Queueing Models and
Service Management
Vol. 7, No. 2, page 31-56, 2024

QMSM
©AU 2023

Optimization Analysis of Ticket Queues with Balking Customers and Single Vacation Policy

Chia-Huang Wu* and Jyun-Lun Shu

Department of Industrial Engineering and Management,

National Yang Ming Chiao Tung University, Hsinchu, Taiwan

(*Received October 2023; accepted June 2024*)

Abstract: Self-checkout services have gained popularity across various industries, offering Service systems issue numbered tickets for upon arrival customers without physical queues are popularly applied in public sectors. These systems are managed by ticketing technology and thus are different from those in common classical queues. This paper introduces a novel ticket queue that accounts for impatient customers and a single vacation policy. A schematic state-transition-rate diagram with the associated flow-balance equations is presented. The block-partitioned infinitesimal generator is provided in matrix form and the corresponding steady-state probabilities are solved recursively using the matrix-geometric method. We also derive explicit expressions of critical metrics relative to the performance measures. Numerical sensitivity analysis and graphical results are presented to assess the influence of various parameters on system characteristics. To reduce the computational complexity and enhance the analysis efficiency, we simplify the model and provide an efficient approximation method. Furthermore, a stepwise regression model is constructed to estimate the expected number of customers in the system without the need for complex matrix manipulations. Finally, applying the NSGA-II algorithm, a triple-objective optimization problem is investigated to determine the optimal operating condition with the minimum cost.

Keywords: balking customers, matrix-geometric method, performance analysis, single vacation, ticket queue, triple-objective optimization

1. Introduction

In recent years, considerable concern has arisen over customer service quality and system operational efficiency in different industries, such as manufacturing, transportation, and network communication. To analyze and manage service systems with different characteristics, numerous investigations in the literature have introduced many queueing models. Part of them assumes customers may leave the system before receiving service due to impatience, i.e., balking behavior. Baccelli *et al.* [1] first proposed a single-server queue with impatient customers. They studied the relationship between the virtual waiting time and the actual offered waiting time. Most impatient customers are also strategic and therefore, the information provided by the system may affect the customers' decision-

* Corresponding author
Email: jacalwu@nycu.edu.tw

making. Mandelbaum and Shimkin [17] investigated abandonments from the queue in an invisible multi-server queue with impatient customers. With the consideration of waiting costs and service benefits, they proved the existence and uniqueness of the equilibrium solution. Li and Wang [16] considered catastrophes in a single-server retrial queue with constant retrial rate. They derived the equilibrium joining/balking strategies under unobservable and observable cases with explicit theoretic proof. Recently, Ke *et al.* [8] discussed retrial and balking behavior in an unreliable service system. They employed the Probabilistic Global Search Lausanne algorithm to solve the cost optimization issue and applied the obtained results to an application of a telephone medical consulting service system. These days, many financial institutions, government agencies, and retail stores lack physical queues. An arriving customer can see the number on the ticket that he or she is issued and the number of tickets currently being served. Based on the difference between the two numbers, it is possible to estimate the current waiting time. In these situations, impatient customers commonly opt not to join the queue (balk), so the actual waiting time is less than expected. Consequently, compared with classical service systems with physical queues, customers' balking more significantly affects the characteristics of ticket queues. Subsequently, we provide a brief review of related literature in ticket queues.

For a single-server Markovian ticket queue with impatient customers and a threshold balking policy, Xu [26] developed an efficient scheme by which to approximate the performance of the ticket queue. Kuzu [12] introduced a multi-server Markovian ticket queue with balking and reneging behaviors and then compared the performance of this ticket queue with a physical queue. In a series of surveys, Kuzu [13] determined that the preference of customers for ticket queues over physical queues translates into increased patience. Based on their investigation of a single-server Markovian ticket queue with reneging customers, Ding *et al.* [3] developed an approximation procedure to numerically solve steady-steady results with an extension to multi-server systems. Considering the situation of customer abandonment, Jennings and Pender [7] conducted a comparative analysis of the ticket queues and standard queues with physical waiting lines. They proved the heavy traffic limit theorem for ticket queues and standard queue processes. Kerner *et al.* [10] proved that no threshold strategy can attain the Nash equilibrium for an infinite Markovian ticket queue with a homogeneous cost-reward function. They also demonstrated that the double threshold strategy is optimal for a cost function and the cost function is an increase function in waiting time. Kuzu and Soyer [15] established a Bayesian model to predict customer abandonments in ticket queues. Based on actual abandonment data collected from a bank, numerical results and managerial insights were presented. Kuzu *et al.* [14] conducted the first empirical analysis of customer behaviors in ticket queues with a focus on forecasting and dynamic decision-making. Hanukov *et al.* [4] studied a ticket queue with a nonhomogeneous population comprising regular and strategic customers. They derived steady-state results with the sojourn time and provided a novel approach to economic analysis aimed at determining the optimal mean orbiting time of strategic customers. For further research on ticket queues, the readers are referred to Hanukov *et al.* [5], Poomrittigul *et al.* [20], and Xiao *et al.* [25].

The above studies have suggested the importance of customer’s balking behavior in evaluating system performance in ticket queues. In addition to customers’ balking, server

vacation is a common attribute in most practical applications. Ke *et al.* [9] and Upadhyaya [24] provided comprehensive reviews of queueing systems with vacations. For a single-server Markovian queue with a single working vacation and multiple vacations, Tian and Wang [22] conducted a pricing analysis in unobservable cases. The customers' equilibrium and the socially optimal strategies were derived under a linear reward-cost structure. Jain *et al.* [6] proposed a generalized Markovian queue by including the characteristics of working vacation, retrial, and customers' balking. They employed the probability generating function technique to obtain the steady-state probabilities as well as several system performance metrics. Tian *et al*. [23] investigated a Markovian queue with working vacation and Bernoulli interruptions. Four scenarios with different levels of system information were considered for examining the behavior of strategic customers. The customer's equilibrium and social-optimal strategies were derived explicitly. Kumar and Jain [11] discussed a single-server queue with balking, a bi-level service network, and a bi-level vacation policy. They proved that a mixed vacation policy can tackle the congestion problem economically and reduce the mean waiting time. Recently, Sun *et al.* [21] analyzed customer balking behavior in single-server observable queues with geometric abandonments and two types of *N* policy. They determined that a residual vacation time reduces the likelihood of customers joining.

Although ticket queues and server vacations are both very common in real-world service systems, however, no research introduces vacation policy in ticket queues. Thus, this research performs the steady-state analysis on a single-server ticket queue with a single vacation policy. The purpose is to examine the effects of server vacation on the system performance. Our contributions include:

(1) This research firstly incorporates the single vacation policy in a ticket queue with balking customers, which has never been examined in literature;
(2) To simplify the analysis procedure, an approximation model is introduced to enhance the computation efficiency and improve implementation convenience;
(3) We establish a regression model for practitioners and system managers to rapidly evaluate the expected number of customers based only on the estimated system parameters without complex calculations.
(4) For management decision-making, the optimal service and vacation rates for the triple-objective problem are tabulated with graphical displays.

In the next subsection, a practical application that the proposed model can be applied is provided.

1.1. Practical application

Ticket queues can be widely observed in different industries, such as pharmacies, government offices, financial institutions, and retrial stores. In these practical service systems, the service provider may temporarily leave the system for other tasks whenever the system becomes empty. These may reduce service efficiency, increase customer waiting time, and influence the system's performance. For example, take-a-number machines for queue call systems are common equipment in banks. The demands from customers who come to the bank can be roughly classified into two categories: document application

processing (business loans, financial advisory, insurance, valuation services, etc) and financial teller service (account handling, cash withdrawal, currency exchange, inter-bank remittances, wealth management, etc). The numbers assigned to customers belonging to two categories are assumed independent. That is, there are two ticket queues in this service system. When an arriving customer finds the number of tickets in front of him/her is larger than a certain threshold, he/she may determine not to join the queue because the anticipated waiting time is too long, i.e., customers' balking behavior. Considering a bank counter responsible for demands of document application, when the service for the last customer is completed and there are no other customers in the corresponding ticket queue, the staff may temporarily leave the job for other secondary tasks (document delivery/archiving, computer file upload, support of financial teller service, to the restroom, etc). Often, these works are processed in batches while waiting for a break, and thus, the staff will return to his/her position, i.e., single vacation policy. With the above characteristics, the proposed ticket queue and the corresponding analysis results can be employed to evaluate the system performance and efficiency of the investigated application.

The remainder of this paper is organized as follows. Section 2 introduces the mathematical notation and model assumptions of the proposed ticket queue. The state transition-rate diagram for the proposed model is provided. Section 3 outlines our steady-state analysis based on the matrix-geometric method as well as an approximation procedure aimed at facilitating the solution process. In Section 4, we outline the closed-form formulations of critical performance metrics as well as sensitivity analysis aimed at identifying decisive parameters under a set of prescribed values. We also present a graphical illustration of the numerical results used to assess the accuracy of the approximations. Furthermore, we present a step-wise regression model for the estimation of the expected number of customers in the system. Finally, a tipple-objective optimization problem for determining the optimal service and vacation rates is performed in Section 5. We apply the NSGA-II algorithm to obtain the Pareto-optimal frontier graphically and tabulate partial non-dominated solutions. Conclusions and avenues for future research are presented in Section 6.

2. Model and preliminaries

This paper investigates a ticket queue with balking customers and single vacation policy. The notation and assumptions underlying the mathematical modeling are as follows:

- The customer arrival process is assumed to follow a Poisson distribution with mean arrival rate λ.
- The service time for each customer is assumed to follow an exponential distribution with mean service rate μ.
- The length of the vacation period is assumed to follow an exponential distribution with mean ϕ^{-1}. When the server completes a vacation with no customers waiting in the queue, he/she awaits idly for a new arrival, i.e., single vacation policy.

- Upon arriving at the system, each customer is issued a numbered ticket. Let D denote the ticket position, which is defined as the number difference between his/her ticket number and the displayed number.
- A new arrival customer will balk if the obtained ticket position is greater than or equal to a given threshold, denoted by K. Thus, the value of K is positively correlated with the patience of the customer. Once a customer decides to join the queue, no reneging is allowed.
- The only information available to customers in the ticket queue is the difference between the number of tickets currently receiving service and the number on their ticket. The number of customers in the system is indicated by N.
- Neither the service provider nor the customers know which customers have balked until the corresponding numbers have been called. The system can only serve one customer at a time on a first-come-first-served basis. The stochastic processes involved in this model (i.e., arrival, service, and vacation) are mutually independent.

In accordance with the above notation and assumptions, the system in a steady state can be defined in vector form as $\mathbf{x} \equiv (\omega, \mathbf{t}) \equiv (\omega, t_1, t_2, ..., t_n)$, where $\omega = 0$ ($\omega = 1$) indicates that the server is working normally (on vacation). The variable $t_i \in \mathbb{N}$ is a positive integer indicating the number of tickets issued between the i^{th} joining customer and the $(i+1)^{th}$ joining customer. As n is equal to the number of customers in the system, we know that the length of the state vector n cannot exceed the balking threshold K. Furthermore, the cumulative number of tickets between the first customer and the $(n-1)^{th}$ customer, $\sum_{i=1}^{n-1} t_i$, must be less than the balking threshold K. The associated state space can be described as follows:

$$\mathbf{S} = \left\{ (\omega, t_1, t_2, ..., t_n) \middle| \omega \in \{0,1\};\ t_i \in \mathbb{N};\ n \in \mathbb{N};\ \sum_{i=1}^{n-1} t_i < K \right\}. \tag{1}$$

Suppose the balking threshold $K = 15$, if we expand the compact state $\mathbf{x} = (0,3,2,2,4)$ as $(0,3,2,2,4) = (0,\ \underline{1,0,0},\ \underline{1,0},\ \underline{1,0},\ \underline{1,0,0,0})$, then the server is working normally. We also know that the first, fourth, sixth, and eighth customers joined the queue, but the other customers did not. The number of customers in the system is $N = 4$ and the ticket position $D = 3+2+2+4 = 11$ indicates the number of customers observed by the next arriving customer, which is significantly less than the number of customers actually in the system. Figure 1 depicts the state transitions for the investigated ticket queue with $K = 3$. For illustration purposes, if the system is currently in the state $\mathbf{x} = (\omega, \mathbf{t}) = (0,1,2)$, the next state after a transition may be $\mathbf{x} = (0,2)$ or $\mathbf{x} = (0,1,3)$, which are associated with the events of customer departure and the arrival of a new customer, respectively. Moreover, if the system is current in $(1,1,2)$, the status will transfer to $(0,1,2)$ once the vacation period

ends. Figure 2 presents a schematic diagram of the state transitions in the proposed ticket queue, where $1_{\{\bullet\}}$ denotes the indicator function and $\|\mathbf{t}\|_1 = \sum_{i=1}^{n} t_i$ denotes the total tickets in $\mathbf{t} = (t_1, t_2, ..., t_n)$.

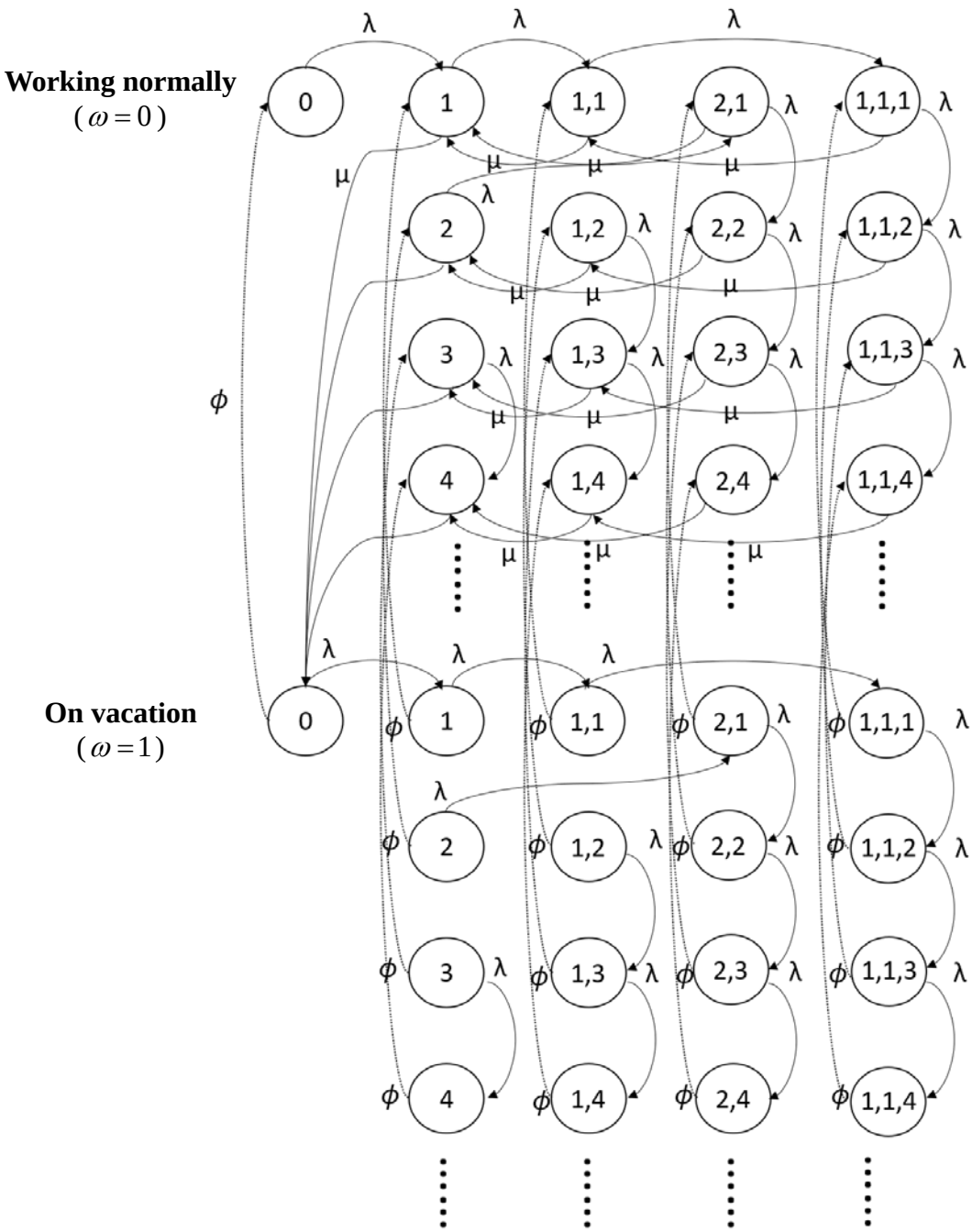


Figure 1. State-transition-rate diagram for the ticket queue with balking customers and single vacation policy ($K = 3$).

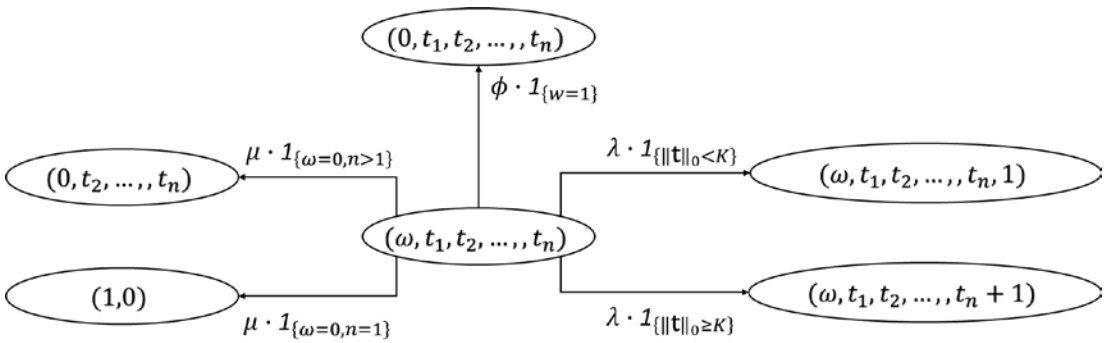


Figure 2. Schematic diagram of state transitions in ticket queue with balking customers and single vacation policy

With the state space provided in Equation (1), the steady-state probability can be denoted by $p_{\mathbf{t}}^{\omega}$, $(\omega, \mathbf{t}) \in \mathbf{S}$. Note that this system forms an ergodic Markov chain as long as the balking limit *K* is finite. The process of analysis can be simplified by partitioning the system state according to the last variable t_n to obtain

$$\mathbf{T}_1 = (0,0) \cup (1,0) \cup \left\{ (\omega, t_1, t_2, ..., t_n) \middle| t_n = 1;\ \omega \in \{0,1\};\ t_i \in \mathbb{N};\ n \in \mathbb{N};\ \sum_{i=1}^{n-1} t_i < K \right\}, \quad (2)$$

and

$$\mathbf{T}_j = \left\{ (\omega, t_1, t_2, ..., t_n) \middle| t_n = j;\ \omega \in \{0,1\};\ t_i \in \mathbb{N};\ n \in \mathbb{N};\ \sum_{i=1}^{n-1} t_i < K \right\}, \quad j = 2,3,.... \quad (3)$$

Note that the number of states collected in $\mathbf{T}_1$ and $\mathbf{T}_j$, $j = 2,3,...$ are $2^K + 2$ and 2^K, respectively. Ordering these states lexicographically allows us to represent distribution $\left\{ p_{(\mathbf{t})}^{\omega},\ (\omega, \mathbf{t}) \in \mathbf{S} \right\}$ using vector $\mathbf{\Pi} = [\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \boldsymbol{\pi}_3, ...]$. This can be expressed as

$$\boldsymbol{\pi}_1 = \left[p_{(0)}^0, p_{(1)}^0, p_{(1,1)}^0, p_{(2,1)}^0, ..., p_{(1,1,...,1)}^0,\ p_{(0)}^1, p_{(1)}^1, p_{(1,1)}^1, p_{(2,1)}^1, ..., p_{(1,1,...,1)}^1 \right], \quad (4)$$

and

$$\boldsymbol{\pi}_j = \left[p_{(j)}^0, p_{(1,j)}^0, p_{(2,j)}^0, ..., p_{(1,1,...,j)}^0,\ p_{(j)}^1, p_{(1,j)}^1, p_{(2,j)}^1, ..., p_{(1,1,...,j)}^1 \right], \quad j = 2,3,..., \quad (5)$$

which are associated with each partitioned sub-space. Based on the above notation and definitions, the proposed ticket queue is a quasi-birth-and-death process (Neuts [19], Chapter 3) with the following block-partitioned infinitesimal generator:

$$\mathbf{Q} = \begin{array}{c} \mathbf{T}_1 \\ \mathbf{T}_2 \\ \mathbf{T}_3 \\ \vdots \\ \mathbf{T}_{K-1} \\ \mathbf{T}_K \\ \mathbf{T}_{K+1} \\ \mathbf{T}_{K+2} \\ \vdots \end{array} \begin{bmatrix} \mathbf{A}_1 & \mathbf{B}_1 & & & & & & & \\ \mathbf{C}_2 & \mathbf{A}_2 & \mathbf{B}_2 & & & & & & \\ \mathbf{C}_3 & & \mathbf{A}_2 & \mathbf{B}_3 & & & & & \\ \vdots & & & \ddots & \ddots & & & & \\ \mathbf{C}_{K-1} & & & & \mathbf{A}_2 & \mathbf{B}_{K-1} & & & \\ \mathbf{C}_K & & & & & \mathbf{A}_2 & \mathbf{B}_K & & \\ \mathbf{C}_K & & & & & & \mathbf{A}_2 & \mathbf{B}_K & \\ \mathbf{C}_K & & & & & & & \mathbf{A}_2 & \ddots \\ \vdots & & & & & & & & \ddots \end{bmatrix}. \tag{6}$$

Let $\mathbf{O}_\ell$ be a square zero matrix of order ℓ and $\mathbf{I}_\ell$ be an identity matrix of order ℓ, while $\mathbf{0}$ is a row zero vector of appropriate dimensions and $\mathbf{e}$ is a unit row vector of appropriate dimensions. Solving the balance equation $\mathbf{\Pi Q} = \mathbf{0}$ and the normalization condition $\mathbf{\Pi e} = 1$ allows us to compute steady-state probability $\mathbf{\Pi}$. Identifying the locations of non-zero elements in $\mathbf{Q}$ is complicated by their relationship with the balking threshold K; to illustrate, we provide a special case where $K = 3$.

When $K = 3$, the submatrices employed in $\mathbf{Q}$ include $\mathbf{A}_1$, $\mathbf{A}_2$, $\mathbf{B}_i$, $i = 1,2,3$ and $\mathbf{C}_i$, $i = 2,3$. Defining $\delta = \lambda + \mu$ and $\tau = \lambda + \phi$ allows us to represent $\mathbf{A}_1$ and $\mathbf{A}_2$ as follows:

$$\mathbf{A}_1 = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{13} & \mathbf{A}_{14} \end{bmatrix} \text{ and } \mathbf{A}_2 = \begin{bmatrix} \mathbf{A}_{21} & \mathbf{O}_4 \\ \mathbf{A}_{23} & \mathbf{A}_{24} \end{bmatrix}. \tag{7}$$

$$\mathbf{A}_{11} = \begin{bmatrix} -\lambda & \lambda & & & \\ & -\delta & \lambda & & \\ & \mu & -\delta & & \lambda \\ & \mu & & -\delta & \\ & & \mu & & -\delta \end{bmatrix}, \quad \mathbf{A}_{14} = \begin{bmatrix} -\tau & \lambda & & & \\ & -\tau & \lambda & & \\ & & -\tau & & \lambda \\ & & & -\tau & \\ & & & & -\tau \end{bmatrix}.$$

$\mathbf{A}_{12}[2,1]$ is a square matrix of order 5 with only one non-zero element $\mathbf{A}_{12}[2,1] = \mu$ and $\mathbf{A}_{13} = \phi\mathbf{I}_5$. Similarly, the submatrices in $\mathbf{A}_2$ are $\mathbf{A}_{23} = \phi\mathbf{I}_4$, $\mathbf{A}_{24} = -\tau\mathbf{I}_4$, and

$$\mathbf{A}_{21} = \begin{bmatrix} -\delta & & & \\ \mu & -\delta & & \\ \mu & & -\delta & \\ & \mu & & -\delta \end{bmatrix}.$$

$\mathbf{B}_1$ is a 10×8 matrix with nonzero elements $\mathbf{B}_1[4,3] = \mathbf{B}_1[5,4] = \mathbf{B}_1[9,7] = \mathbf{B}_1[10,8] = \lambda$. $\mathbf{B}_2$ is a 8×8 diagonal matrix with diagonal elements $[0, \lambda, \lambda, \lambda, 0, \lambda, \lambda, \lambda]$ and $\mathbf{B}_3 = \lambda\mathbf{I}_8$.

$\mathbf{C}_2$ is a 8×10 matrix with nonzero elements $\mathbf{C}_2[2,4]=\mathbf{C}_2[6,10]=\lambda$ and $\mathbf{C}_2[1,6]=\mu$.
$\mathbf{C}_3$ is also a 8×10 matrix with only one nonzero element $\mathbf{C}_3[1,6]=\mu$.

3. Steady-state analysis

Newly-arriving customers balk if and only if the obtained ticket position exceeds the balking threshold, which means that if K is finite, then the system will remain stable, regardless of the traffic intensity. Applying the matrix analytic method and expanding the balance equation $\mathbf{\Pi Q}=\mathbf{0}$ implies the following:

$$\boldsymbol{\pi}_1\mathbf{A}_1+\sum_{i=2}^{K-1}\boldsymbol{\pi}_i\mathbf{C}_i+\sum_{i=K}^{\infty}\boldsymbol{\pi}_i\mathbf{C}_K=\mathbf{0}, \tag{8}$$

$$\boldsymbol{\pi}_i\mathbf{B}_i+\boldsymbol{\pi}_{i+1}\mathbf{A}_2=\mathbf{0},\ i=1,2,...,K-1, \tag{9}$$

$$\boldsymbol{\pi}_i\mathbf{B}_K+\boldsymbol{\pi}_{i+1}\mathbf{A}_2=\mathbf{0},\ i=K,K+1,.... \tag{10}$$

Based on a technique similar to the well-known matrix-geometric method, we can define the rate matrix as $\mathbf{R}=\mathbf{B}_K\left(-\mathbf{A}_2^{-1}\right)$. Thus, Equations (8)-(10) imply the following:

$$\boldsymbol{\pi}_i=\boldsymbol{\pi}_K\mathbf{R}^{i-K},\ i=K+1,K+2,..., \tag{11}$$

$$\boldsymbol{\pi}_{i+1}=\boldsymbol{\pi}_1\phi_1\phi_2...\phi_i=\boldsymbol{\pi}_1\prod_{j=1}^{i}\phi_j,\ \ i=1,2,3,...,K-1, \tag{12}$$

where $\phi_j=\mathbf{B}_j\left(-\mathbf{A}_2^{-1}\right),\ \ i=1,2,3,...$. Substituting these equations into Equation (8) results in the following matrix-form balance equation for $\boldsymbol{\pi}_1$:

$$\boldsymbol{\pi}_1\left[\mathbf{A}_1+\sum_{i=2}^{K-1}\prod_{j=1}^{i-1}\phi_j\mathbf{C}_i+\prod_{j=1}^{K-1}\phi_j\left(\mathbf{I}_{2^K}-\mathbf{R}\right)^{-1}\mathbf{C}_K\right]=\boldsymbol{\pi}_1\mathbf{\Psi}=\mathbf{0}. \tag{13}$$

The normalization condition can be represented as

$$\boldsymbol{\pi}_1\left\{\mathbf{e}^{\mathrm{T}}+\left[\sum_{i=2}^{K-1}\prod_{j=1}^{i-1}\phi_j+\prod_{j=1}^{K-1}\phi_j\left(\mathbf{I}-\mathbf{R}\right)^{-1}\right]\mathbf{e}^{\mathrm{T}}\right\}=\boldsymbol{\pi}_1\mathbf{z}=1. \tag{14}$$

Thus, $\boldsymbol{\pi}_1$ can be obtained by simultaneously solving Equations (13) and (14).

3.1 Approximation method

With Figure 1, it can be expected that the system state transition is very complicated once the value of the parameter K is large. The fact that the number of states considered in the subspace $\mathbf{T}_1$ is 2^K+2 makes it an exponential function of parameter K. For

example, if $K=5$, then the cardinality of $\mathbf{T}_1$ is $|\mathbf{T}_1|=2^5+2=34$. If $K=10$, then $|\mathbf{T}_1|=2^{10}+2=1026$ is approximately thirty times as much as the former. This means that the computation time and memory space required for the above procedure rapidly increases with the value of K. Thus, we developed a heuristic solution to improve efficiency in computing the ticket queue for large K values. Keeping track only of the total number of customers who join or balk while disregarding the sequence would result in a system of reduced complexity, which disregards intermix situations. The fact that the balking behavior in the reduced system is identical to that in the original system, implies their stochastic characteristics are also similar. The state space for the reduced ticket queue can be expressed as follows:

$$\mathbf{S}^R=(0,0)\cup(1,0)\cup\left\{(\omega,J,B)\middle|\omega\in\{0,1\};\ J=0,1,2,...,K;\ B\in\mathbb{N}\right\}. \tag{15}$$

where $(0,0)$ and $(1,0)$ respectively indicate that the system is empty when the server is idle and busy. Variables J and B respectively denote the number of joining customers (including the customer being served, if any) and the number of balking customers in the system. The associated steady-state probability is denoted by q_0^0, q_0^1, and $q_{(J,B)}^{\omega}$, $(\omega,J,B)\in\mathbf{S}^R$. The state space $\mathbf{S}^R$ can be partitioned as:

and

$$\mathbf{T}_j^R=\left\{(\omega,J,B)\middle|\omega\in\{0,1\};\ J=0,1,2,...,K;\ B=j-1\right\},\quad j=2,3,.... \tag{17}$$

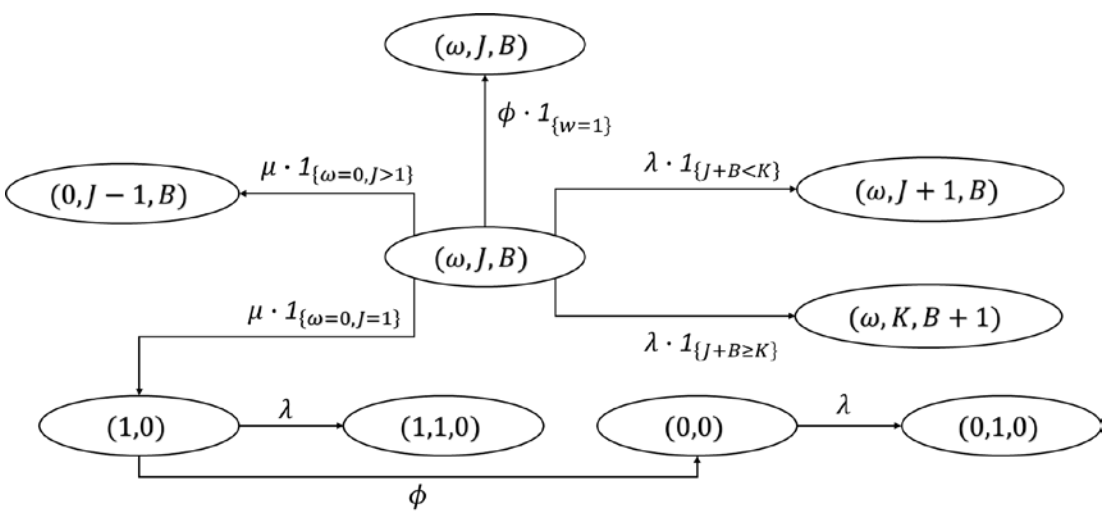


Figure 3. Schematic diagram of state transitions of the reduced ticket queue

Unlike the cardinality mentioned above $|\mathbf{T}_1|=2^K+2$, the cardinality of $\mathbf{T}_1^R$ is $|\mathbf{T}_1^R|=2K+2$, which is a linear function of K. For example, if $K=10$, then the cardinality

of $\mathbf{T}_1^R$ is $\left|\mathbf{T}_1^R\right| = 22$, which is only 2.1% of $\left|\mathbf{T}_1\right| = 1026$. Figure 3 presents a schematic diagram of the reduced ticket queue. Note that the state transitions in the reduced system are significantly simpler than those in the original system. The approximated steady-state probabilities obtained from the reduced system are denoted by $\mathbf{\Pi}^a = \left[\boldsymbol{\pi}_1^a, \boldsymbol{\pi}_2^a, \boldsymbol{\pi}_3^a, ...\right]$, where

$$\boldsymbol{\pi}_1^a = \left[q_0^0, q_{(1,0)}^0, q_{(2,0)}^0, ..., q_{(K,0)}^0, q_0^1, q_{(1,0)}^1, q_{(2,0)}^1, ..., q_{(K,0)}^1\right], \tag{18}$$

and

$$\boldsymbol{\pi}_j^a = \left[q_{(1,j-1)}^0, q_{(2,j-1)}^0, ..., q_{(K,j-1)}^0, q_{(1,j-1)}^1, q_{(2,j-1)}^1, ..., q_{(K,j-1)}^1\right], \quad j = 2,3,..., \tag{19}$$

which are associated with each partitioned sub-space. With the exception of the two additional states, this is similar to the solution procedure mentioned earlier for the original system, wherein the matrix-analytic method can be used to establish steady-state probabilities for the reduced system.

4. System performance

In the previous section, we outline an approximation of steady-state probabilities for the proposed ticket queue. In this section, to evaluate the service level, we define several critical system performance metrics and derive corresponding closed-form formulations. For convenience, we respectively use $\mathbf{0}_\ell$ and $\mathbf{e}_\ell$ to denote the row zero vector and row identity vector of order ℓ.

- The exact and approximated expected number of (joining) customers in the system are respectively denoted by $E(N)$ and $E(N)^a$, as follows:

$$E(N) = \sum_{\omega=0}^{1} \sum_{i=1}^{K} \sum_{\|\mathbf{t}\|_0 = i} i p_{\mathbf{t}}^{\omega}, \tag{18}$$

$$E(N)^a = \boldsymbol{\pi}_1^a \times \mathbf{v}_0 + \sum_{j=2}^{\infty} \boldsymbol{\pi}_j^a \mathbf{v}, \tag{19}$$

  where $\mathbf{v}_0 = [0,1,2,...,K,0,1,2,...,K]^{\mathrm{T}}$ and $\mathbf{v} = [1,2,...,K,1,2,...,K]^{\mathrm{T}}$.

- The exact and approximated probability that the server is idle are respectively denoted by P_{ID} and P_{ID}^a, as follows:

$$P_{ID} = p_{(0)}^0 + p_{(0)}^1 = \boldsymbol{\pi}_1 \times [1, \mathbf{0}_{2^{K-1}}, 1, \mathbf{0}_{2^{K-1}}], \tag{20}$$

$$P_{ID}^a = q_0^0 + q_0^1 = \boldsymbol{\pi}_1^a \times [1, \mathbf{0}_K, 1, \mathbf{0}_K]. \tag{21}$$

- The probability of a given customer balking is denoted by P_b, as follows:

$$P_b = \boldsymbol{\pi}_1 \mathbf{u}_{2^K} + \boldsymbol{\pi}_1 \left\{ \sum_{i=1}^{K-2} \prod_{j=1}^{i} \phi_j + \prod_{j=1}^{K-1} \phi_j \left(\mathbf{I}_{2K} - \mathbf{R} \right)^{-1} \right\} \mathbf{u}_{2^K-1}, \tag{22}$$

where $\mathbf{u}_\ell = [\mathbf{0}_\ell, 1, \mathbf{0}_\ell, 1]^{\mathrm{T}}$.

The approximate probability of a given customer balking, P_b^a, can be evaluated in terms of $\boldsymbol{\pi}_1^a$ and the approximation of matrices ϕ_j and $\mathbf{R}$.

4.1 Sensitivity analysis

Sensitivity analysis was used to investigate the influence of each system parameter on various metrics of system performance. In the following, we provide numerical evidence justifying the use of approximations of the original ticket queue derived using the reduced system. These results are helpful for managers seeking to identify variables crucial to decision-making. We began with the following initial parameter values: $K = 5$, $\lambda = 5$, $\mu = 5$, and $\phi = 8$. We then adjusted the value of one parameter at a time (from 1 to 9). The results are presented in Figures 4 and 5 in the form of snapshots of $E(N)$ and P_b (solid lines) and the corresponding $E(N)^a$ and P_b^a values (dashed lines).

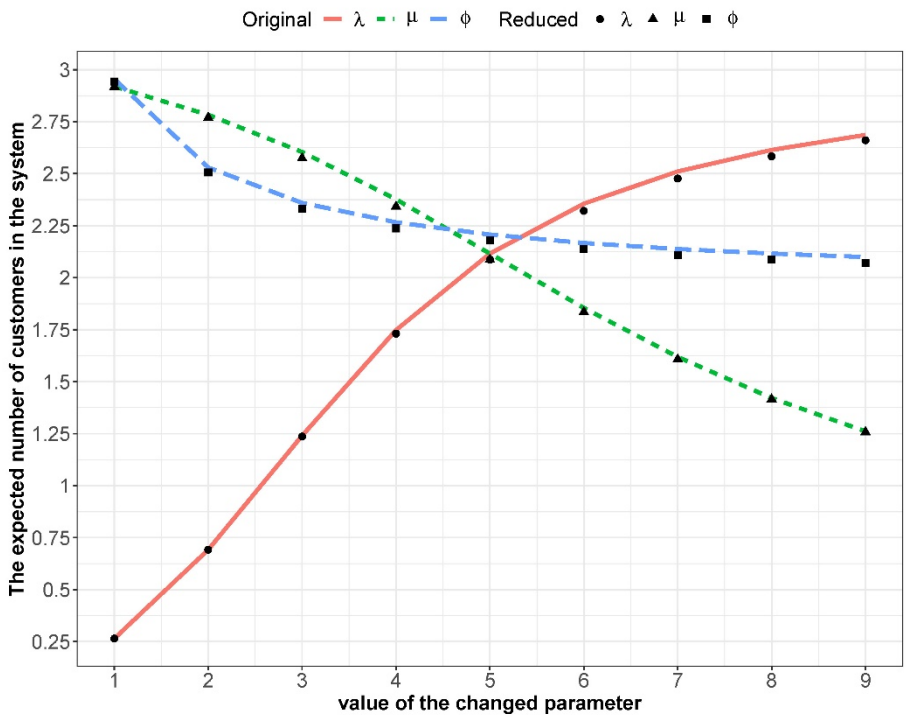


Figure 4. $E(N)$ and $E(N)^a$ as functions of λ, μ and ϕ from 1(1)9

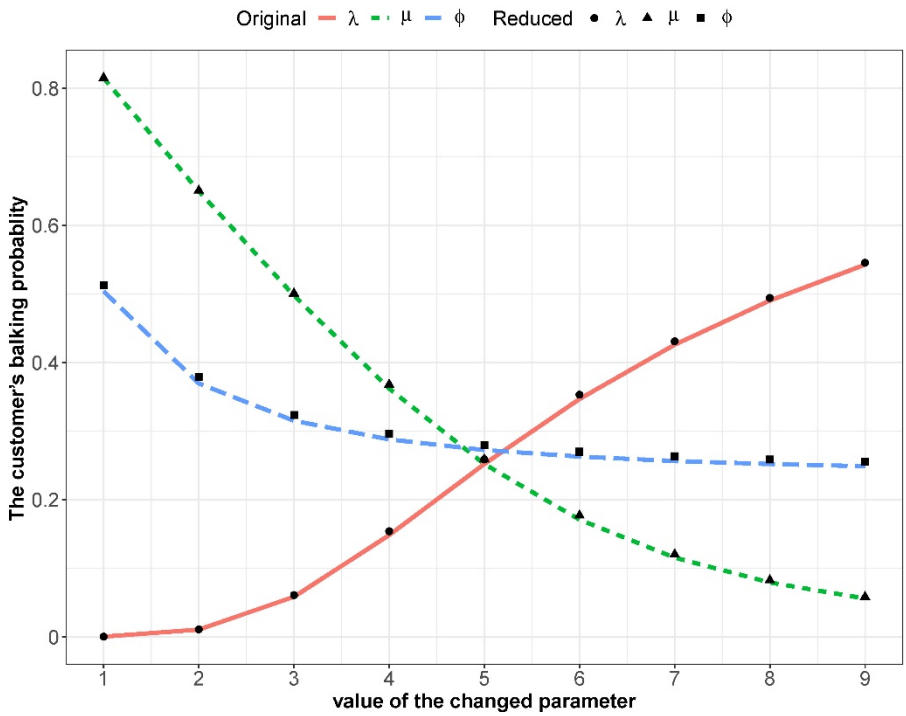


Figure 5. P_b and P_b^a as functions of λ, μ and ϕ from 1(1)9

Table 1. APE $E[N]^a$ and P_b^a (in %) when changing one parameter at a time

	$\left\|E[N]^a - E[N]\right\| / E[N] \times 100\%$			$\left\|P_b^a - P_b\right\| / P_b \times 100\%$		
	λ	μ	ϕ	λ	μ	ϕ
1	<0.0001	0.0435	0.4057	0.5083	0.0036	1.6893
2	0.0383	0.4793	0.9316	2.9124	0.1230	2.4716
3	0.3818	1.1425	1.1683	4.2586	0.6299	2.6949
4	0.9823	1.4832	1.2741	3.7428	1.5721	2.7413
5	1.3872	1.3872	1.3266	2.6756	2.6756	2.7348
6	1.4755	1.0636	1.3561	1.7819	3.5840	2.7149
7	1.3635	0.7231	1.3746	1.1700	4.0985	2.6940
8	1.1711	0.4596	1.3872	0.7739	4.2179	2.6756
9	0.9671	0.2830	1.3964	0.5196	4.0469	2.6602

As shown in Figures 4 and 5, $E(N)$ and P_b significantly increased with the mean arrival rate, particularly when λ was less than μ. Increasing the service rate was shown to reduce queue length and reduce the likelihood of balking. Shortening the vacation period did not provide significant benefits. Table 1 lists the calculated absolute percentage errors (APE) $\left|E[N]^a - E[N]\right| / E[N] \times 100\%$ and $\left|P_b^a - P_b\right| / P_b \times 100\%$ as an indication of approximation accuracy. Note that the reduced system provides excellent approximation

results, as indicated by an APE of less than 5% and a worst-case APE of 4.3%. Table 1 also shows that APEs are concave functions of λ and μ, due to the low probability of joining and balking customers intermixing when the arrival rate or service rate is significantly high.

To assess the compound effect of impatience and the lengths of the arrival/service/vacation periods, we set $K = 2(2)8$ and changed the value of λ (μ and ϕ) from 1 to 9. The APE results indicating approximation accuracy are listed in Tables 2-4. The $E(N)$ and P_b curves in Figures 6-11 reveal the following patterns:

(1) For patient customers and a high K value, $E(N)$ and P_b increased with an increase in λ and decreased with an increase in or ϕ. This indicates that the effects of system parameters are more pronounced when the ticket queue includes a larger number of patient customers. In contrast, for impatient customers and a small K value, P_b will be very high. Under these conditions, the probability of balking can be reduced by improving service efficiency or shortening the vacation period.

(2) Overall, the effects of shortening the vacation period in reducing the likelihood of balking are somewhat limited since under a single vacation policy, queue length is determined mainly by service efficiency.

(3) When the mean arrival rate is low or the service rate is high, the low likelihood of balking diminishes the effect of K on system length. Note that the effects of K value did not vary with changes in ϕ.

(4) Our numerical results reveal that the behavior of a reduced ticket queue is similar to that of the original system in terms of the expected number of customers and the probability of balking. Note that when the K value was large, the accuracy of the approximations was lower, as indicated by APE values of 0% to 7.17%.

Taken together, these results indicate that the value of the parameter K must be estimated and monitored carefully, as it has a direct bearing on system performance, approximation accuracy, and estimates of customer patience. This indicates that to strengthen service capabilities, managers should focus on the training of employees rather than on shortening the vacation period.

Table 2. APE values of $E[N]^a$ and P_b^a (in %) under $\lambda = 1(1)9$ and $K = 2(2)8$

	$\left\|E[N]^a - E[N]\right\| / E[N] \times 100\%$				$\left\|P_b^a - P_b\right\| / P_b \times 100\%$			
λ	$K=2$	$K=4$	$K=6$	$K=8$	$K=2$	$K=4$	$K=6$	$K=8$
1		0.0008	<0.0001	<0.0001		0.7215	0.3163	0.1059
2		0.0503	0.0213	0.0044		2.5060	2.9231	2.4148
3		0.3395	0.3515	0.2227		3.0038	5.2145	6.3577
4		0.7152	1.1485	1.2656		2.4908	4.8740	6.8111
5	<0.0001	0.9395	1.7632	2.3452	<0.0001	1.7984	3.4435	4.6994
6		0.9894	1.8952	2.5667		1.2421	2.2056	2.7744
7		0.9316	1.7151	2.2143		0.8526	1.3841	1.5931
8		0.8264	1.4232	1.7081		0.5903	0.8753	0.9261
9		0.7094	1.1285	1.2467		0.4144	0.5629	0.5501

Table 3. APEs of $E[N]^a$ and P_b^a (in %) under $\mu = 1(1)9$ and $K = 2(2)8$

	$\left\|E[N]^a - E[N]\right\| / E[N] \times 100\%$				$\left\|P_b^a - P_b\right\| / P_b \times 100\%$			
μ	$K=2$	$K=4$	$K=6$	$K=8$	$K=2$	$K=4$	$K=6$	$K=8$
1		0.0658	0.0244	0.0050		0.0055	0.0022	0.0007
2		0.4252	0.4691	0.3735		0.1181	0.1136	0.0844
3		0.8300	1.3606	1.5832		0.4974	0.6959	0.7127
4		1.0027	1.8956	2.5457		1.1162	1.9182	2.3562
5	<0.0001	0.9395	1.7632	2.3452	<0.0001	1.7984	3.4435	4.6994
6		0.7574	1.2720	1.4730		2.3679	4.6844	6.5740
7		0.5574	0.7849	0.7232		2.7335	5.2872	7.1629
8		0.3887	0.4449	0.3127		2.8889	5.2765	6.6567
9		0.2630	0.2436	0.1299		2.8747	4.8647	5.6308

Table 4. APEs of $E[N]^a$ and P_b^a (in %) under $\phi = 1(1)9$ and $K = 2(2)8$

	$\left\|E[N]^a - E[N]\right\| / E[N] \times 100\%$				$\left\|P_b^a - P_b\right\| / P_b \times 100\%$			
ϕ	$K=2$	$K=4$	$K=6$	$K=8$	$K=2$	$K=4$	$K=6$	$K=8$
1		0.2113	0.6499	1.2002		0.8539	2.6801	4.7851
2		0.5425	1.3101	1.9269		1.4204	3.5235	5.3535
3		0.7313	1.5514	2.1284		1.6588	3.6509	5.2037
4		0.8286	1.6502	2.2132		1.7542	3.6200	5.0261
5	<0.0001	0.8806	1.6994	2.2633	<0.0001	1.7900	3.5642	4.8959
6		0.9100	1.7288	2.2984		1.8010	3.5140	4.8061
7		0.9279	1.7486	2.3247		1.8017	3.4741	4.7438
8		0.9395	1.7632	2.3452		1.7984	3.4435	4.6994
9		0.9475	1.7744	2.3614		1.7938	3.4200	4.6670

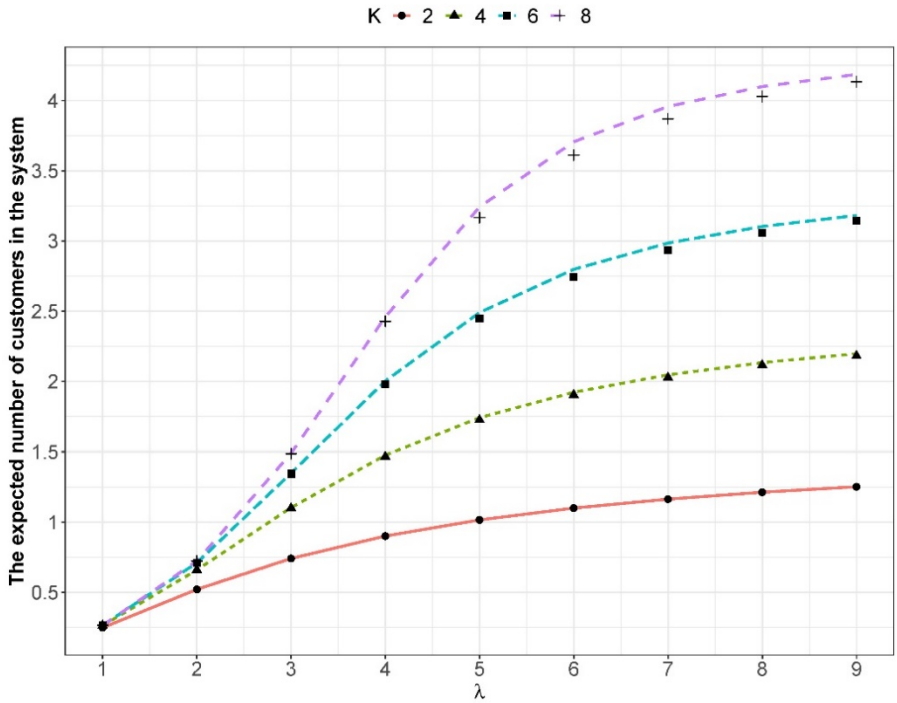


Figure 6. $E(N)$ and $E(N)^a$ as functions of $\lambda = 1(1)9$ and $K = 2(2)8$ (bottom to top)

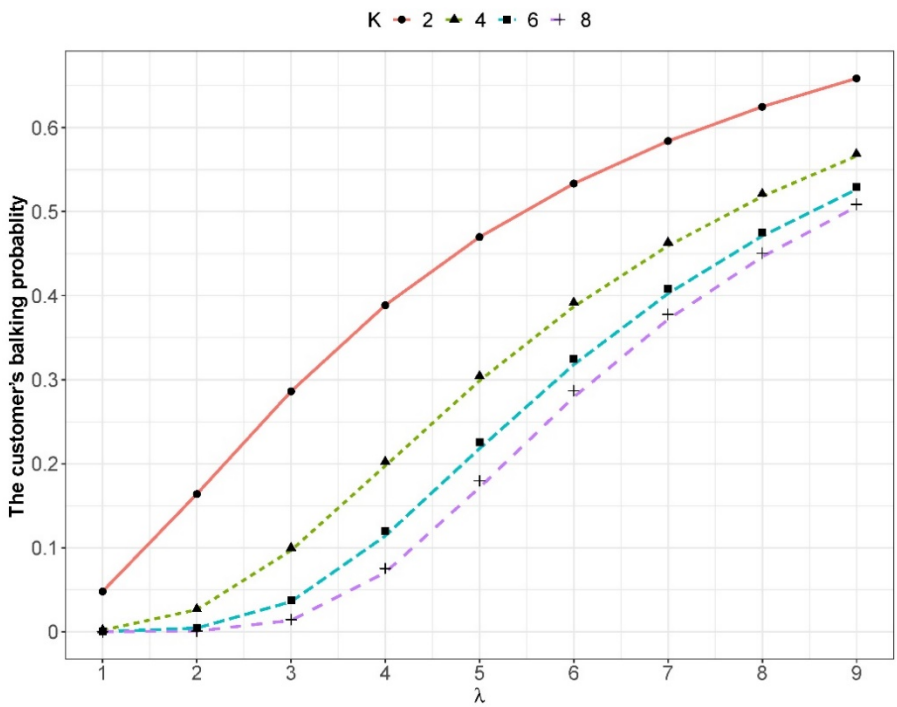


Figure 7. P_b and P_b^a as functions of $\lambda = 1(1)9$ and $K = 2(2)8$ (bottom to top)

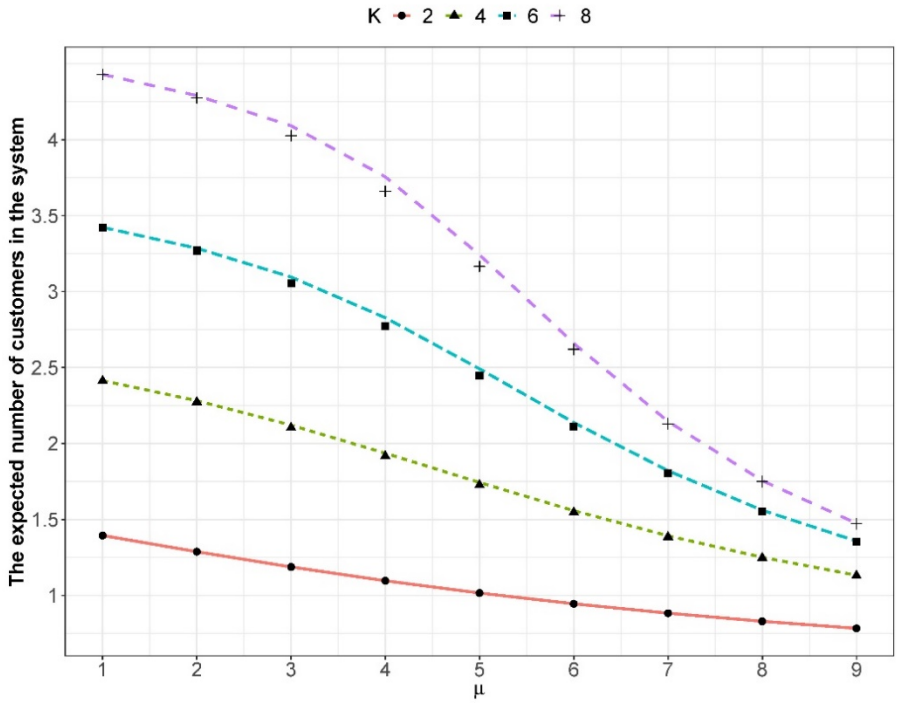


Figure 8. $E(N)$ and $E(N)^a$ as functions of $\mu = 1(1)9$ and $K = 2(2)8$ (bottom to top)

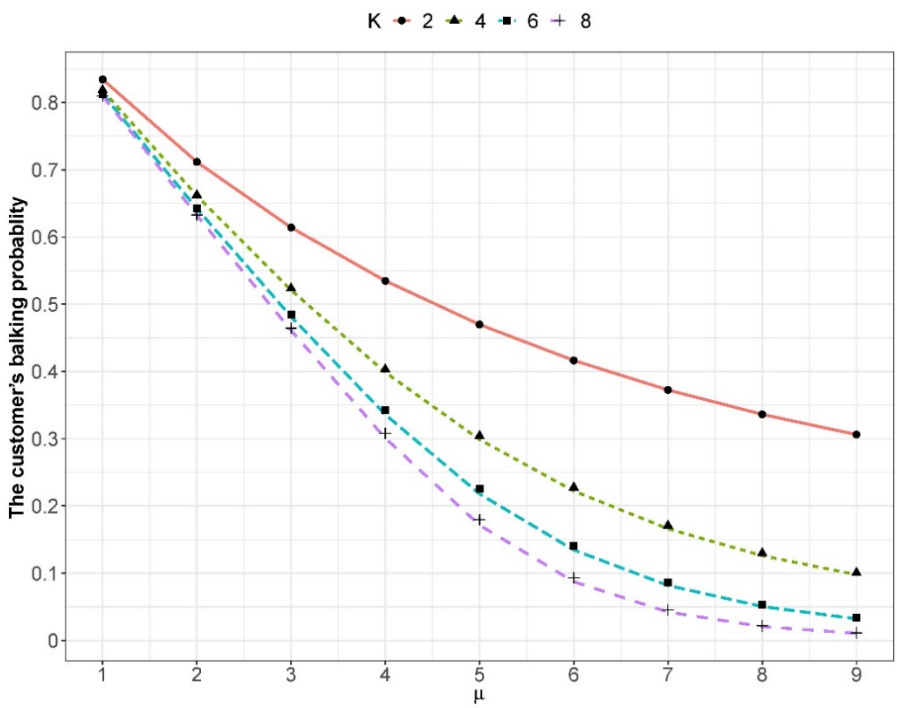


Figure 9. P_b and P_b^a against values of $\mu = 1(1)9$ and $K = 2(2)8$ (bottom to top)

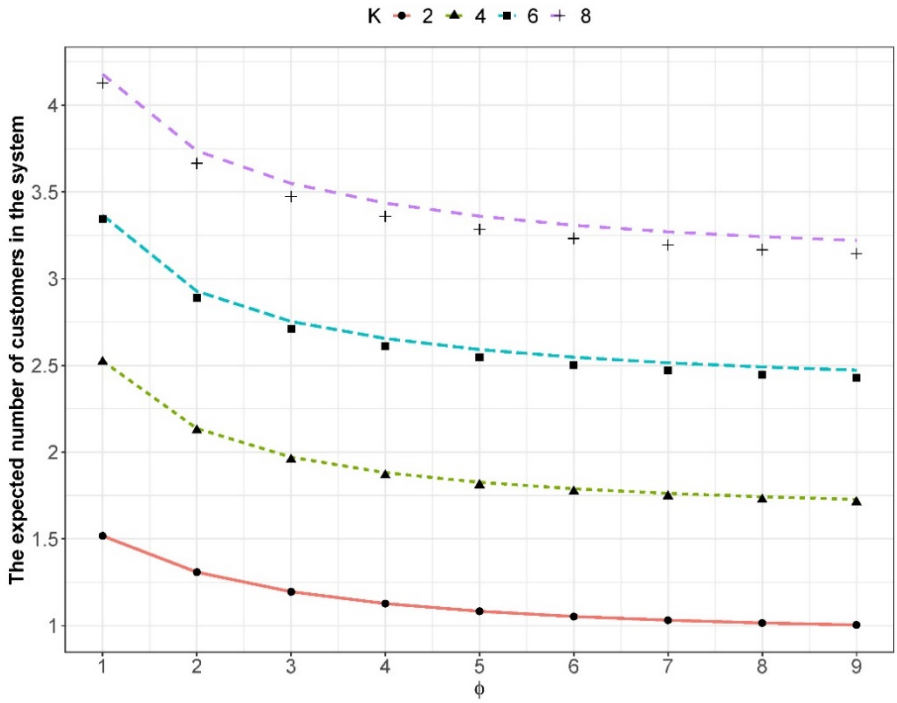


Figure 10. $E(N)$ and $E(N)^a$ as functions of $\phi = 1(1)9$ and $K = 2(2)8$ (bottom to top)

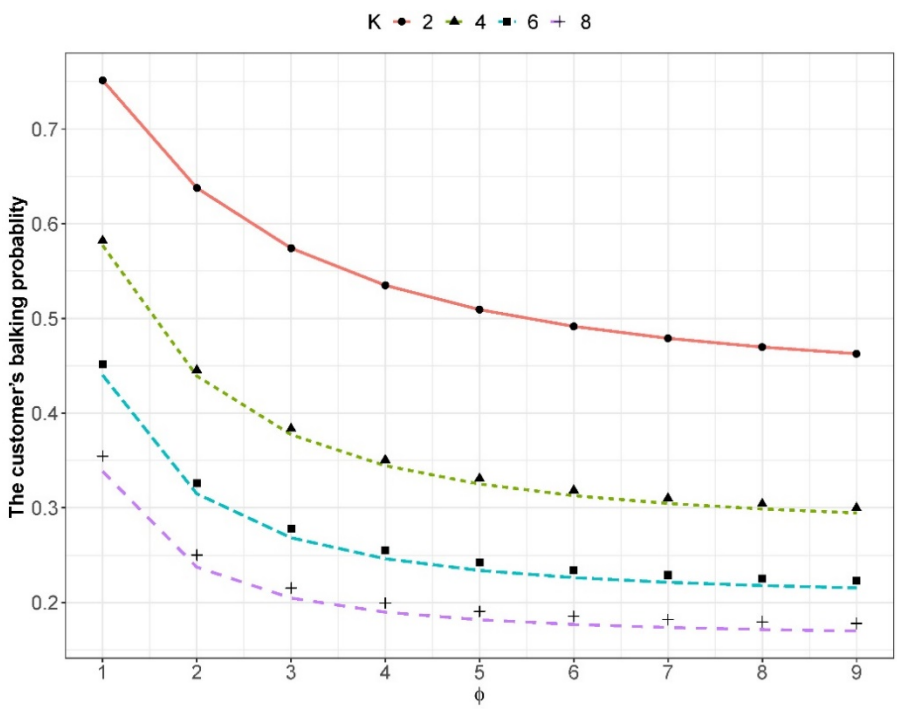


Figure 11. P_b and P_b^a as functions of $\phi = 1(1)9$ and $K = 2(2)8$ (bottom to top)

4.2 Regression model

For both the original and reduced systems, the complexity of the matrix analytical approach could lead to difficulties in implementation. Note also that matrix manipulations and recursive algorithms necessitate the use of a suitably enabled computer. Thus, we formulated regression models to determine the correlation between system parameters (independent variables) and $E[N]$ (dependent variable). $E[N]$ was calculated under a range of parameter values, including $K = 2(1)8$, $\lambda = 1(1)8$, $\mu = 1(1)8$, $\phi = 1(1)8$, which resulted in a total of $7 \times 8 \times 8 \times 8 = 3,584$ combinations. As shown in the figures above, $E[N]$ is a complex non-linear function of individual parameters and corresponding interactions. Thus, we expanded on the original variable set by adding traffic intensity $\rho = \lambda / \mu$ and weighted traffic intensity $\rho' = K\lambda / \mu$. For each independent variable (K, λ, μ, ϕ, ρ and ρ'), we applied reciprocal, natural logarithmic, and exponential functions to generate three additional variables. This resulted in a regression model with $6 \times (1+3) = 24$ independent variables. We derived the regression model via backward stepwise regression, as follows:

$$\begin{aligned} E[N] \approx & -0.0857 + 0.2426K - 0.0458\mu - 0.8275\rho + 0.0429\rho' + 1.0769\rho'^{-1} \\ & -1.8989K^{-1} + 0.4319\lambda^{-1} + 1.1793\mu^{-1} + 0.6781\phi^{-1} \\ & +2.4001\ln(\lambda) - 1.324\ln(\mu) - 0.0575\ln(\phi). \end{aligned} \tag{23}$$

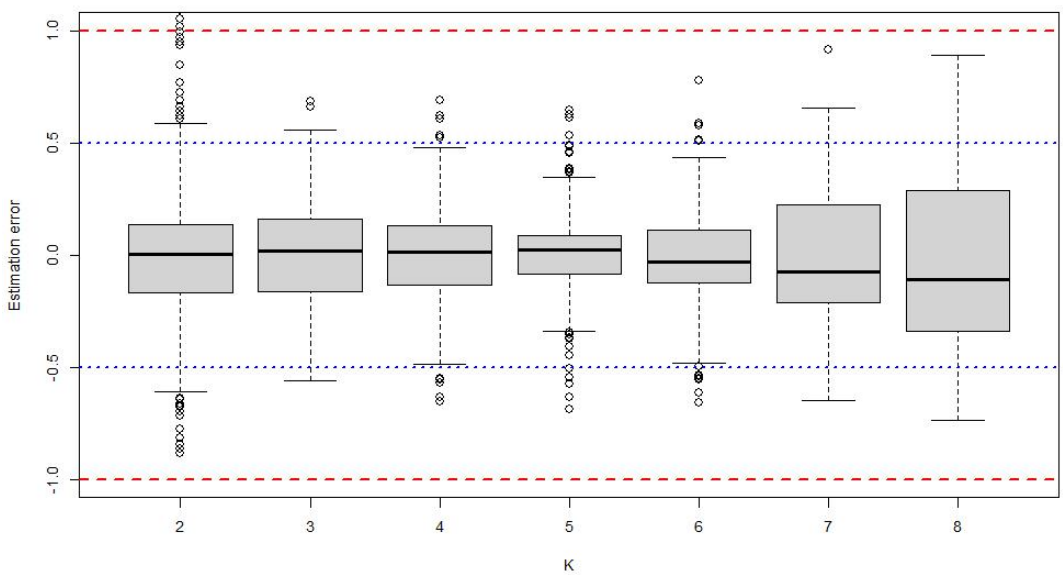


Figure 12. Estimation error as a function of parameter K

The resulting model achieved a multiple R-squared $R^2 = 0.9485$ and an adjusted R-squared $R^2 = 0.9483$, indicating that Equation (23) may produce useful predictions,

denoted by $E[N]^r$. As shown in Figure 12, the estimation error $E[N]^r - E[N]$ varied between -0.8829 and 1.1651 as a function of K. Estimation accuracy dropped off when K was extremely small or extremely large; however, under moderate conditions, the model provides meaningful guidance for managers in evaluating $E[N]$ without the need for troublesome analysis.

5. Optimization Analysis

The earlier analysis examines the steady-state probabilities and the system performance measurement. This section explores the issue of operating cost optimization. First, the cost elements per unit time are listed below

- C_h: the holding cost per customer;
- C_s: the service cost at a specific service rate;
- C_v: the vacation cost at a specific vacation rate;
- C_b: the cost incurred when a balking customer refuses join to the system.

Based on these cost elements, a function expressing the total operating cost per unit time can be formulated as:

$$TC(\mu,\phi) = C_h E[N] + C_s\mu + C_v\phi + C_b\lambda P_b . \tag{24}$$

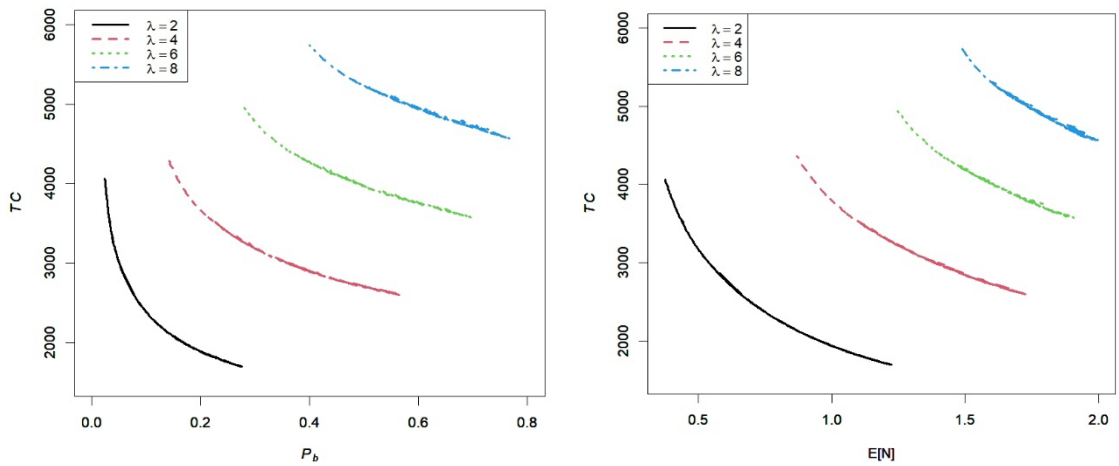
In addition to the purpose of cost minimization, the expected number of customers in the system $E[N]$ and the balking probability P_b are taken into account as the second and the third objective functions. For the developed triple-objective optimization problem, our goal is to establish the Pareto-optimal solutions formed by non-dominated solutions. A non-dominated solution is one in which no one objective function can be improved without a simultaneous detriment to at least one of the other objectives (Nayak, 2020). The mathematical model can be expressed as

$$\min_{\mu_L \le \mu \le \mu_U,\ \phi_L \le \phi \le \phi_U} \left[TC, P_b, E[N]\right], \tag{25}$$

where $[\mu_L, \mu_U]$ and $[\phi_L, \phi_U]$ indicate the search intervals. Since the decision variables, service and vacation rates, are both continuous variables, we apply the nondominated sorting genetic algorithm II (NSGA-II) for this task. NSGA-II is a popular heuristic algorithm introduced by Deb *et al.* (2002), which is an enhanced genetic algorithm specifically for multi-objective optimization problems. It is designed based on a fast nondominated sorting procedure, elitist-preserving approach, and crowding distance calculation. In this investigation, the parameters in the NSGA-II algorithm are population size 200, number of generations 50, mutation probability 0.2, and crossover rate 0.7.

5.1 Numerical Results

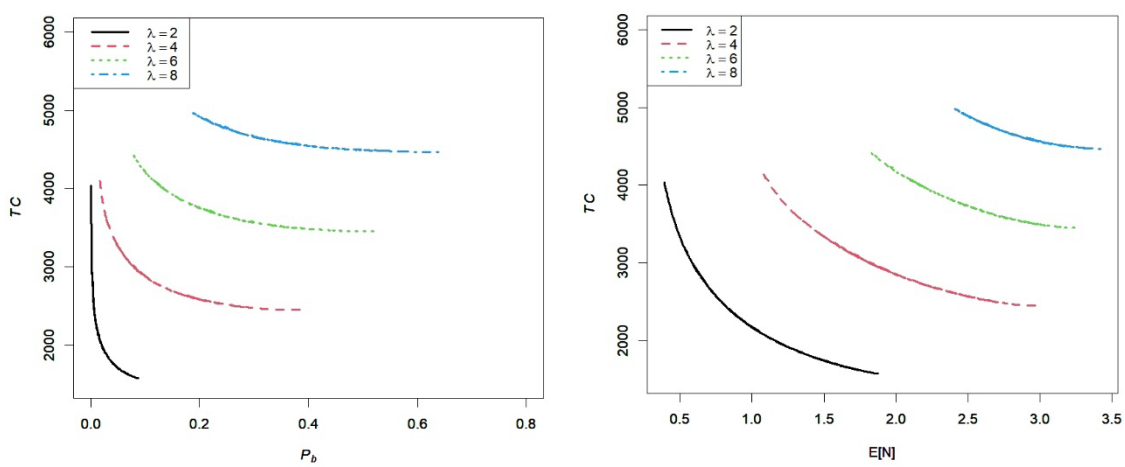
Given the cost elements $C_h = 100$, $C_s = 300$, $C_v = 200$ and $C_b = 500$, the non-dominated solutions are obtained by NSGA-II algorithm with the search intervals $\mu \in [3,10]$ and $\phi \in [2,5]$. The Pareto frontiers for $K = 3$ and $K = 6$ with different arrival rates are displayed as a function of $E[N]$ and P_b in Figures 13-14. For practical purposes, partial non-dominated solutions expressing the optimal service rate and vacation rate are tabulated in Tables 5 and 6.



(a) (b)

Figure 13. Pareto-optimal frontiers at $K = 3$ and different arrival rates

(a) TC versus P_b; (b) TC versus $E[N]$



(a) (b)

Figure 14. Pareto-optimal frontier at $K = 6$ and different arrival rates

(a) TC versus P_b; (b) TC versus $E[N]$

In the numerical results, it can be observed that

(1) The expected cost is a decreasing, but not linear, function of the expected number of customers and the balking probability. To reduce the balking probability, cost will increase significantly, particularly, when the required balking probability is very low.

(2) When the mean arrival rate rises and leads to a higher traffic intensity, the expected cost also becomes larger since the system loading is heavier. Moreover, the expected number of customers and the balking probability both increase.

(3) When the customers are more patient with a higher value of K, the balking probability decreases and the expected cost notably decreases but the expected number of customers increases. Enhancing the service rate is more beneficial than reducing the vacation period, particularly, in the improvement of the balking probability.

Table 5. Non-dominated solutions with different arrival rates and $K = 3$.

		$\lambda = 2$					$\lambda = 4$		
μ	ϕ	$E[N]$	P_b	TC	μ	ϕ	$E[N]$	P_b	TC
3.001	2.000	1.224	0.276	1698.4	3.000	2.000	1.727	0.564	2600.5
4.032	2.927	0.924	0.156	2043.7	4.033	2.477	1.570	0.457	2776.5
5.051	3.367	0.753	0.102	2366.2	5.056	3.676	1.344	0.337	3059.5
6.149	4.650	0.568	0.056	2887.5	6.034	4.320	1.195	0.264	3321.9
7.019	4.986	0.494	0.042	3193.8	7.013	4.995	1.059	0.206	3621.2
8.042	4.905	0.446	0.034	3471.6	8.005	4.896	0.990	0.179	3837.1
9.007	4.930	0.408	0.028	3757.0	9.074	5.000	0.916	0.152	4118.4
9.999	5.000	0.375	0.024	4061.0	10.00	5.000	0.869	0.137	4360.9
		$\lambda = 6$					$\lambda = 8$		
μ	ϕ	$E[N]$	P_b	TC	μ	ϕ	$E[N]$	P_b	TC
3.000	2.001	1.906	0.696	3577.9	3.000	2.000	1.999	0.767	4567.7
4.032	2.984	1.756	0.592	3759.0	4.009	3.082	1.872	0.683	4738.3
5.025	3.384	1.661	0.520	3911.8	5.010	3.443	1.807	0.624	4868.8
6.004	4.052	1.546	0.447	4107.8	6.033	4.193	1.712	0.557	5048.0
7.062	4.975	1.412	0.372	4369.7	7.002	4.937	1.622	0.497	5238.7
8.002	4.926	1.354	0.337	4532.6	8.134	4.967	1.566	0.454	5407.9
9.012	4.998	1.292	0.304	4743.0	9.052	4.990	1.525	0.425	5566.0
10.00	5.000	1.243	0.279	4960.7	10.00	5.000	1.486	0.399	5745.6

Table 6. Non-dominated solutions with different arrival rates and $K = 6$.

$\lambda = 2$					$\lambda = 4$				
μ	ϕ	$E[N]$	P_b	TC	μ	ϕ	$E[N]$	P_b	TC
3.000	2.000	1.875	0.088	1575.2	3.291	2.000	2.967	0.384	2451.4
4.023	2.327	1.356	0.032	1840.0	4.015	2.000	2.781	0.296	2474.1
5.055	3.278	0.923	0.009	2273.7	5.009	2.542	2.397	0.182	2614.6
6.118	4.435	0.656	0.003	2791.0	6.017	3.444	1.954	0.098	2885.0
7.056	4.992	0.536	0.001	3170.3	7.015	4.137	1.613	0.055	3202.7
8.163	4.875	0.473	0.001	3472.1	8.081	4.998	1.309	0.029	3612.3
9.103	4.959	0.426	0.001	3765.9	9.147	4.988	1.167	0.020	3898.7
10.00	5.000	0.393	0.001	4039.8	10.00	5.000	1.079	0.016	4140.0
$\lambda = 6$					$\lambda = 8$				
μ	ϕ	$E[N]$	P_b	TC	μ	ϕ	$E[N]$	P_b	TC
3.891	2.000	3.239	0.520	3450.4	3.898	2.000	3.420	0.638	4464.5
4.020	2.000	3.225	0.507	3450.7	4.033	2.000	3.413	0.628	4464.9
5.005	2.067	3.093	0.418	3477.8	5.011	2.453	3.281	0.540	4482.6
6.033	2.631	2.839	0.311	3553.6	6.049	2.737	3.154	0.458	4508.3
7.003	3.104	2.598	0.232	3677.3	7.005	3.241	2.992	0.378	4561.6
8.035	4.224	2.246	0.149	3926.1	8.018	3.714	2.815	0.305	4650.7
9.022	4.960	1.979	0.101	4200.6	9.004	4.181	2.634	0.244	4777.6
10.00	5.000	1.826	0.079	4419.2	10.00	5.000	2.408	0.184	4978.2

6. Conclusions

This study investigated a ticket queue with balking customers and a single vacation policy. The system states are explicitly described by one indicator variable with vectors of various orders. We present a state-transition diagram and a list of flow-balance equations in matrix form. The steady-state probabilities are solved using the matrix-geometric method and recursive technique. Suitable partitioning of the state space makes it possible to determine the state-state distribution of system size using matrix-geometric and recursive techniques. To simplify the solution process, we developed a reduced stochastic model to approximate steady-state results. The resulting probabilities can then be used to derive expressions of the expected number of balking customers and customers in the system. Sensitivity analysis was used to examine the effects of various parameters on the two system performance metrics. The approximation accuracy of the proposed reduced model was excellent, as evidenced by the absolute percentage errors. Graphs revealed the importance of evaluating customer patience and their balking threshold. It appears that system length is affected mainly by service efficiency rather than the vacation rate. We also developed a backward step-wise regression model with added variables and feature extractions to make it possible for practitioners to estimate system length without tedious matrix manipulations. A cost function was formulated based on the system performance metrics. The efficient

NSGA-II was applied to solve a triple-objective optimization problem considering the expected operating cost, the expected number of customers, and the balking probability. Numerical results with graphical illustrations were provided for use by managers. Future research could consider reneging, a hybrid vacation policy, and/or cases involving unreliable servers.

References

[1] Baccelli, F., Boyer, P., & Hebuterne, G. (1984). Single-server queues with impatient customers. *Advances in Applied Probability*, 16(4), 887-905.

[2] Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. A. M. T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2), 182-197.

[3] Ding, D., Ou, J., & Ang, J. (2015). Analysis of ticket queues with reneging customers. *Journal of the Operational Research Society*, 66(2), 231-246.

[4] Hanukov, G., Anily, S., & Yechiali, U. (2020). Ticket queues with regular and strategic customers. *Queueing Systems*, 95(1-2), 145-171.

[5] Hanukov, G., Hassoun, M., & Musicant, O. (2021). On the benefits of providing timely information in ticket queues with balking and calling times. *Mathematics*, 9(21), 2753.

[6] Jain, M., Dhibar, S., & Sanga, S. S. (2022). Markovian working vacation queue with imperfect service, balking and retrial. *Journal of Ambient Intelligence and Humanized Computing*, 1-17.

[7] Jennings, O. B., & Pender, J. (2016). Comparisons of ticket and standard queues. *Queueing Systems*, 84, 145-202.

[8] Ke, J. C., Liu, T. H., Su, S., & Zhang, Z. G. (2022). On retrial queue with customer balking and feedback subject to server breakdowns. *Communications in Statistics-Theory and Methods*, 51(17), 6049-6063.

[9] Ke, J. C., Wu C. H., & Zhang Z. G. (2010). Recent developments in vacation queueing models: A short survey. *International Journal of Operations Research* 7, 3-8.

[10] Kerner, Y., Sherzer, E., & Yanco, M. A. (2017). On non-equilibria threshold strategies in ticket queues. *Queueing Systems*, 86(3), 419-431.

[11] Kumar, A., & Jain, M. (2023). M/M/1 queue with bi-level network process and bi-level vacation policy with balking. *Communications in Statistics-Theory and Methods*, 52(15), 5502-5526.

[12] Kuzu, K. (2010). Analytical and empirical investigations of ticket queues: Implications for system performance, customer perceptions, and behavior. Ph.D. Thesis, The Pennsylvania State University.

[13] Kuzu, K. (2015). Comparisons of perceptions and behavior in ticket queues and physical queues. *Service Science*, 7(4), 294-314.

[14] Kuzu, K., Gao, L., & Xu, S. H. (2019). To wait or not to wait: The theory and practice of ticket queues. *Manufacturing and Service Operations Management*, 21(4), 853-874.

[15] Kuzu, K., & Soyer, R. (2018). Bayesian modeling of abandonments in ticket queues. *Naval Research Logistics*, 65(6-7), 499-521.

[16] Li, K., & Wang, J. (2021). Equilibrium balking strategies in the single-server retrial queue with constant retrial rate and catastrophes. *Quality Technology and Quantitative Management*, 18(2), 156-178.

[17] Mandelbaum, A., & Shimkin, N. (2000). A model for rational abandonments from invisible queues. *Queueing Systems*, 36(1), 141-173.

[18] Nayak, S. 2020. Fundamentals of optimization techniques with algorithms. Academic Press.

[19] Neuts, M. 1981. *Matrix Geometric Solution in Stochastic Models*. Johns Hopkins University Press, Baltimore, MD.

[20] Poomrittigul, S., Koomsubsiri, A., Aung, H. L., Sasithong, P., & Wuttisittikulkij, L. (2020, January). Ticket machine queuing system design application for service efficiency simulation and comparison. In *2020 International Conference on Electronics, Information, and Communication (ICEIC)* (pp. 1-4). IEEE.

[21] Sun, W., Zhang, Z., & Li, S. (2023). Comparisons of customer balking behavior in observable queues with *N* policies and geometric abandonments. *Quality Technology and Quantitative Management*, 20(3), 307-333.

[22] Tian, R. & Wang, Y. (2020). Optimal strategies and pricing analysis in M/M/1 queues with a single working vacation and multiple vacations. *RAIRO-Operations Research*, 54(6), 1593-1612.

[23] Tian, R., Zhang, Z. G., & Su, S. (2022). On Markovian queues with single working vacation and Bernoulli interruptions. *Probability in the Engineering and Informational Sciences*, 36(3), 616-643.

[24] Upadhyaya, S. (2016). Queueing systems with vacation: An overview. *International Journal of Mathematics in Operational Research*, 9, 167-213.

[25] Xiao, L., Xu, S. H., Yao, D. D., & Zhang, H. (2022). Optimal staffing for ticket queues. *Queueing Systems*, 102(1-2), 309-351.

[26] Xu, S. X., Gao L., Ou J. (2007). Service performance analysis and improvement for a ticket queue with balking customers. *Management Science.* 53, 971-990.