ANALYTICAL AND COMPUTATIONAL ASPECTS OF A MULTI-SERVER QUEUE WITH IMPATIENCE UNDER DIFFERENTIATED WORKING VACATIONS POLICY

[1]Aimen Dehimi, [2]Mohamed Boualem, [3]Amina Angelika Bouchentouf, [4]Sofiane Ziani, [5]Louiza Berdjoudj

•

[1]University of Bejaia, Faculty of Exact Sciences, Applied Mathematics Laboratory, 06000 Bejaia, Algeria
[2]University of Bejaia, Faculty of Technology, Research Unit LaMOS, 06000 Bejaia, Algeria
[3]Laboratory of Mathematics, Djillali Liabes University of Sidi Bel Abbes, 22000 Sidi Bel Abbes, Algeria
[4]University of Bejaia, Research Unit LaMOS, 06000 Bejaia, Algeria
[5]University of Bejaia, Faculty of Exact Sciences, Research Unit LaMOS, 06000 Bejaia, Algeria
[1]aimen.dehimi@univ-bejaia.dz, [2]mohammed.boualem@univ-bejaia.dz, [3]bouchentouf_amina@yahoo.fr
[4]sofiane.ziani@univ-bejaia.dz, [5]louiza.berdjoudj@univ-bejaia.dz

Abstract

A multi-server queueing system with synchronous differentiated working vacation policy, Bernoulli schedule vacation interruption, and customer impatience (balking and reneging) is studied. The system consists of c servers and a finite capacity N, where customers arrive according to a Poisson process and are served in the chronological order of their arrival. When the system becomes empty, servers wait for a random duration before entering a type-1 working vacation, during which service is provided at a reduced rate. If customers are present in the system at the moment of service achievement during this period, the vacation is interrupted. With a certain probability, servers return to the regular busy period; otherwise, they continue the working vacation. Upon completion of the working vacation, if the system is still empty, servers can take another working vacation of shorter duration, named type-2 working vacation; otherwise, they switch to the regular busy period. Customer impatience is considered during both the normal busy period and working vacations. A recursive analysis method is used to find the steady-state probabilities of the system. Then, some important performance measures are obtained. Furthermore, an optimal operational policy for the model is developed to minimize the total expected cost. The Grey Wolf Optimization (GWO) meta-heuristic approach is employed to determine the optimal service rates for both working vacations and normal busy periods. Finally, several numerical examples are provided to validate and support the theoretical findings.

Keywords: Multi-server queue, differentiated vacations, impatience, GWO algorithm, cost optimization.

1. Introduction

Queueing models have gained considerable attention due to their significance in shaping and evaluating telecommunication systems, computer systems, and production management [8, 22, 23].

The concept of a server vacation queue has garnered extensive research attention, primarily due to its unique characteristic of allowing the server to utilize idle time for various tasks, such as maintenance, service industries, production and manufacturing systems, or just taking a break [7]. For example, the growing utilization of wireless cellular networks has led to a substantial surge in energy consumption. To address this issue and promote the development of energy-efficient wireless cellular networks, researchers have introduced the concept of hibernation or sleeping for base stations (BS) during periods of inactivity. This approach is akin to the concept of a server going on vacation, where the BS temporarily reduces its power consumption when there are no active users in the network. In the classical server vacation queue, a server temporarily ceases its service during a designated vacation period [9, 10, 21]. It is important to note that some systems are designed with the presence of an alternate server that operates at a different, often lower, service rate when the primary server takes a vacation. Such a system is commonly referred to as a working vacation queue. In most working vacation policies, the server typically returns to its regular service rate after the vacation period ends, but only if there are customers waiting in the system. The idea of a working vacation was initially introduced by [24], where they proposed that the server does not entirely cease its operations during a vacation but continues providing service to the queueing system at a reduced rate. This concept has paved the way for various working vacation policies, enhancing the flexibility and efficiency of queueing system designs. These models have been discussed by different authors [3, 4, 12, 16, 26].

In numerous practical scenarios involving congestion, there are occasions when urgent events take place during vacation. As a result, the servers must interrupt their vacation and resume work instead of utilizing the remaining vacation time. Otherwise, such a situation incurs a substantial cost in terms of waiting customers. The concept of vacation interruption was initially introduced by [14]. Subsequently, [15] conducted a study on a $GI/M/1$ queue utilizing a supplementary variable method. Further discussions on an $M/PH/1$ queue, considering working vacations and vacation interruption, can be found in [2]. For more in-depth studies, additional references include [11, 13, 17, 18], and references therein.

In the past two decades, there has been a significant focus on the subject of impatient customers within queueing theory. This research area has proven to be intriguing and challenging, particularly in the context of globalization, hospital emergency rooms handling critical patients, and other relevant domains. As a result, the topic of queueing models with server vacations and impatient customers has garnered significant attention in the literature [1, 5, 8, 20, 25].

The main aim of this work is to conduct an analytical and optimization analysis of a finite capacity queue with multi-server and impatient customers (balking and reneging behaviors), incorporating vacation interruption and differentiated working vacations. The suggested queueing model presents promising applications across diverse sectors, including call centers, telecommunications and manufacturing, where servers can experience periods of downtime. In this research, the steady-state probabilities of queue length when servers are in working vacations period (type-1 and type-2), and in normal busy period are investigated using the recursive analysis approach. Several important performance measures are derived from these probabilities. Optimization in queueing systems is crucial in practical applications. In this study, the optimization problem tackled is complex and challenging, as the objective function is nonlinear on the service rates. To address this issue, the GWO algorithm is employed to determine the optimal service rates for both working vacations and normal busy periods, aiming to minimize the expected total cost. The GWO algorithm is known for its high performance in both unconstrained and constrained problems [19]. It has shown competitive results compared to well-established heuristics in swarm intelligence. Notably, the application of the GWO algorithm in queueing theory is relatively scarce in the existing literature. The present work can be considered as an extension to the research in reported [6], where the steady-state distributions were investigated in the case of a single server. By applying the GWO algorithm and considering the multi-server case, this paper contributes to the understanding and analysis of the considered model. Finally, numerical examples are presented to evaluate the behavior and performance of the proposed queueing system. These numerical results provide insights and support our findings.

The structure of the paper is as follows: Section 2 provides a detailed description of the queueing model being studied. In Section 3, we derive the steady-state distributions of queue sizes during different server periods, including working vacation and normal busy periods. Section 4 gives explicit formulas for different performance measures of the queueing model. Moving on to Section 5, we analyze the effect of various system parameters on the performance measures through graphics. In Section 6, we address an optimization problem related to service rates using the GWO algorithm and present numerical results. Finally, section 7 gives a general conclusion and perspectives.

2. Overview and analysis of the proposed framework

We investigate an $M/M/c/K$ queue with impatience, operating under the differentiated working vacations along with vacation interruption. The fundamental assumptions underpinning this queueing system are outlined as:

- Customers enter the system in line with a Poisson process characterized by a rate of α.
- The time during a normal busy period of each server follows an exponential distribution and is denoted by service rate μ_1.
- Customers are served in accordance with FCFS (First-Come-First-Served) discipline and the capacity of the system is considered to be finite, say K.
- The time during working vacations of each server follows an exponential distribution and is denoted by service rate μ_2 ($\mu_2 < \mu_1$).
- The queueing system under consideration involves multiple servers, denoted by c, when the system has no customers the servers wait for a random duration of time before leaving collectively for type-1 working vacation. Subsequently , When the servers return from their working vacation and find the system non-empty, they change their service rate from μ_2 to μ_1 and a normal busy period starts. If the servers return to find an empty queue, they immediately leave for another working vacation.
- The waiting time for the servers follows an exponential distribution with rate Δ.
- Following the completion of the waiting time duration, they begin an initial type-1 working vacation exponentially distributed with parameter Φ_1. Once they return from the initial type-1 working vacation, if there are no customers in the queue, they transition to type-2 vacation which follows an exponential distributions characterized by parameter Φ_2. Otherwise, they return to the normal busy period and start serving customers in the queue.
- Upon a customer's arrival during the vacation period, within this phase. Upon completing a service, if there are customers in the queue, the servers follow the Bernoulli distribution. They may opt to interrupt the vacation and move to the normal busy period, a choice determined by the probability denoted as β'. Alternatively, the servers may choose to continue the vacation, a decision made with the complementary probability $\beta = 1 - \beta'$. It is crucial to note that the vacation service rate exclusively applies to the first arriving customer during the working vacation period.
- Upon customer arrival, a decision is made based on the following probabilities: The customer opts to either join the queue with a probability denoted as ψ_k or decide not to and balk, with the complementary probability expressed as $\psi_k' = 1 - \psi_k$. This decision-making process occurs when there are already k customers ahead in the queue, where $c \leq k \leq K$. It is important to note that the probabilities ψ_k satisfy the conditions $0 \leq \psi_{k+1} \leq \psi_k \leq 1$, $c \leq k \leq K-1$, $\psi_0 = 1, ..., \psi_{c-1} = 1$, and $\psi_K = 0$.
- During the normal busy period or either type-1 or type-2 working vacations, customers are governed by impatience timers: T_0, T_1, or T_2, respectively. These timers follow exponential distributions with parameters ξ_0, ξ_1, and ξ_2. In practical terms, if a customer's service doesn't commence before the timer expires, they will abandon the queue (renege), and their return is not anticipated.

The variables introduced are mutually independent.

2.1. Real-world implementation of the model

The considered queueing system finds practical application in technical software product support centers. Customers seeking assistance with technical issues contact the support center, arriving randomly over time according to a Poisson process (α). During regular operating hours, support agents attend to customers, with service times following an exponential distribution at rate μ_1. when there is no call in the system, the support agents are allowed to remain in an inactive state for a random period (waiting time). After that, support agents enter type-1 working vacation, where service capacity decreases to μ_2. Upon return from a working vacation, if there are non-calls, agents transition to a type-2 working vacation. At the time during both type-1 and type-2 working vacation modes, if some calls are present in the system, the support agents can continue operating with probability β or they will switch to the normal busy period with probability $\beta^{'} = 1 - \beta$ and be processed immediately (working vacation interruption). Calls decide whether to join the queue with probability ψ_k and $1 - \psi_k$ denotes the probability that they decide to balk when there are $n \geq c$ incoming calls in front of them in the system. Additionally, during various operational phases, customers are subject to impatience timers (T_0, T_1, T_2), abandoning the queue if service doesn't commence before timer expiration (reneging).

3. Examination of the probabilities in a steady-state

We consider the bi-variate process $(S(t),\ L(t))_{(t\geq 0)}$, where $L(t)$ is the number of customers in the system at time t, and $S(t)$ defines the state of the servers at time t and takes one of three values, such as $S(t) = 0$: when the servers are in normal busy period at time t, and $S(t) = 1$ (resp. $S(t) = 2$): when the servers are in type-1 (resp. in type-2) working vacation period at time t.

The joint probability $P_{j,k} = \lim_{t\to\infty} P\{S(t) = j,\ L(t) = k,\ (j,k) \in \Omega\}$, denote the steady-state probabilities of the system. Figure 1 shows the transition diagram of the considered model. Next, to avoid overloading mathematical expressions, the following notations are used:

$$\varsigma_k = \begin{cases} 0, & k = 0, 1, \\ k\beta^{'}\mu_2, & 2 \leq k \leq c-1, \\ c\beta^{'}\mu_2, & k \geq c, \end{cases} \quad \varphi_{0,k} = \begin{cases} \mu_1, & k = 1, \\ k\mu_1 + (k-1)\xi_0, & 2 \leq k \leq c, \\ c\mu_1 + (k-1)\xi_0, & k \geq c+1, \end{cases}$$

$$\zeta_{j,k} = \begin{cases} \mu_2, & k = 1, \\ k\beta\mu_{2j} + (k-1)\xi_j, & 2 \leq k \leq c-1, \\ c\beta\mu_{2j} + (k-1)\xi_j, & k \geq c. \end{cases}$$

Using the principle of balance equations

$$(\alpha + \Delta)P_{0,0} = \mu_1 P_{0,1},\ \ k = 0, \tag{1}$$

$$\begin{aligned} &(\alpha + k\mu_1 + (k-1)\xi_0)P_{0,k} = \alpha P_{0,k-1} + ((k+1)\mu_1 + k\xi_0)P_{0,k+1} + \Phi_1 P_{1,k} \\ &+(k+1)\beta'\mu_2 P_{1,k+1} + \Phi_2 P_{2,k},\ \ 1 \leq k \leq c-1, \end{aligned} \tag{2}$$

$$\begin{aligned} (\alpha\psi_c + c\mu_1 + (c-1)\xi_0)P_{0,c} = (c\mu_1 + c\xi_0)P_{0,c+1} + \alpha P_{0,c-1} + \Phi_1 P_{1,c} \\ +c\beta'\mu_2 P_{1,c+1} + \Phi_2 P_{2,c} + c\beta'\mu_2 P_{2,c+1}, \end{aligned} \tag{3}$$

$$\begin{aligned} &(\alpha\psi_k + c\mu_1 + (k-1)\xi_0)P_{0,k} = \alpha\psi_{k-1}P_{0,k-1} + (c\mu_1 + k\xi_0)P_{0,k+1} + \Phi_1 P_{1,k} \\ &+c\beta'\mu_2 P_{1,k+1} + \Phi_2 P_{2,k} + c\beta'\mu_2 P_{2,k+1},\ \ c+1 \leq k \leq K-1, \end{aligned} \tag{4}$$

$$(c\mu_1 + (K-1)\xi_0)P_{0,K} = \alpha\psi_{K-1}P_{0,K-1} + \Phi_1 P_{1,K} + \Phi_2 P_{2,K}, \tag{5}$$

$$(\alpha + \Phi_1)P_{1,0} = \Delta P_{0,0} + \mu_2 P_{1,1},\ \ k = 0, \tag{6}$$

$$(\alpha + k\mu_2 + (k-1)\xi_1 + \Phi_1)P_{1,k} = \alpha P_{1,k-1} + ((k+1)\beta\mu_2 + k\xi_1)P_{1,k+1},\ \ 1 \leq k \leq c-1, \tag{7}$$

$$(\alpha\psi_c + c\mu_2 + (c-1)\xi_1 + \Phi_1)P_{1,c} = \alpha P_{1,c-1} + (c\beta\mu_2 + c\xi_1)P_{1,c+1}, \tag{8}$$

$$(\alpha\psi_k + c\mu_2 + (k-1)\xi_1 + \Phi_1)P_{1,k} = \alpha\psi_{k-1}P_{1,k-1} + (c\beta\mu_2 + k\xi_1)P_{1,k+1}, c+1 \leq k \leq K-1, \tag{9}$$

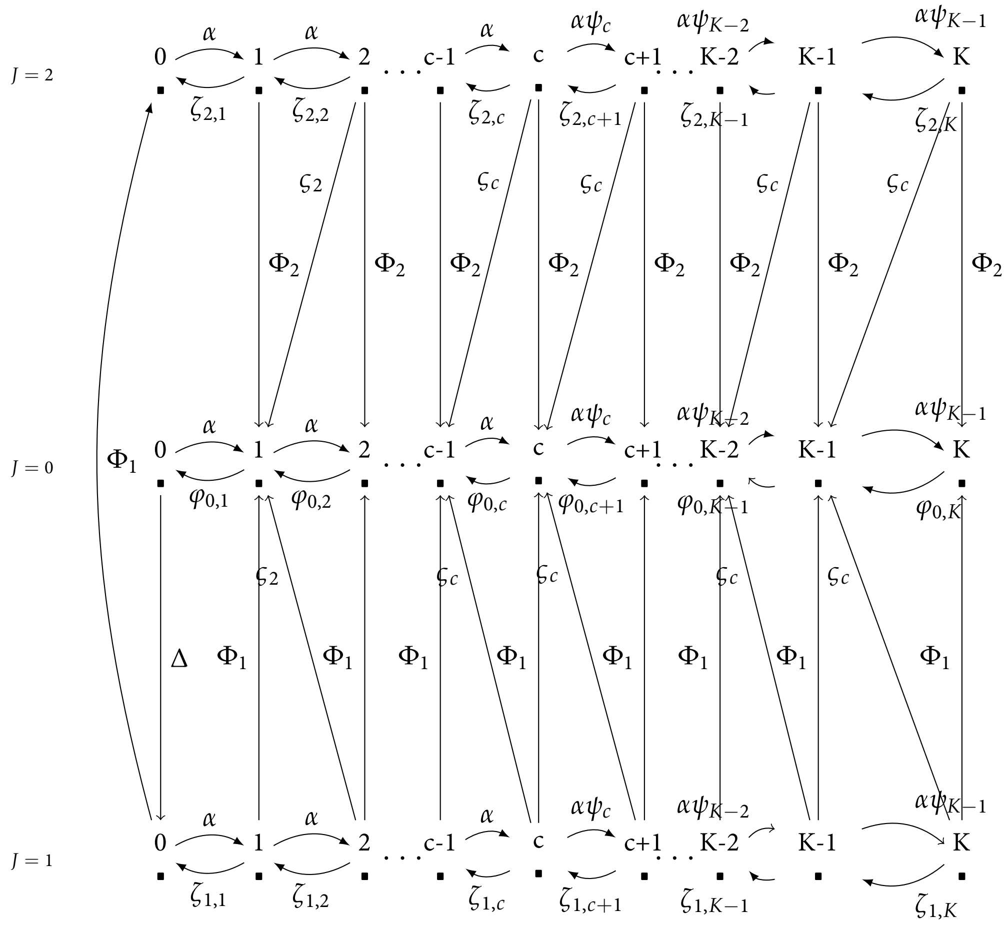


Figure 1: *State transition rate diagram*

$$(c\mu_2 + (K-1)\xi_1) + \Phi_1)P_{1,K} = \alpha\psi_{K-1}P_{1,K-1}, \tag{10}$$

$$\alpha P_{2,0} = \Phi_1 P_{1,0} + \mu_2 P_{2,1}, \;\; k = 0, \tag{11}$$

$$(\alpha + k\mu_2 + (k-1)\xi_2 + \Phi_2)P_{2,k} = \lambda P_{2,k-1} + ((k+1)\beta\mu_2 + k\xi_2)P_{2,k+1}, \;\; 1 \leq k \leq c-1, \tag{12}$$

$$(\alpha\psi_c + c\mu_2 + (c-1)\xi_2) + \Phi_2)P_{2,c} = \alpha P_{2,c-1} + (c\beta\mu_2 + c\xi_2)P_{2,c+1}, \tag{13}$$

$$(\alpha\psi_k + c\mu_2 + (k-1)\xi_2) + \Phi_2)P_{2,k} = \alpha\psi_{k-1}P_{2,k-1} + (c\beta\mu_2 + k\xi_2)P_{2,k+1}, c+1 \leq k \leq K-1, \tag{14}$$

$$(c\mu_2 + (K-1)\xi_2 + \Phi_2)P_{2,K} = \alpha\psi_{K-1}P_{2,K-1}, \tag{15}$$

The normalizing condition is

$$\sum_{k=0}^{K} (P_{0,k} + P_{1,k} + P_{2,k}) = 1. \tag{16}$$

Now, we present the solution of the equations above in the following theorem.

Theorem 1. The probabilities describing the system size in different operational periods, namely the type-2 working vacation period ($P_{2,k}$), type-1 working vacation period ($P_{1,k}$), and normal busy period ($P_{0,k}$), in the steady-state are respectively expressed as follows:

$$P_{2,k} = \theta_k P_{2,K} = \theta_k \left(\sum_{k=0}^{K} (\theta_k + \Theta_1\delta_k + \Theta_2\omega_k - \Gamma_k) \right)^{-1}, \;\; k = 0, 1, 2, ..., K. \tag{17}$$

$$P_{1,k} = \Theta_1 \delta_k P_{2,K}. \tag{18}$$

$$P_{0,k} = (\Theta_2 \omega_k - \Gamma_k) P_{2,K}, \tag{19}$$

wher e

$$\theta_k = \begin{cases} 1, & k = K, \\ \frac{c\mu_2+(K-1)\xi_2+\Phi_2}{\alpha\psi_{K-1}}, & k = K-1, \\ \frac{\alpha\psi_{k+1}+c\mu_2+\Phi_2+k\xi_2}{\alpha\psi_k}\theta_{k+1} - \frac{(c\beta\mu_2+(k+1)\xi_2)}{\alpha\psi_k}\theta_{k+2}, & c \le k < K-1, \\ \frac{\alpha\psi_{k+1}+(k+1)\mu_2+\Phi_2+k\xi_2}{\alpha}\theta_{k+1} - \frac{((k+1)\beta\mu_2+(k+1)\xi_2)}{\alpha}\theta_{k+2}, & k = c-1, \\ \frac{\alpha+(k+1)\mu_2+\Phi_2+k\xi_2}{\alpha}\theta_{k+1} - \frac{((k+2)\beta\mu_2+(k+1)\xi_2)}{\alpha}\theta_{k+2}, & 0 \le k \le c-2, \end{cases} \tag{20}$$

$$\delta_k = \begin{cases} 1, & k = K, \\ \frac{c\mu_2+(K-1)\xi_1+\Phi_1}{\alpha\psi_{K-1}}, & k = k-1, \\ \frac{\alpha\psi_{k+1}+c\mu_2+\Phi_1+k\xi_1}{\alpha\psi_k}\delta_{k+1} - \frac{(c\beta\mu_2+(k+1)\xi_1)}{\alpha\psi_k}\delta_{K+2}, & c \le k < K-1, \\ \frac{\alpha\psi_{k+1}+(k+1)\mu_2+\Phi_1+k\xi_1}{\alpha}\delta_{k+1} - \frac{((k+1)\beta\mu_2+(k+1)\xi_1)}{\alpha}\delta_{k+2}, & k = c-1, \\ \frac{\alpha+(k+1)\mu_2+k\xi_1+\Phi_1}{\alpha}\delta_{k+1} - \frac{((k+2)\beta\mu_2+(k+1)\xi_1)}{\alpha}\delta_{k+2}, & 0 \le k \le c-2, \end{cases} \tag{21}$$

$$\Theta_1 = \frac{\alpha\theta_0 - \mu_2\theta_1}{\Phi_1\delta_0}. \tag{22}$$

$$\omega_k = \begin{cases} 1, & k = K, \\ \frac{c\mu_1+(K-1)\xi_0}{\alpha\psi_{K-1}}, & k = K-1, \\ \frac{\alpha\psi_{k+1}+c\mu_1+k\xi_0}{\alpha\psi_k}\omega_{k+1} - \frac{(c\mu_1+(k+1)\xi_0)}{\alpha\psi_k}\omega_{k+2}, & c \le k < K-1, \\ \frac{\alpha\psi_{k+1}+(k+1)\mu_1+k\xi_0}{\alpha}\omega_{k+1} - \frac{((k+1)\mu_1+(k+1)\xi_0)}{\alpha}\omega_{k+2}, & k = c-1, \\ \frac{\alpha+(k+1)\mu_1+k\xi_0}{\alpha}\omega_{k+1} - \frac{((k+2)\mu_1+(k+1)\xi_0)}{\alpha}\omega_{k+2}, & 0 \le k \le c-2, \end{cases}$$

$$\Gamma_k = \begin{cases} 0, & k = K, \\ \frac{\Phi_1\Theta_1+\Phi_2}{\alpha\psi_{K-1}}, & k = K-1, \\ \frac{\Theta_1(\Phi_1\delta_{k+1}+c\beta'\mu_1\delta_{k+2})+(\Phi_2\theta_{k+1}+c\beta'\mu_2\theta_{k+2})}{\alpha\psi_k}, & c \le k < K-1, \\ \frac{\Theta_1(\Phi_1\delta_{k+1}+(k+1)\beta'\mu_1\delta_{k+2})+(\Phi_2\theta_{k+1}+(k+1)\beta'\mu_2\theta_{k+2})}{\alpha}, & k = c-1, \\ \frac{\Theta_1(\Phi_1\delta_{k+1}+(k+2)\beta'\mu_2\delta_{k+2})+(\Phi_2\theta_{k+1}+(k+2)\beta'\mu_2\theta_{k+2})}{\alpha}, & 0 \le k \le c-2, \end{cases}$$

$$\Theta_2 = \frac{\Theta_1(\alpha+\Phi_1)\delta_0 - \Theta_1\mu_1\delta_1 + \Delta\Gamma_0}{\Delta\omega_0}, \tag{23}$$

and

$$P_{2,K} = \left(\sum_{k=0}^{K} (\theta_k + \Theta_1 \delta_k + \Theta_2 \omega_k - \Gamma_k) \right)^{-1}. \tag{24}$$

Proof. The stationary probabilities, denoted as $P_{2,k}$, $P_{1,k}$, and $P_{0,k}$, are determined using equations (1) – (15), expressed in terms of $P_{2,K}$. To calculate $P_{2,k}$, we use a recursive approach to solve equations (12) – (15). This leads us to derive expressions (17) and (20).

For $P_{1,k}$, we find it to be equal to $\Theta_1 \delta_k P_{1,K}$, with δ_k defined by (21). Utilizing equation (11), we obtain equations (18) and (22).

By solving equations (2) – (5), we can express $P_{0,k}$ in terms of both $P_{0,K}$ and $P_{2,K}$. Further, with the assistance of equation (6), we deduce $P_{0,k}$ as a function of $P_{2,K}$, as given in (19).

Finally, we ensure that these probabilities satisfy the normalization condition (see equation (16)), which leads us to equation (24). ■

4. Metrics of system performance

▷ The probabilities associated with different server states–normal busy period, type-1 working vacation, and type-2 working vacation–are defined as follows:

$$P_{bn} = P_{2,K} \sum_{k=0}^{K} (\Theta_2 \omega_k - \Gamma_k), \qquad P_{wv1} = \Theta_1 P_{2,K} \sum_{k=0}^{K} \delta_k, \qquad P_{wv2} = P_{2,K} \sum_{k=0}^{K} \theta_k.$$

▷ The probabilities of the servers being idle during the busy period (P_{id}) and actively working during the normal busy period (P_{wn}) are expressed as follows:

$$P_{id} = (\Theta_2 \omega_0 - \Gamma_0) P_{2,K}.$$

$$P_{wn} = 1 - \left[P_{2,K} \left((\Theta_2 \omega_0 - \Gamma_0) + \Theta_1 \sum_{k=0}^{K} \delta_k + \sum_{k=0}^{K} \theta_k \right) \right]. \tag{25}$$

▷ The expressions for the expected number of customers in the system (L_s) and in the queue (L_q) are defined as follows:

$$L_s = P_{2,K} \left[\sum_{k=0}^{K} (\Theta_2 k \omega_k - k\Gamma_k + \Theta_1 k \delta_k + k\theta_k) \right]. \tag{26}$$

$$L_q = P_{2,K} \left[\sum_{k=c}^{K} (\Theta_2 (k-c) \omega_k - (k-c)\Gamma_k + \Theta_1 (k-c) \delta_k + (k-c)\theta_k) \right]. \tag{27}$$

▷ The expression for E_{cs} (expected number of customers served per time unit) is given by:

$$\begin{aligned} E_{cs} = P_{2,K} & \left[\mu_1 \Theta_2 \sum_{k=1}^{c-1} k\omega_k - \mu_1 \sum_{k=1}^{c-1} k\Gamma_k + c\mu_1 \Theta_2 \sum_{n=c}^{K} \omega_k - c\mu_1 \sum_{k=c}^{K} \Gamma_k \right] \\ + P_{2,K} & \left[\mu_2 \Theta_1 \sum_{k=1}^{c-1} k\delta_k + \mu_2 \sum_{k=1}^{c-1} k\theta_k + c\mu_2 \Theta_1 \sum_{k=c}^{K} \delta_k + c\mu_2 \sum_{k=c}^{K} \theta_k \right]. \end{aligned} \tag{28}$$

▷ The expressions for the expected waiting time of customers in the system (W_s) and in the queue (W_q) are given by:

$$W_s = \frac{L_s}{\lambda'}, \text{ and } W_q = \frac{L_q}{\lambda'}, \quad \text{where } \lambda' = \lambda - B_r. \tag{29}$$

▷ The expected reneging rate:

$$R_r = P_{2,K} \left[\sum_{k=1}^{K} (\xi_0 \Theta_2 (k-1) \omega_k - \xi_0 (k-1) \Gamma_k + \xi_1 \Theta_1 (k-1) \delta_k) + \xi_2 \sum_{k=1}^{K} (k-1) \theta_k \right]. \tag{30}$$

▷ The expected balking rate:

$$B_r = \alpha P_{2,K} \left[\sum_{k=c}^{K} (\Theta_2 \psi_k' \omega_k - \psi_k' \Gamma_k + \Theta_1 \psi_k' \delta_k + \psi_k' \theta_k) \right]. \tag{31}$$

5. Numerical results

This section presents various numerical examples to illustrate the influence of different parameters, including α,Φ_1, ξ_0, Φ_2, c, ξ_1, K, ξ_2, on the performance metrics of the queueing model (P_{wv1}, P_{wv2}, P_{wn}, P_{id}, B_r, R_r, E_{cs}, L_s, L_q, W_s, λ'). To do this, we use the probability of non-balking defined as: $\psi_k = 1 - \frac{k}{K}$.

- Scenario 1: We fix $\alpha = 0.01 : .01 : 5$, $\beta' = 0.3$, $\Phi_1 = 1.15$, $\Phi_2 = 1.8$, $\xi_0 = 0.7$, $\xi_1 = 1.1$, $\xi_2 = 1.5$. We consider the following cases:

– Case 1: $\mu_1 = 2.5$, $\mu_2 = 1$, $\Delta = 0.3$, $c = [1; 2; 3; 4]$, $K = 10$.

– Case 2: $\mu_1 = 2.5$, $\mu_2 = 1$, $\Delta = 0.3$, $c = 3$, $K = [10; 15; 20; 25]$.

- Scenario 2: We fix $\Phi_1 = 0.01 : .01 : 2.5$, $\alpha = 2$, $\mu_1 = 2.5$, $\mu_2 = 1$, $\Delta = 0.3$, $\Phi_2 = 1.8$, $c = 3$, $K = 10$. We study the following cases :

– Case 1: $\xi_0 = [0.6; 0.9; 1.2; 1.5]$, $\xi_1 = 1.1$, $\xi_2 = 1.5$, $\beta' = 0.3$.

– Case 2: $\xi_0 = 0.7$, $\xi_1 = [0.8; 1.1; 1.4; 1.7]$, $\xi_2 = 1.5$, $\beta' = 0.3$.

– Case 3: $\xi_0 = 0.7$, $\xi_1 = 1.1$, $\xi_2 = [1.5; 1.8; 2.1; 2.4]$, $\beta' = 0.3$.

- Scenario 3: We fix $\Phi_2 = 0.01 : .01 : 3$, $\alpha = 2$, $\mu_1 = 2.5$, $\mu_2 = 1$, $\Delta = 0.3$, $\Phi = 1.15$, $c = 3$, $K = 10$. We study the following cases :

– Case 1: $\xi_0 = [0.6; 0.9; 1.2; 1.5]$, $\xi_1 = 1.1$, $\xi_2 = 1.5$, $\beta' = 0.3$.

– Case 2: $\xi_0 = 0.7$, $\xi_1 = [0.8; 1.1; 1.4; 1.7]$, $\xi_2 = 1.5$, $\beta' = 0.3$.

– Case 3: $\xi_0 = 0.7$, $\xi_1 = 1.1$, $\xi_2 = [1.5; 1.8; 2.1; 2.4]$, $\beta' = 0.3$.

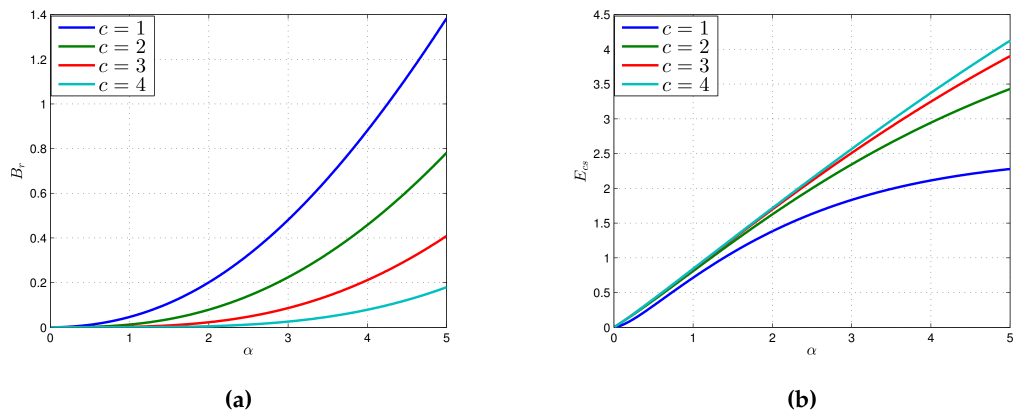
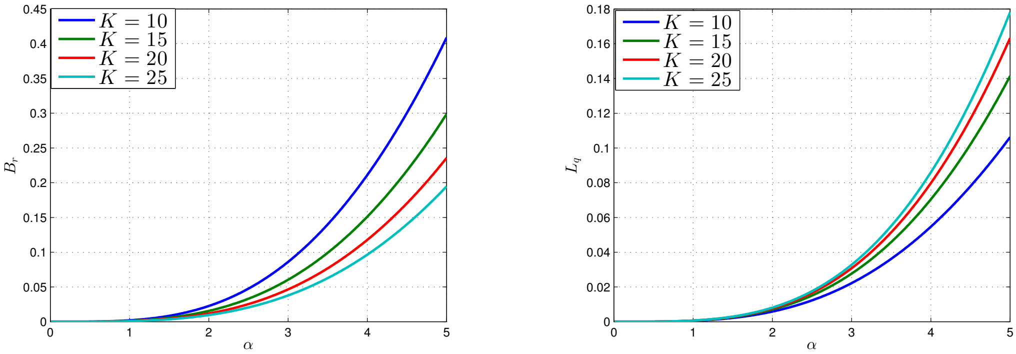


Figure 2: *B_r and E_{cs} vs. α for different values of c*

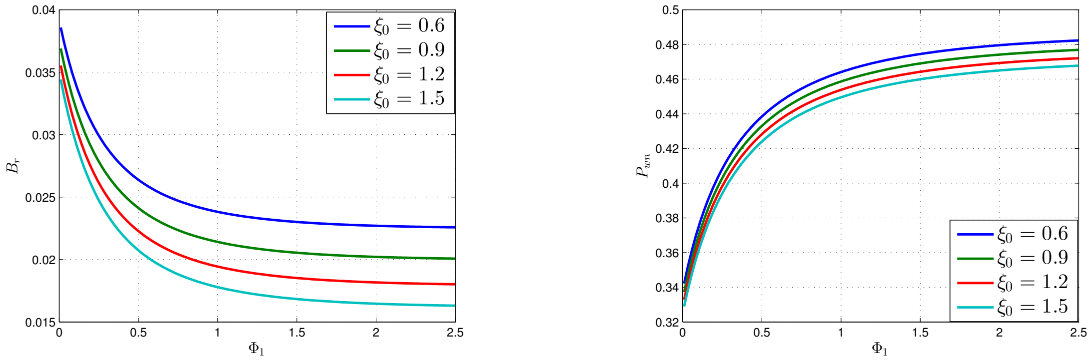
Discussion of Results

▷ Effect of α (arrival rate): Along with the increasing value of α, several factors are significantly affected. The system size increases, leading to an augmentation in the probability of working during the normal busy period P_{wn}. Additionally , the average balking B_r (see Figures 2a and 3a), mean number of served customer E_{cs} (see Figure 2b), mean number of customers



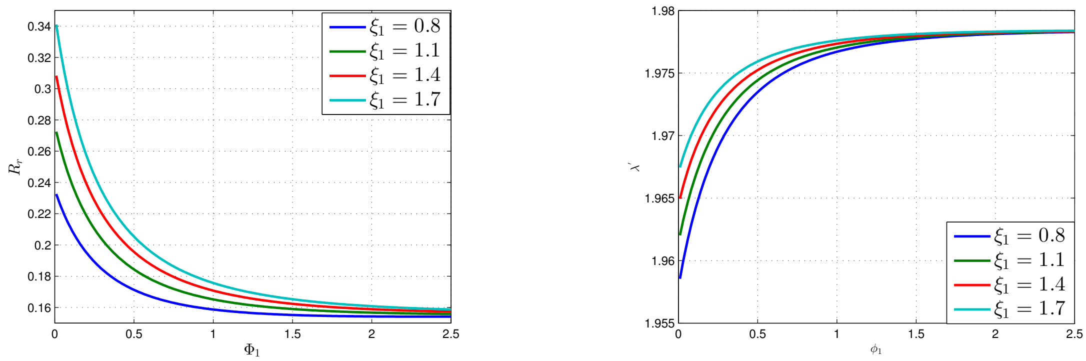
(a) **(b)**

Figure 3: *B_r and L_q vs. α for different values of K*



(a) **(b)**

Figure 4: *B_r and P_{wn} vs. Φ_1 for different values of ξ_0*

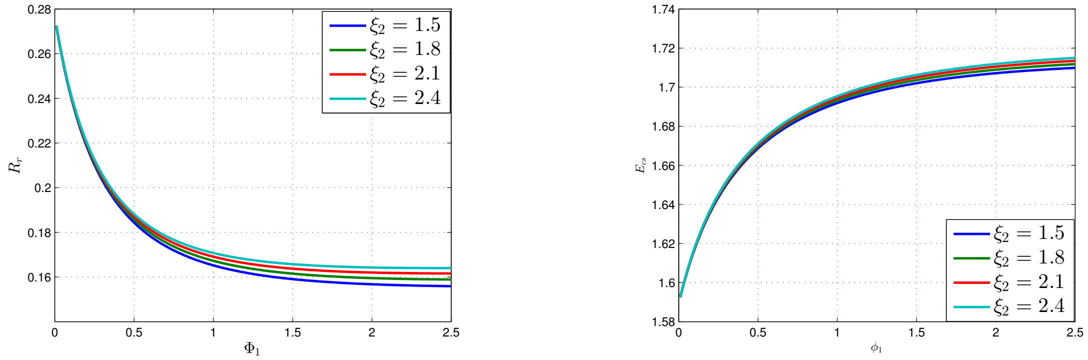


(a) **(b)**

Figure 5: *R_r and λ' vs. Φ_1 for different values of ξ_1*

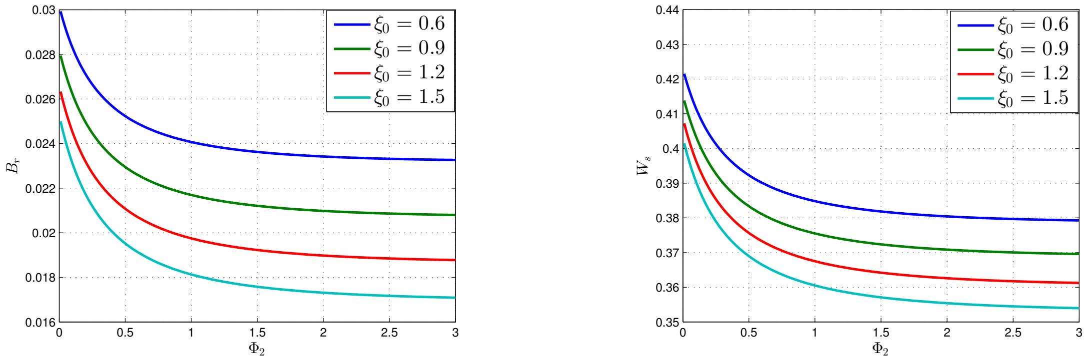
in the queue L_q (see Figure 3b) all increase. Conversely, the probabilities P_{wv1}, P_{wv2} and P_{id} decrease As a result, the average waiting time of a customers in the system decreases. This can be attributed to the effective arrival rate λ' increasing faster than the mean number of customers in the system (L_s).

▷ Effect of c (number of servers): There is clear evidence that as the parameter c increases, the



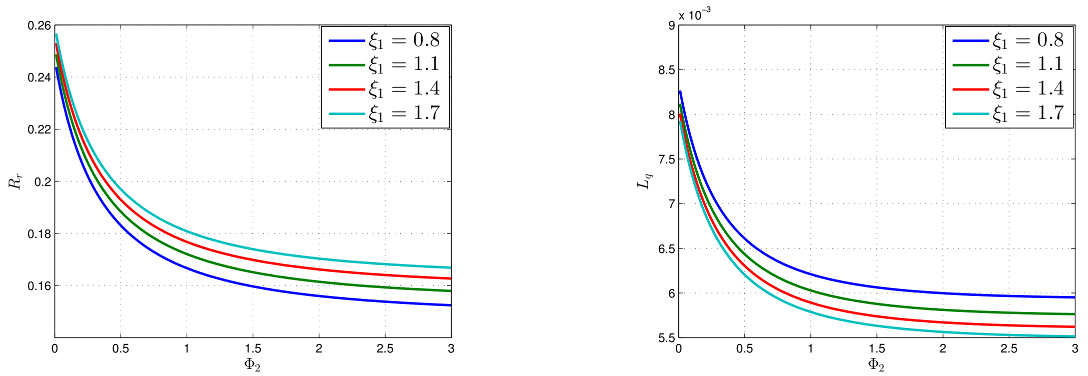
(a) **(b)**

Figure 6: *R_r and E_{cs} vs. Φ_1 for different values of ξ_2*



(a) **(b)**

Figure 7: *B_r and W_s vs. Φ_2 for different values of ξ_0*



(a) **(b)**

Figure 8: *R_r and L_q vs. Φ_2 for different values of ξ_1*

quantity L_q decreases. Moreover, a larger number of servers leads to a higher number of customers being served (see Figure 2b), thereby resulting in a reduced average balking rate (cf. Figure 2a).

▷ Effect of K (system capacity): The system's large capacity of the parameter K encourages more customers to join the queue, hoping to be served, which leads to a decrease in the average

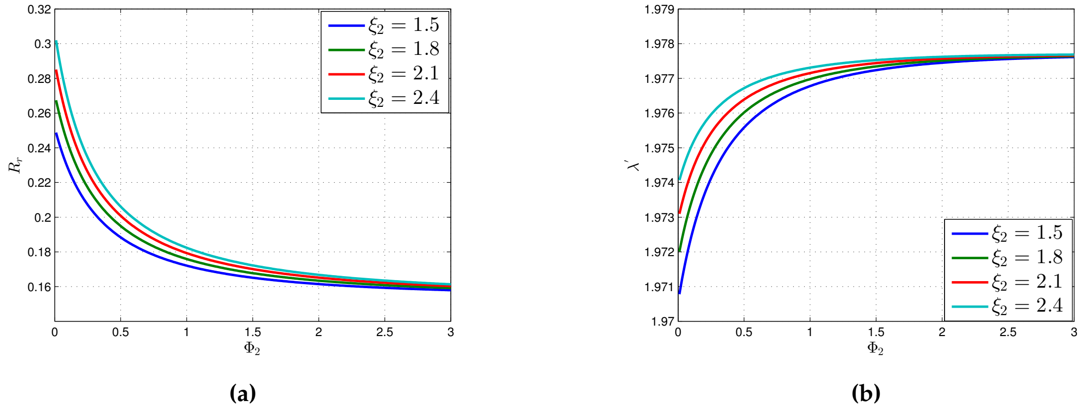


Figure 9: *R_r and λ' vs. Φ_2 for different values of ξ_2*

value B_r (see Figure 3a). Furthermore, as the systems capacity increases, the average number of customers in the queue also increases (cf. Figure 3b). Thus, there is a significant increase in the mean waiting time for customers.

▷ Effect of working vacation rates (Φ_i): By increasing the working vacations rates ϕ_i, $(i = 1, 2)$, the system tends to transition quickly to the normal busy period (see Figure 4b) where customers are served much faster (see Figure 6b). This leads to a decrease in the mean waiting time of the customers (cf. Figure 7b). Then, the system becomes rapidly empty. Consequently, the average value B_r decreases (see Figures 4a and 7a), which implies a growth in effective arrivals (cf. Figures 5b, and 9b). Moreover, higher working vacation rates correspond to, lower average reneging rate (see Figures 5a, 6a, 8a, 9a, and 9b), resulting in smaller mean number of customers in the queue L_q (cf. Figure 8b).

▷ Effect of parameters ξ_0, ξ_1, and ξ_2 (impatience rates) : Increasing impatience rates, whether during busy normal period or working vacations period, results in increased average value R_r (see Figures 5a, 6a, 8a, and 9a) as well as increased mean number E_{cs} (cf. Figure 6b). Additionally, higher impatience rates lead to, a decrease in the average value W_s (see Figure 7b). Consequently, due to this impatience, there is a decrease in the number of customers both L_s and L_q (cf. Figure 8b). This results in a reduced average balking rate and an increased effective arrival rate (see Figures 4a, 7a, 5b, and 9b).

6. Cost optimization

6.1. Cost model

In this section, we propose a model for the costs incurred in our queueing model. In this context, we start by defining the total expected cost per unit of time of the system as:

$$\Upsilon(\mu_1, \mu_2) = C_{wn}P_{wn} + C_{id}P_{id} + C_{wv}(P_{wv1} + P_{wv2}) + C_qL_q + C_rR_r + C_bB_r + c\mu_1C_{\mu_1} + c\mu_2C_{\mu_2},$$

where,

- C_{wn} (resp. C_{id}) denotes the cost per unit time when the servers are working (resp. idle) during normal busy period,
- C_{wv} (resp. C_q) is the cost per unit time when the servers are on type-1 or type-2 working vacation period (resp. when a customer joins the queue and waits for service),
- C_r (resp. C_b) is the cost per unit time when a customer reneges (resp. balks),

- C_{μ_1} (resp. C_{μ_2}) denotes the cost per service per unit time during normal busy period (resp. during type-1 or type-2 working vacation period).

6.2. Grey Wolf Optimizer

The GWO algorithm is one of the recent advancements in swarm intelligence optimization (see [19]). Is inspired by grey wolves in nature, which search for the optimal way to hunt prey. The GWO algorithm uses the same mechanism found in nature, where it follows the hierarchy of the pack to organize the different roles in the pack of wolves. In addition, the GWO algorithm is promising for complex optimization problems. This meta-heuristic algorithm efficiently explores the search space and converges to the optimal solution by simulating the hunting behavior of grey wolves. Its simplicity, versatility and proven success make it an invaluable tool for researchers in a variety of fields. We use this novel technique to globally search (μ_1, μ_2) until the minimum value of $\Upsilon(\mu_1, \mu_2)$ is achieved.

6.3. Numerical Cost Optimum

The main goal is to identify optimal service rates μ_1 and μ_2 in order to minimize the expected cost function. Because optimization problems are complex and highly non-linear, they are challenging to solve analytically. However, we can utilize appropriate nonlinear optimization techniques to determine the optimal solutions in the cost model. In this case, we fix the parameters and employ the grey wolf optimization algorithm to search for the optimal values (u_1^*, u_2^*) for the service rates. The optimization problem can be written as:

$$\min_{\mu_1,\mu_2} \Upsilon(\mu_1, \mu_2)$$
$$\text{s.t} \begin{cases} \mu_1 - \mu_2 > 0, \\ \mu_2 > 0, \\ (\mu_1, \mu_2) \in \mathbb{R}_+^2. \end{cases}$$

The objective is to evaluate the cost function Υ in accordance to parameters μ_1 and μ_2 to minimize the total expected cost incurred by the system using Grey Wolf Optimizer.

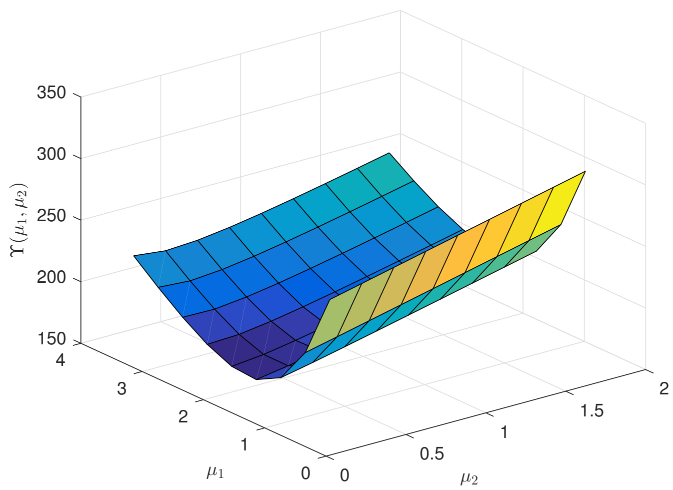


Figure 10: $\Upsilon(\mu_1, \mu_2)$ *vs.* μ_1 *and* μ_2

Figure 10 effectively visualizes the convexity of the objective function Υ according to service rates μ_1 and μ_2.

Then, in what follows, the optimal solutions are given by applying the GWO meta-heuristic for various system parameters. To do this, we fix the parameters as: $C_s = 45$, $C_{id} = 20$, $C_{wv} = 30$, $C_q = 40$, $C_r = 35$, $C_b = 25$, $C_{\mu_1} = 10$, $C_{\mu_2} = 5$.

Table 1: *The optimal* (μ_1^*, μ_2^*) *and* $\Upsilon^*(\mu_1^*, \mu_2^*)$ *for various values of* α *and* K, *when* $\alpha = 8:1:10$, $\Delta = 0.5$, $\beta' = 0.5$, $\Phi_1 = 0.4$, $\Phi_2 = 0.8$, $K = [20; 24; 28]$, $c = 3$, $\xi_0 = 0.6$, $\xi_1 = 0.9$, $\xi_2 = 1.4$.

K	α	μ_1^*	μ_2^*	$\Upsilon^*(\mu_1^*, \mu_2^*)$
	8	3.2914	0.4088	214.5547
20	9	3.6627	0.4387	232.1802
	10	4.0280	0.4648	249.4850
	8	3.3526	0.4278	215.1095
24	9	3.7278	0.4581	232.7175
	10	4.0999	0.4857	249.9891
	8	3.3939	0.4384	216.5381
28	9	3.7759	0.4713	233.1391
	10	4.1548	0.5005	250.3924

Table 2: *The optimal* (μ_1^*, μ_2^*) *and* $\Upsilon^*(\mu_1^*, \mu_2^*)$ *for various values of* Δ *when* $\alpha = 9$, $\Delta = 0.2 : 0.2 : 0.8$, $\beta' = 0.4$, $\Phi_1 = 0.4$, $\Phi_2 = 0.8$, $K = 24$, $c = 3$, $\xi_0 = 0.6$, $\xi_1 = 0.9$, $\xi_2 = 1.4$.

Δ	μ_1^*	μ_2^*	$\Upsilon^*(\mu_1^*, \mu_2^*)$
0.2	3.7858	0.2393	228.1624
0.4	3.7433	0.3986	231.5360
0.6	3.7129	0.5071	233.7100
0.8	3.6941	0.5948	235.3126

Table 3: *The optimal* (μ_1^*, μ_2^*) *and* $\Upsilon^*(\mu_1^*, \mu_2^*)$ *for various values of* β', *when* $\alpha = 9$, $\Delta = 0.5$, $\beta' = 0.3 : 0.2 : 0.9$, $\Phi_1 = 0.4$, $\Phi_2 = 0.8$, $K = 24$, $c = 3$, $\xi_0 = 0.6$, $\xi_1 = 0.9$, $\xi_2 = 1.4$.

β'	μ_1^*	μ_2^*	$\Upsilon^*(\mu_1^*, \mu_2^*)$
0.3	3.6668	0.5143	235.5446
0.5	3.7264	0.4575	232.7175
0.7	3.7621	0.4140	230.9342
0.9	3.7850	0.3798	229.6789

- From Table 1, it can be clearly seen that the optimum expected cost $\Upsilon^*(\mu_1^*, \mu_2^*)$ exhibits a significant increase as the values of the arrival rate α and finite capacity K increases.
- From Table 2, can be observed that as Δ, value increases, the minimum expected cost increases. This observation clearly indicates that increasing the waiting rate of servers is an expensive endeavor.
- From Table 3, when interruption probability (β') increases, the minimum expected cost decreases. So, a higher interruption probability positively affects the overall expected cost of the system.

7. Conclusion

This paper focused on the analysis of a finite-space multi-server queue where customers exhibit impatience under a synchronous differentiated working vacation policy. Specifically, the customers are assumed to be impatient during the normal busy period, as well as during type-1 and type-2 working vacations. The main goal of the analysis is to determine the steady-state probabilities of the system size under different server states, including the normal busy period and the working vacations (type-1 and type-2). This is achieved through the application of recursive analysis techniques. We derived important system performance measures that provide valuable insights into the behavior and efficiency of the considered multi-server queueing system.

A GWO algorithm is performed to determine the optimal service rates for both working vacations and normal busy periods aiming to minimize the expected total cost. The problem at hand is formulated as a nonlinear optimization problem, and several numerical examples are provided to illustrate the effectiveness of the proposed approach. The focus of the analysis is on conducting a cost optimization study, where the effect of different system parameters and cost elements is investigated. The numerical examples and cost optimization analysis presented in this study shed light on the significance of system parameters and cost elements in queueing systems. Overall, this study contributes to the understanding and optimization of queueing systems, highlighting the potential advantages of cost optimization techniques in various real-life and industrial settings.

The model discussed in the paper can be extended to handle more complex scenarios, such as an unreliable multi-server queue with heterogeneous customers, which introduces additional complexity to the problem. While this extension increases the dimension of the problem significantly. It is also possible to relax the exponential assumptions by considering phase-type distributions for service times.

References

[1] Afroun, F., Aïssani, D., Hamadouche, D., and Boualem, M. (2018). Q-matrix method for the analysis and performance evaluation of unreliable $M/M/1/N$ queueing model. *Mathematical Methods in the Applied Sciences*, 41:9152–9163.

[2] Baba, Y. (2010). The $M/PH/1$ queue with working vacations and vacation interruption. *Journal of Systems Science and Systems Engineering*, 19:496–503.

[3] Bouchentouf, A.A. and Guendouzi A., and Kandoucib, A. (2019). Performance and economic study of heterogeneous $M/M/2/N$ feedback queue with working vacation and impatient customers. *ProbStat Forum*, 12:15–35.

[4] Bouchentouf, A.A., Boualem, M., Yahiaoui, L., and Ahmad, H. (2022). A multi-station unreliable machine model with working vacation policy and customers impatience. *Quality Technology & Quantitative Management*, 19:766–796.

[5] Bouchentouf, A.A., Cherfaoui, M., and Boualem, M. (2019). Performance and economic analysis of a single server feedback queueing model with vacation and impatient customers. *Opsearch*, 56:300–323.

[6] Bouchentouf, A. A., Guendouzi, A., and Majid, S. (2020). On impatience in Markovian $M/M/1/N/DWV$ queue with vacation interruption. *Croatian Operational Research Review*, 11:21–37.

[7] Bouchentouf, A.A., Cherfaoui, M., and Boualem, M. (2021). Analysis and performance evaluation of Markovian feedback multi-server queueing model with vacation and impatience. *American Journal of Mathematical and Management Sciences*, 40:261–282.

[8] Chettouf, A., Bouchentouf, A.A., and Boualem, M. (2024). A Markovian Queueing Model for Telecommunications Support Center with Breakdowns and Vacation Periods. *Oper. Res. Forum* 5, article number 22. https://doi.org/10.1007/s43069-024-00295-y

[9] Doshi, B. T. (1986). Single server queues with vacation: A survey. *Queueing Systems*, Vol. 1, No. 1, 29–66.

[10] Doshi, B. T. (1990). Single server queues with vacations. *Stochastic Analysis of the Computer and Communication Systems*, 217–264.

[11] Jyothsna, K., Laxmi, P, V., and Kumar, V. P. (2022). Analysis of $GI/M/1/N$ and $GI/Geo/1/N$ queues with balking and vacation interruptions. *Journal of Mathematical Modeling*, 10:569-585.

[12] Prakati, P. (2024). $M/M/C$ queue with multiple working vacations and single working vacation under encouraged arrival with impatient customers. *Reliability: Theory and Applications*, 19 :650-662.

[13] Laxmi, P. V., and Jyothsna, K. (2015). Impatient customer queue with Bernoulli schedule vacation interruption. *Computers and Operations Research*, 56:1–7.

[14] Li, J., and Tian, N. (2007). The $M/M/1$ queue with working vacations and vacation interruptions. *Journal of Systems Science and Systems Engineering*, 16:121–127.

[15] Li, J. H., Tian, N. S., and Ma, Z. Y. (2008). Performance analysis of $GI/M/1$ queue with working vacations and vacation interruption. *Applied Mathematical Modelling*, 32:2715–2730.

[16] Majid, S., and Manoharan, P. (2018). Impatient customers in an $M/M/c$ queue with single and multiple synchronous working vacations. *Pakistan Journal of Statistics and Operation Research*, 14:571–594.

[17] Majid, S., Bouchentouf, A. A., and Guendouzi, A. (2021). Analysis and optimisation of a $M/M/1/WV$ queue with Bernoulli schedule vacation interruption and customers impatience. *Acta Universitatis Sapientiae, Mathematica*, 13:367–395.

[18] Majid, S., Manoharan, P., and Ashok, A. (2018). An $M/M/1$ queue with working vacation and vacation interruption under Bernoulli schedule. *International Journal of Engineering and Technology*, 7:448–454.

[19] Mirjalili, S., Mirjalili, S. M., and Lewis, A. (2014). Grey wolf optimizer . *Advances in Engineering Software*, 69:46–61.

[20] Sasikala, S., and Abinaya, V. (2023). The $M/M/c/N$ interdependent inter Arrival queueing model with controllable arrival rates, reverse balking and impatient Customers. *Arya Bhatta Journal of Mathematics and Informatics*, 15:1–10.

[21] Shekhar, C., Varshney, S., and Kumar, A. (2021). Matrix-geometric solution of multi-server queueing systems with Bernoulli scheduled modified vacation and retention of reneged customers: A meta-heuristic approach. *Quality Technology & Quantitative Management*, 18:39–66.

[22] Schwarz, M., Sauer, C., Daduna, H., Kulik, R., and Szekli, R. (2006). $M/M/1$ queueing systems with inventory. *Queueing Systems*, 54:55–78.

[23] Stolletz, R. (2012). Performance analysis and optimization of inbound call centers. *Springer Science and Business Media*, 528:1–25. https://doi.org/10.1007/978-3-642-55506-0.

[24] Servi, L. D., and Finn, S. G. (2002). $M/M/1$ queues with working vacations ($M/M/1/WV$). *Performance Evaluation*, 50:41–52.

[25] Yue, D., and Yue, W. (2009). Analysis of an $M/M/c/N$ queueing system with balking, reneging, and synchronous vacations. *Advances in Queueing Theory and Network Applications*, 165–180.

[26] Ziad, I., Laxmi, P. V., Bhavani, E. G., Bouchentouf, A. A., and Majid, S. (2023). A matrix geometric solution of a multi-server queue with waiting servers and customers impatience under variant working vacation and vacation interruption. *Yugoslav Journal of Operations Research*, 3:389–407. https://doi.org/10.2298/YJOR220315001Z.