, 3, \dots, q_2 - 2$, $r \geq 1$, $v \geq 1$, $r + v + k + \sum_{i=1}^{k-1} c_i \leq q_2$ to $(r, v, c_1, c_2, \dots, c_k, 0)$ with rate $\lambda(1 - \psi(r + v + k + \sum_{i=1}^k c_i))$.
- (x) From $(r, v, c_1, c_2, \dots, c_k)$, $k = 1, 2, 3, \dots, q_2 - 1$, $r \geq 0$, $v \geq 1$, $r + v + k + \sum_{i=1}^{k-1} c_i \leq q_2$ to $(r, v, c_1, c_2, \dots, c_k + 1)$ with rate $\lambda\psi(r + v + k + \sum_{i=1}^k c_i)$.

Cases (i), (iii), (vi), (vii), and (viii) indicate transitions within a given level, activated by, respectively, arrival of a real customer (case (i)), departure of a real customer (cases (iii) and (vi)), and clearance of a virtual customer (cases (vii) and (viii)). Cases (ii) and (x) indicate transitions from level i to level $i + 1$, $i = 0, 1, 2, \dots$, activated by arrival of a virtual customer. Case (iv) indicates transitions from level i to level $i - 1$, $i = 1, 2, 3, \dots$, activated by clearance of a virtual customer. Cases (v) and (ix) indicate transitions from level i to level 0 , $i = 1, 2, 3, \dots$, activated by an arrival of a real customer. Let $D_{i,0}$, $i = 0, 1, 2, \dots, q_2 - 1$ denote the square matrix of order $2^{q_2} \times 2^{q_2}$ composed of the transition rates from level i to level 0 activated by an arrival of a real customer (cases (i), (v) and (ix)), where, in each matrix $D_{i,0}$, the states are arranged in order described in Feature 1. Let D denote the square matrix of order $2^{q_2} \times 2^{q_2}$ composed of the transitions within level, i.e., transitions caused by service completion (σ) and by clearance completion (μ), again, the states arranged in order described in Feature 1.

3.2. Steady State Analysis

Let Q denote the infinitesimal generator matrix of the Markovian process described above. The matrix Q is given by

$$Q = \begin{pmatrix} B_{0,0} & B_{0,1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ B_{1,0} & A_1 & B_{1,2} & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ B_{2,0} & A_2 & A_1 & B_{2,3} & 0 & 0 & 0 & 0 & 0 & \cdots \\ B_{3,0} & 0 & A_2 & A_1 & B_{3,4} & 0 & 0 & 0 & 0 & \cdots \\ \vdots & \vdots & & \ddots & \ddots & \ddots & 0 & 0 & 0 & \cdots \\ B_{q_2-1,0} & 0 & \cdots & 0 & A_2 & A_1 & B_{q_2-1,q_2} & 0 & 0 & \cdots \\ 0 & 0 & \cdots & 0 & 0 & A_2 & A_1 & A_0 & 0 & \cdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & A_2 & A_1 & A_0 & \cdots \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \ddots & \ddots & \ddots \end{pmatrix}$$

where the matrices $B_{i,j}$, A_i , all of size $2^{q_2} \times 2^{q_2}$, are given as follows

$$\begin{aligned} B_{i,i+1} &= \lambda X_{2^{q_2} \times 2^{q_2}}^{i,i+1}, \quad i = 0, 1, 2, \dots, q_2 - 1, \\ A_0 &= \lambda I_{2^{q_2} \times 2^{q_2}}, \\ A_2 &= \mu X_{2^{q_2} \times 2^{q_2}}^1, \\ A_1 &= D + X_{2^{q_2} \times 2^{q_2}}^2, \\ B_{1,0} &= D_{1,0} + A_2, \\ B_{i,0} &= D_{i,0}, \\ B_{0,0} &= D + X_{2^{q_2} \times 2^{q_2}}^3, \end{aligned}$$

where

$$\begin{aligned} X_{2^{q_2} \times 2^{q_2}}^1 &= \text{diag} \{1 \ 0 \ 0 \ \cdots \ 0\}, \\ X_{2^{q_2} \times 2^{q_2}}^2 &= -\text{diag} \left\{ \left[(A_0 + A_2 + D) \vec{e}^T \right]^T \right\}, \\ X_{2^{q_2} \times 2^{q_2}}^3 &= -\text{diag} \left\{ \left[(B_{0,1} + D) \vec{e}^T \right]^T \right\}, \end{aligned}$$

$\vec{e} = (1 \ 1 \ \cdots \ 1)^T$ denotes a column vector of ones,

$\vec{u}_i$ denotes a column vector that indicates the number of customers of each of 2^{q_2}

phases in level i , arranged in order described in Feature 1, i.e., $r + v + k + \sum_{i=1}^k c_k$,

$i = 0, 1, 2, \dots, q_2 - 1$, $X_{2^{q_2} \times 2^{q_2}}^{i,i+1} = \text{diag}\{u_i\}$, $\text{diag}\{\cdots\}$ denotes the diagonal matrix with the diagonal entries listed in the brackets, and $I_{\times}$ denotes the identity matrix of having size indicated in the suffix.

Note that, following the order of phases described in Feature 1, the matrix A_1 is lower triangular.

Let $\vec{p}_i$, $i = 0, 1, 2, \dots$, be the probability vectors of the system's states of level i (arranged in order described in Feature 1). Further, denote the vector of all probabilities by $\vec{p} = (\vec{p}_0 \ \vec{p}_1 \ \vec{p}_2 \ \cdots)$. Then, the balance equations can be written as

$$\vec{p}Q = \vec{0} \text{ and } \sum_{i=0}^{\infty} \vec{p}_i \vec{e} = 1$$

For further analysis, the elements of matrices A_0 , A_1 , and A_2 are denoted as follows:

$$A_0 \equiv [a_0^{v,t}], \quad A_1 \equiv [a_1^{v,t}], \quad A_2 \equiv [a_2^{v,t}], \quad v = 1, 2, 3, \dots, 2^{q_2}, \quad t = 1, 2, 3, \dots, 2^{q_2}$$

Theorem 1. The system's stability condition is $\lambda < \mu$.

Proof. According to Hanukov and Yechiali [58], when each of the matrices A_0 , A_1 and A_2 are lower triangular (which is the case in our model), the stability condition is given by $a_0^{1,1} < a_2^{1,1}$, which, in our model, results in $\lambda < \mu$. $\square$

Theorem 1 shows that the stability condition is not affected by real customers' service rate, σ . This result is explained by the fact that the number of real customers in the system is bounded. Let R be the matrix satisfying

$$A_0 + RA_1 + R^2A_2 = 0_{2^{q_2} \times 2^{q_2}}$$

In general, the matrix R is calculated via successive substitutions; see [59,60]. However, in some special cases, the matrix R can be obtained directly. One case is when A_2 is of rank 1, satisfying $A_2 = \vec{c} \cdot \vec{r}$, where $\vec{c}$ is a column vector and $\vec{r}$ is a row vector normalized by $\vec{r} \cdot \vec{e} = 1$ (see [21]). In this case, R can be calculated by $R = -A_0(A_1 + A_0\vec{e} \cdot \vec{r})^{-1}$. In our model, $A_2 = \vec{c} \cdot \vec{r}$ with $\vec{c} = (\mu, 0, 0, \dots, 0)^T$ and $\vec{r} = (1, 0, 0, \dots, 0)$.

Another more general case (see [58]) is when each of the three matrices A_0 , A_1 , and A_2 is lower triangular, as is the case in the current model. In such a case, the entries of $R \equiv [r^{v,t}]$ are given explicitly by

$$r^{v,t} = 0, \quad v < t \quad (1)$$

$$r^{v,v} = \begin{cases} \frac{-a_1^{v,v} - \sqrt{(a_1^{v,v})^2 - 4a_0^{v,v}a_2^{v,v}}}{2a_2^{v,v}}, & a_2^{v,v} > 0, a_0^{v,v} > 0 \\ 0, & a_2^{v,v} > 0, a_0^{v,v} = 0, \\ \frac{-a_0^{v,v}}{a_1^{v,v}}, & a_2^{v,v} = 0 \end{cases} \quad \forall v \quad (2)$$

$$r^{v,t} = -\frac{a_0^{v,t} + \sum_{k=t+1}^v r^{v,k}a_1^{k,t} + \sum_{\tau=t+1}^{v-1} r^{v,\tau}r^{\tau,t}a_2^{t,t} + \sum_{k=t+1}^v \sum_{\tau=k}^v r^{v,\tau}r^{\tau,k}a_2^{k,t}}{a_1^{t,t} + a_2^{t,t}(r^{t,t} + r^{v,v})}, \quad v > t \quad (3)$$

Then, the steady state probability vectors satisfy

$$\vec{p}_i = \vec{p}_{q_2} R^{i-q_2}, \quad i = q_2, q_2 + 1, q_2 + 2, \dots \quad (4)$$

In order to calculate those probability vectors, one needs first to obtain the vectors $\vec{p}_i, i = 0, 1, \dots, q_2 - 1$. This is achieved by considering the corresponding vector equations from the set $\vec{p}Q = \vec{0}$ and $\sum_{i=0}^{\infty} \vec{p}_i \vec{e} = 1$. Thus,

$$\begin{aligned} \sum_{j=0}^{q_2-1} \vec{p}_j B_{j,0} &= \vec{0}, \\ \vec{p}_{i-1} B_{i-1,i} + \vec{p}_i A_1 + \vec{p}_{i+1} A_2 &= \vec{0}, \quad i = 1, 2, 3, \dots, q_2 - 1, \\ \vec{p}_{q_2-1} B_{q_2-1,q_2} + \vec{p}_{q_2} (A_1 + RA_2) &= \vec{0}, \\ \vec{p}_{q_2} [I - R]^{-1} \vec{e} &= 1. \end{aligned}$$

In the next two theorems, an alternative representation of stability condition is given.

Theorem 2. The stability condition is given by $r^{1,1} < 1$.

Proof. Since A_0 , A_1 , and A_2 are all lower triangular, then $a_1^{1,1} = -(a_0^{1,1} + a_2^{1,1})$. Thus, according to the first term in Equation (2), we have $r^{1,1} = a_0^{1,1}/a_2^{1,1}$, which, together with Theorem 1, proves the claim. $\square$

The next theorem shows that the representation of the stability condition introduced in Theorem 2 is valid also when matrices A_0 , A_1 , and A_2 are upper triangular.

Theorem 3. *If each of the matrices A_0 , A_1 , and A_2 are upper triangular, the stability condition is given by $r^{n,n} < 1$.*

Proof. Since A_0 , A_1 , and A_2 are all upper triangular, then $a_1^{n,n} = -(a_0^{n,n} + a_2^{n,n})$. Thus, according to the first term in Equation (2), we have $r^{n,n} = a_0^{n,n} / a_2^{n,n}$. According to Hanukov and Yechiali [58], when each of the matrices A_0 , A_1 , and A_2 are upper triangular, the stability condition is given by $a_0^{n,n} < a_2^{n,n}$, which proves the claim. $\square$

Note that $r^{1,1}$ and $r^{n,n}$ are, respectively, left-upper and right-lower entries of R .

3.3. Performance Measures

The system performance is evaluated through several variables: Customers are interested in their expected flow time (sum of waiting time and service time) and their probability to receive service (complementary to the balking probability). The server performance is evaluated by its utilization. Finally, we also measure the total (ticket) queue length.

As mentioned, the only difference between levels is the value of states' last component (v or c_k). Thus, when this value is excluded, each phase has the same number of customers in the system in all its states. Let $\vec{u}_Q$ be the column vector that indicates the number of customers of each of 2^{q_2} phases, arranged in order described in Feature 1, i.e., $r + v + k + \sum_{i=1}^{k-1} c_k$. Let $\vec{i}_{2^{q_2}} = (i \ i \ \dots \ i)^T$ be a 2^{q_2} -dimensional column vector. Let L be the mean total number of customers in the system, which can be calculated as follows:

$$\begin{aligned} L &= \sum_{i=0}^{q_2-1} \vec{p}_i \left(\vec{u}_Q + \vec{i}_8 \right) + \sum_{i=q_2}^{\infty} i \vec{p}_i \vec{e} + \sum_{i=q_2}^{\infty} \vec{p}_i \vec{u}_Q = \sum_{i=0}^{q_2-1} \vec{p}_i \left(\vec{u}_Q + \vec{i}_8 \right) + \sum_{i=q_2}^{\infty} i \vec{p}_{q_2} R^{i-q_2} \vec{e} + \sum_{i=q_2}^{\infty} \vec{p}_{q_2} R^{i-q_2} \vec{u}_Q \\ &= \sum_{i=0}^{q_2-1} \vec{p}_i \left(\vec{u}_Q + \vec{i}_8 \right) + \vec{p}_{q_2} (q_2 [I - R]^{-1} + R [I - R]^{-2}) \vec{e} + \vec{p}_{q_2} [I - R]^{-1} \vec{u}_Q. \end{aligned} \quad (5)$$

Let N be the mean number of real customers in the system. In order to calculate N , let $\vec{u}_N$ be the column vector indicating the number of real customers of each of 2^{q_2} phases, arranged in order described in Feature 1, i.e., $r + k$. Thus,

$$N = \sum_{i=0}^{q_2-1} \vec{p}_i \vec{u}_N + \sum_{i=q_2}^{\infty} \vec{p}_i \vec{u}_N = \sum_{i=0}^{q_2-1} \vec{p}_i \vec{u}_N + \sum_{i=q_2}^{\infty} \vec{p}_{q_2} R^{i-q_2} \vec{u}_N = \sum_{i=0}^{q_2-1} \vec{p}_i \vec{u}_N + \vec{p}_{q_2} [I - R]^{-1} \vec{u}_N.$$

The effective arrival rate of real customers is given by:

$$\lambda_{eff} = \sum_{i=0}^{q_2-1} p_i D_{i,0} \vec{e}.$$

Then, using Little's law, the expected flow time of a customer who has decided to join the queue is obtained by:

$$F = \frac{N}{\lambda_{eff}} \quad (6)$$

Let SL be the service level, which is the portion of customers who join the queue, and given by:

$$SL = \frac{\lambda_{eff}}{\lambda} \quad (7)$$

The server utilization is given by:

$$U = 1 - p_{0,0} \quad (8)$$

Let U_{eff} be the server effective utilization, defined as the proportion of time the server is busy processing real customers (excluding the calling time for virtual customers). We note that the portion of utilized time assigned to real customers can be expressed by:

$$\frac{SL/\sigma}{SL/\sigma + (1-SL)/\mu}.$$

This allows formulating the effective utilization as:

$$U_{eff} = U \cdot \frac{SL/\sigma}{SL/\sigma + (1-SL)/\mu} \quad (9)$$

4. The Early Information Policy (EIP)

In the EIP, the queue length is provided to the customer upon arrival, and the decision of leaving is taken before a ticket is issued. In this case, the ticket queue describes accurately the customer queue, and there is no loss of server capacity to call balking customers. Such a system is a birth-and-death process where the states represent the queue length of non-balking customers. The arrival rate to a given state is the general arrival rate conditioned by the probability for the customer to stay in the queue. The birth process is, therefore, state-dependent with rate $\lambda_q = \lambda(1 - \psi(q))$. The death process is state-independent with rate σ .

Let us denote π_q , the steady-state probability of being in state q and $\rho = \lambda/\sigma$. Solving the balance equations of state q is straightforward (see Kleinrock [61], p. 92) and leads to:

$$\pi_0 = \frac{1}{1 + \sum_{q=1}^{q_2} \rho^q \prod_{j=0}^{q-1} (1 - \psi(j))}, \quad \pi_q = \pi_0 \rho^q \prod_{j=0}^{q-1} (1 - \psi(j)).$$

The expected service level can be expressed as:

$$SL_{ear} = 1 - \sum_{q=0}^{q_2} \pi_q \psi(q) \quad (10)$$

The expected value of the ticket queue length is:

$$L_{ear} = \sum_{q=0}^{q_2} q \pi_q \quad (11)$$

Based on Little's law with an average arrival rate of $\lambda \cdot SL$, we can compute the expected flow time as:

$$F_{ear} = \frac{Q}{\lambda \cdot SL} \quad (12)$$

The server utilization is given by:

$$U_{ear} = 1 - \pi_0 \quad (13)$$

5. Numerical Results: LIP vs. EIP

In the following, all the calculations were conducted in Maple 2021. Our work focuses on the impact of providing delayed information on the system. We, therefore, compare each of the aforementioned five systems' performance under the EIP vs. their counterpart

under the LIP over a range of traffic intensity values. We assume that the balking function is piecewise linear as follows:

$$\psi(q) = \begin{cases} 0 & q \leq q_1 \\ \frac{q-q_1}{q_2-q_1} & q_1 < q \leq q_2 \\ 1 & q_2 < q \end{cases} \quad (14)$$

For the sake of illustrating the behavior of our model, we use the following parameter values as an example: $\lambda = 25$, $\sigma = 20$, $\mu = 30$, $q_1 = 1$, $q_2 = 3$. Based on these values, the performance measures can be calculated by the following procedure: (i) we obtain the rate-matrix R by using Equations (1)–(3); (ii) then, we calculate the steady state probabilities using Equation (4); (iii) we use Equations (5)–(9) to get the following performance measures for the late information policy: $L = 6.717$, $F = 0.089$, $SL = 0.213$, $U = 0.922$, $U_{eff} = 0.267$; and, finally, (iv) we use Equations (10)–(13) to get the following performance measures for the early information policy: $L_{ear} = 1.525$, $F_{ear} = 0.076$, $SL_{ear} = 0.796$, $U_{ear} = 0.791$. We illustrate the improvement achieved by adopting the early information policy by calculating the ratio between corresponding performance measures as presented above.

$$L_{EIPvsLIP} = \frac{L_{ear}}{L} = \frac{1.525}{6.717} = 0.227 \quad (15)$$

$$F_{EIPvsLIP} = \frac{F_{ear}}{F} = \frac{0.076}{0.089} = 0.854 \quad (16)$$

$$SL_{EIPvsLIP} = \frac{SL_{ear}}{SL} = \frac{0.796}{0.213} = 3.737 \quad (17)$$

$$U_{EIPvsLIP} = \frac{U_{ear}}{U} = \frac{0.791}{0.922} = 0.859 \quad (18)$$

$$U_{eff,EIPvsLIP} = \frac{U_{ear}}{U_{eff}} = \frac{0.791}{0.267} = 2.963 \quad (19)$$

Equation (15) shows that the mean number of customers present in the system decreases when adopting the early information policy. Similarly, Equation (16) shows that the mean time the customer spends in the system decreases with the EIP. Equation (17) shows that the service level is higher with the EIP. Equation (18) shows that, with the EIP, the server idle time increases; however, Equation (19) shows that, with the LIP, the server spends more time actually serving customers.

We expand the comparison study between the EIP and LIP policies by providing the following sensitivity analysis: all rates (μ, σ, λ) are measured in units (customers) per minute. The calling rate is kept constant at $\mu = 30$, and the service rate σ was chosen in set $\{3, 10, 20\}$, resulting in ratios between both rates in $\{1.5, 3, 10\}$. The smaller the σ , the less significant the calling time compared to real customer service time. The arrival rate λ evolved relative to both service rates. Two values of λ were chosen slightly under σ , namely, $\lambda \in \{0.75\sigma, 0.95\sigma\}$ to represent under-stressed scenarios, i.e., situations in which the queue would have been stable even without balking. The rest of the scenarios that we denote as stressed are characterized by an arrival rate higher than σ . In these cases, the balking mechanism allows for the system to stabilize. The arrival rate values were set as $\lambda = \sigma + \gamma(\mu - \sigma)$ with $\gamma \in \{0, 0.25, 0.5, 0.75, 0.9, 0.99\}$. The two highest λ values were chosen to express severe system congestion. The balking function's parameters are assumed to be $q_1 = 1$ and $q_2 = 3$.

Figures 2–6 present, respectively, for each of the five performance levels, the ratio of the performance level under the EIP vs. its counterpart under the LIP as a function of the traffic level and for each of the three service rates. Notice that, because the ratio values span over a very large range for the service level and for the effective utilization, a logarithmic scale was used. Moreover, the abscissa present values of γ for stressed scenarios (between the two vertical lines marking $\gamma = 0$ and $\gamma = 1$), except for the two smallest values that

stand for under stressed scenarios $\{0.75\sigma, 0.95\sigma\}$. Figures 2 and 4 clearly show the steep degradation in the performance of the LIP as the traffic intensifies and reaches saturation. At this point, under the EIP, both the service level and the effective utilization are more than two orders of magnitude higher than their LIP counterparts, meaning that the LIP essentially paralyzes the system at higher levels of congestion. Although it may be difficult to get a correct impression of the effects of the difference between the service and calling rates on the log scale, for stressed scenarios, the values of the effective utilization ratios for $\sigma = 20$ are between 36% and 47% higher than for $\sigma = 3$. This is somehow not a surprise since the longer the calling times relative to the service time, the less efficient the server is expected to be. All this happens while the server utilization levels stay comparable (Figure 3). In terms of the customer experience, although the queue length under the EIP is shorter than under the LIP, with the gap between them increasing with the level of congestion (Figure 5), Figure 6 shows that the flow time remains comparable between the two policies when the calling times are relatively small, with a clear advantage to the EIP when the calling rate gets closer to the service rate.

Figure 2. Ratio of service level EIP vs. LIP as a function of traffic intensity.

Figure 3. Ratio of utilization EIP vs. LIP as a function of traffic intensity.

Figure 4. Ratio of effective utilization EIP vs. LIP as a function of traffic intensity.

Figure 5. Ratio of queue length EIP vs. LIP as a function of traffic intensity.

Figure 6. Ratio of flow time EIP vs. LIP as a function of traffic intensity.

To conclude this short comparative section, let us state that, while the server may experience about the same level of utilization under both the EIP and LIP and the sojourn time of the customers that decide to stay in the queue may also remain comparable, the difference between the two information policies is mostly felt in the service level, with the LIP crippling the system as the traffic increases up to almost total paralysis for heavy congestion.

6. Conclusions

In this work, we address the damage done to the performance of ticket queues. We first demonstrated how ignoring the creation of virtual customers proves to be a poor management policy. The lateness of information has a significant impact on most aspects of the system performance from the customer point of view (service level, flow time) while being slightly detrimental to the operation from the server point of view (utilization). We then studied the sensitivity of the results to the system characteristics. The importance of the information policy on the service level was shown to be heavily dependent on the level of traffic intensity (the arrival rate relative to both calling and service rates). In cases where the calling times and service times are close, we showed that the late information policy is even more damaging to the system and that the loss in terms of the flow time and effective utilization is even greater as compared to the early information policy.

The practical solution to the costly phenomenon presented here may seem simple to implement: try to provide the customer with the highest quality of information prior to her entry in the queue. In our simplified model, this meant informing the customer of the queue size before printing a ticket. However, customers tend to leave not only upon arrival but also if the waiting time is too long or the apparent queue progression is too slow (reneging). The damage in terms of the calling time is exactly the same, which means that every effort to reduce the creation of virtual customers as much as possible should be considered: improving the information quality by giving an estimate of the waiting time, allowing for customers to go carry out errands and be prompted by messaging about their approaching time of service, allowing them to fix “appointments”, or making their stay more comfortable.

Another type of solution could involve eliminating virtual customers from the ticket queue, for example, by prompting them to rescan their ticket close to their time of service, or, if a time comes when queues are managed through customer mobile devices, by identifying customers that leave the queue.

While some of these solutions may be costless on new systems (it only takes simple programming to display the estimate of the expected waiting time), implementing them on existing systems may require significant expenses. This is even more relevant when considering possible improvements to the waiting room. The proposed analysis allows one to estimate the worthiness of any such improvement.

The proposed methodology could be broadened in several different directions. First and foremost, future research can tackle the problem of reneging in the same type of partially observable queues with calling times and using similar tools. Second, by using simulation to overcome analytical difficulties, we could tackle more complex, maybe intractable variations on the present system by including, for example, both balking and reneging, blocking, multiple servers, etc. Finally, it would only be natural to bring the ideas presented here to the field, conduct naturalistic studies on real-life instances of this problem, and consider the effectiveness of different solutions.

Author Contributions: Conceptualization, G.H., M.H. and O.M.; methodology, G.H.; software, G.H.; validation, G.H., M.H. and O.M.; formal analysis, G.H.; investigation, G.H., M.H. and O.M.; writing—original draft preparation, G.H., M.H. and O.M.; writing—review and editing, G.H., M.H. and O.M.; visualization, G.H., M.H. and O.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Gans, N.; Koole, G.; Mandelbaum, A. Telephone call centers: Tutorial, review, and research prospects. *Manuf. Serv. Oper. Manag.* **2003**, *5*, 79–141. [\[CrossRef\]](#)
- Hanukov, G.; Avinadav, T.; Chernonog, T.; Spiegel, U.; Yechiali, U. A queueing system with decomposed service and inventoried preliminary services. *Appl. Math. Model.* **2017**, *47*, 276–293. [\[CrossRef\]](#)
- Hanukov, G.; Avinadav, T.; Chernonog, T.; Spiegel, U.; Yechiali, U. Improving efficiency in service systems by performing and storing “preliminary services”. *Int. J. Prod. Econ.* **2018**, *197*, 174–185. [\[CrossRef\]](#)
- Goswami, V.; Panda, G. Synchronized abandonment in discrete-time renewal input queues with vacations. *J. Ind. Manag. Optim.* **2021**. [\[CrossRef\]](#)
- Panda, G.; Goswami, V.; Banik, A.D. Equilibrium and socially optimal balking strategies in Markovian queues with vacations and sequential abandonment. *Asia-Pac. J. Oper. Res.* **2016**, *33*, 1650036. [\[CrossRef\]](#)
- Pacheco, A.; Simões, M.L.; Milheiro-Oliveira, P. Queues with server vacations as a model for pretimed signalized urban traffic. *Transp. Sci.* **2017**, *51*, 841–851. [\[CrossRef\]](#)
- Takhedmit, B.; Abbas, K. A parametric uncertainty analysis method for queues with vacations. *J. Comput. Appl. Math.* **2017**, *312*, 143–155. [\[CrossRef\]](#)
- Hanukov, G.; Avinadav, T.; Chernonog, T.; Yechiali, U. Performance improvement of a service system via stocking perishable preliminary services. *Eur. J. Oper. Res.* **2019**, *274*, 1000–1011. [\[CrossRef\]](#)
- Hanukov, G.; Avinadav, T.; Chernonog, T.; Yechiali, U. A multi-server system with inventory of preliminary services and stock-dependent demand. *Int. J. Prod. Res.* **2021**, *59*, 4384–4402. [\[CrossRef\]](#)
- Kumar, R.; Ghosh, S.; Banik, A.D. Numerical study on transient behaviour of finite bulk arrival or service queues with multiple working vacations. *Int. J. Math. Oper. Res.* **2021**, *18*, 384–403. [\[CrossRef\]](#)
- Goswami, V.; Panda, G. Performance analysis of renewal input queues with multiple vacations and synchronized abandonment. *Int. J. Manag. Sci. Eng. Manag.* **2021**, 1–13. [\[CrossRef\]](#)
- Panda, G.; Goswami, V. Strategic customers in Markovian queues with vacations and synchronized abandonment. *ANZIAM J.* **2020**, *62*, 89–120. [\[CrossRef\]](#)
- Laxmi, P.V.; Rajesh, P. Analysis of variant working vacations on batch arrival queues. *Opsearch* **2016**, *53*, 303–316. [\[CrossRef\]](#)
- Wang, Y.; Tian, R. Strategic analysis in Markovian queues with a single working vacation and multiple vacations. *Int. J. Innov. Comput. Inf. Control* **2020**, *16*, 349–366.
- Shin, Y.W. Algorithmic approach to Markovian multi-server retrial queues with vacations. *Appl. Math. Comput.* **2015**, *250*, 287–297. [\[CrossRef\]](#)
- Hanukov, G.; Anily, S.; Yechiali, U. Ticket queues with regular and strategic customers. *Queueing Syst.* **2020**, *95*, 145–175. [\[CrossRef\]](#)
- Kuzu, K. Comparisons of perceptions and behavior in ticket queues and physical queues. *Serv. Sci.* **2015**, *7*. [\[CrossRef\]](#)
- Larson, R.C. OR forum—Perspectives on queues: Social justice and the psychology of queueing. *Oper. Res.* **1987**, *35*, 895–905. [\[CrossRef\]](#)
- Jennings, O.B.; Pender, J. Comparisons of ticket and standard queues. *Queueing Syst.* **2016**, *84*, 145–202. [\[CrossRef\]](#)
- Guo, P.; Hassin, R. Strategic behavior and social optimization in Markovian vacation queues: The case of heterogeneous customers. *Eur. J. Oper. Res.* **2012**, *222*, 278–286. [\[CrossRef\]](#)
- Yu, Q.; Allon, G.; Bassamboo, A. How do delay announcements shape customer behavior? An empirical study. *Manag. Sci.* **2017**, *63*, 1–20. [\[CrossRef\]](#)
- Hu, M.; Li, Y.; Wang, J. Efficient ignorance: Information heterogeneity in a queue. *Manag. Sci.* **2018**, *64*, 2650–2671. [\[CrossRef\]](#)
- Klimenok, V.; Dudin, A.; Dudina, O.; Kochetkova, I. Queueing System with Two Types of Customers and Dynamic Change of a Priority. *Mathematics* **2020**, *8*, 824. [\[CrossRef\]](#)
- Palmer, R.; Utley, M. On the modelling and performance measurement of service networks with heterogeneous customers. *Ann. Oper. Res.* **2020**, *293*, 237–268. [\[CrossRef\]](#)
- Dudin, A.N.; Lee, M.H.; Dudina, O.; Lee, S.K. Analysis of priority retrial queue with many types of customers and servers reservation as a model of cognitive radio system. *IEEE Trans. Commun.* **2016**, *65*, 186–199. [\[CrossRef\]](#)
- Gavirneni, S.; Kulkarni, V.G. Self-selecting priority queues with Burr distributed waiting costs. *Prod. Oper. Manag.* **2016**, *25*, 979–992. [\[CrossRef\]](#)
- Sun, B.; Dudina, O.S.; Dudin, S.A.; Samouylov, K.E. Optimization of admission control in tandem queue with heterogeneous customers and pre-service. *Optimization* **2018**, *69*, 165–185. [\[CrossRef\]](#)
- Jiang, T. Analysis of a tollbooth tandem queue with two-class customers and two heterogeneous dedicated servers. *Asia-Pac. J. Oper. Res.* **2018**, *35*, 1850043. [\[CrossRef\]](#)

29. Klimenok, V.; Dudin, A.; Vishnevsky, V. Priority Multi-Server Queueing System with Heterogeneous Customers. *Mathematics* **2020**, *8*, 1501. [\[CrossRef\]](#)
30. Sun, W.; Li, S.; Cheng-Guo, E. Equilibrium and optimal balking strategies of customers in Markovian queues with multiple vacations and N-policy. *Appl. Math. Model.* **2016**, *40*, 284–301. [\[CrossRef\]](#)
31. Morozov, E.; Rumyantsev, A.; Dey, S.; Deepak, T.G. Performance analysis and stability of multiclass orbit queue with constant retrial rates and balking. *Perform. Eval.* **2019**, *134*, 102005. [\[CrossRef\]](#)
32. Zirem, D.; Boualem, M.; Adel-Aissanou, K.; Aïssani, D. Analysis of a single server batch arrival unreliable queue with balking and general retrial time. *Qual. Technol. Quant. Manag.* **2019**, *16*, 672–695. [\[CrossRef\]](#)
33. Ke, J.-C.; Chang, F.-M.; Liu, T.-H. M/M/c balking retrial queue with vacation. *Qual. Technol. Quant. Manag.* **2019**, *16*, 54–66. [\[CrossRef\]](#)
34. Li, Q.; Guo, P.; Wang, Y. Equilibrium analysis of unobservable M/M/n priority queues with balking and homogeneous customers. *Oper. Res. Lett.* **2020**, *48*, 674–681. [\[CrossRef\]](#)
35. Kumar, R.; Sharma, S. Transient analysis of an M/M/c queueing system with balking and retention of reneging customers. *Commun. Stat. Methods* **2018**, *47*, 1318–1327. [\[CrossRef\]](#)
36. Cui, S.; Su, X.; Veeraraghavan, S. A model of rational retrials in queues. *Oper. Res.* **2019**, *67*, 1699–1718. [\[CrossRef\]](#)
37. Sunar, N.; Tu, Y.; Ziya, S. Pooled vs. Dedicated Queues when Customers Are Delay-Sensitive. *Manag. Sci.* **2021**, *67*. [\[CrossRef\]](#)
38. Chen, P.; Zhou, Y. Equilibrium balking strategies in the single server queue with setup times and breakdowns. *Oper. Res.* **2015**, *15*, 213–231. [\[CrossRef\]](#)
39. Sun, W.; Li, S.; Tian, N. Equilibrium and optimal balking strategies of customers in unobservable queues with double adaptive working vacations. *Qual. Technol. Quant. Manag.* **2017**, *14*, 94–113. [\[CrossRef\]](#)
40. Haviv, M.; Oz, B. Self-regulation of an unobservable queue. *Manag. Sci.* **2018**, *64*, 2380–2389. [\[CrossRef\]](#)
41. Yu, S.; Liu, Z.; Wu, J. Equilibrium strategies of the unobservable M/M/1 queue with balking and delayed repairs. *Appl. Math. Comput.* **2016**, *290*, 56–65. [\[CrossRef\]](#)
42. Lingenbrink, D.; Iyer, K. Optimal signaling mechanisms in unobservable queues. *Oper. Res.* **2019**, *67*, 1397–1416. [\[CrossRef\]](#)
43. Kim, B.; Kim, J. Comparison of socially optimal and profit maximizing prices in an unobservable queue with heterogeneous waiting costs. *Oper. Res. Lett.* **2021**, *49*, 207–211. [\[CrossRef\]](#)
44. Hassin, R.; Roet-Green, R. The impact of inspection cost on equilibrium, revenue, and social welfare in a single-server queue. *Oper. Res.* **2017**, *65*, 804–820. [\[CrossRef\]](#)
45. Benioudakis, M.; Burnetas, A.; Ioannou, G. Lead-time quotations in unobservable make-to-order systems with strategic customers: Risk aversion, load control and profit maximization. *Eur. J. Oper. Res.* **2021**, *289*, 165–176. [\[CrossRef\]](#)
46. Lingenbrink, D.; Iyer, K. Optimal signaling mechanisms in unobservable queues with strategic customers. In Proceedings of the Proceedings of the 2017 ACM Conference on Economics and Computation, Cambridge, MA, USA, 26–30 June 2017; p. 347.
47. Goswami, V.; Panda, G. Optimal customer behavior in observable and unobservable discrete-time queues. *J. Ind. Manag. Optim.* **2021**, *17*, 299. [\[CrossRef\]](#)
48. Wang, F.; Wang, J.; Zhang, Z.G. Strategic behavior and social optimization in a double-ended queue with gated policy. *Comput. Ind. Eng.* **2017**, *114*, 264–273. [\[CrossRef\]](#)
49. Tian, R. Social optimization and pricing strategies in unobservable queues with delayed multiple vacations. *Math. Probl. Eng.* **2019**, *2019*, 4684957. [\[CrossRef\]](#)
50. Xu, S.H.; Gao, L.; Ou, J. Service performance analysis and improvement for a ticket queue with balking customers. *Manag. Sci.* **2007**, *53*, 971–990. [\[CrossRef\]](#)
51. Ding, D.; Ou, J.; Ang, J. Analysis of ticket queues with reneging customers. *J. Oper. Res. Soc.* **2015**, *66*, 231–246. [\[CrossRef\]](#)
52. Liu, Z.; Cao, C.; Gao, S. Equilibrium Joining Strategies in the Geo/GeoK/1 Queueing System. *Mathematics* **2019**, *7*, 1029. [\[CrossRef\]](#)
53. Hassin, R. *Rational Queueing*; CRC Press: Boca Raton, FL, USA, 2016; ISBN 1498745288.
54. Perlman, Y.; Yechiali, U. The impact of infection risk on customers' joining strategies. *Saf. Sci.* **2021**, *138*, 105194. [\[CrossRef\]](#)
55. Hanukov, G.; Avinadav, T.; Chernonog, T.; Yechiali, U. A service system with perishable products where customers are either fastidious or strategic. *Int. J. Prod. Econ.* **2020**, *228*, 107696. [\[CrossRef\]](#)
56. Kuzu, K.; Gao, L.; Xu, S.H. To wait or not to wait: The theory and practice of ticket queues. *Manuf. Serv. Oper. Manag.* **2019**, *21*, 853–874. [\[CrossRef\]](#)
57. Kuzu, K.; Soyer, R. Bayesian modeling of abandonments in ticket queues. *Nav. Res. Logist.* **2018**, *65*, 499–521.
58. Hanukov, G.; Yechiali, U. Explicit solutions for continuous-time QBD processes by using relations between matrix geometric analysis and the probability generating functions method. *Probab. Eng. Inf. Sci.* **2021**, *35*, 565–580. [\[CrossRef\]](#)
59. Ramswami, V.; Latouche, G. A general class of Markov processes with explicit matrix-geometric solutions. *Oper.-Res.-Spektrum* **1986**, *8*, 209–218. [\[CrossRef\]](#)
60. Neuts, M.F. *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*; Courier Corporation: Baltimore, MA, USA, 1994; ISBN 0486683427.
61. Kleinrock, L. *Queueing Systems, Volume I: Theory*; Wiley: New York, NY, USA, 1975.