RAIRO-Oper. Res. 59 (2025) 2779–2803
https://doi.org/10.1051/ro/2024192

RAIRO Operations Research
www.rairo-ro.org

MODELING AND OPTIMIZATION OF AN $M/M/1/K/WV$ RETRIAL QUEUE WITH DUAL-PHASE SERVICE, CUSTOMER'S BALKING UNDER F-POLICY

Sudeep Singh Sanga* and Ramesh Vankudothu

Abstract. This paper examines an $M/M/1/K/WV$ queue, incorporating features such as retrial queue, working vacation (WV), admission control F-policy, customer balking, and dual-phase service. Upon arrival, customers observing an extended queue length have the option to refrain from joining the queue, a phenomenon known as balking. Alternatively, if a customer decides to join the queue but finds the service facility occupied, they are directed to a virtual space termed the retrial orbit. After a random duration, customers in the retrial orbit make subsequent attempts to avail the service. In our study, dissatisfied customers with the initial phase of service have the option to choose a secondary phase of service. The F-policy adeptly regulates customer congestion by limiting admissions when the system capacity is reached and resuming admissions once the capacity decreases to a predefined threshold value F. Introducing the working vacation concept in our study ensures continuous service provision during the vacation period, albeit at a reduced rate compared to the standard operational pace. This multifaceted approach contributes to a more nuanced and flexible management of customer influx in the system. The model is mathematically formulated by developing steady-state difference equations, which are then solved using the matrix analytical method. Diverse system metrics are subsequently derived, providing valuable insights for system managers to effectively analyze and navigate the queueing management and decision-making processes. We also present a numerical example to illustrate the impact and behavior of input parameters on a range of system metrics. Furthermore, we employ the adaptive neuro-fuzzy inference system (ANFIS) to validate our analytical findings against the ANFIS results. Establishing a bivariate cost function is a noteworthy aspect of our research, and we minimize cost function using both the quasi-Newton method (QNM) and genetic algorithm (GA), enhancing the robustness and reliability of our optimization approach. The practical justification of the model is demonstrated in an automotive service facility (ASF) where malfunctioning bikes (MBs) arrive for servicing.

Mathematics Subject Classification. 60K25, 60K30.

Received June 15, 2024. Accepted September 21, 2024.

1. Introduction

In many service systems where customer queues are prevalent, customers seek immediate service, while servers strive to attend to customers promptly. However, managing the flow of customers can be challenging, especially in cases with a large volume of customers. The high demand for service combined with limited server capacity often leads to delays and queue formations, impacting customer satisfaction and operational efficiency in the

Keywords. Retrial queue, working vacation, F-policy, balking, dual phase service, ANFIS, cost, QNM, GA.

Department of Mathematics, SV National Institute of Technology, Surat, Gujarat, India.
*Corresponding author: ssanga@amhd.svnit.ac.in

service system. Efficient queue management strategies and optimization techniques are essential to reduce these challenges and improve the system's overall performance. Retrial queues (RQs) are important in mitigating congestion problems at service systems' servers. When an arriving customer finds the server busy, they move to a virtual waiting area known as the retrial orbit (RO). Customers make subsequent service requests from the RO, attempting to secure service from the server again once it becomes available. This mechanism helps manage server congestion by temporarily diverting waiting customers to an alternative virtual queue, facilitating a more efficient utilization of server resources and reducing customer wait times. The application of RQ extends across various sectors, including call centers, telecommunications networks, railway ticket counters, transportation and automobile systems, and more. To illustrate the practical significance of RQs, let us consider a scenario at a bank counter where a queue forms. When customers arrive at the bank and encounter a busy cashier (server), they may proceed to a designated waiting area (retrial orbit). After a waiting period, customers return to the cash counter, anticipating that the cashier will be available to assist them. Significant contributions to the study of RQs in the literature have been made by notable researchers such as Falin and Templeton [1], Falin [2], and Artalejo and Gomez [3]. These works have played a key role in advancing our understanding of retrial queueing systems.

In addition to retrial queues, admission control of customers is an important aspect of addressing congestion problems in service systems. Admission control is key, particularly when the system has finite capacity limitations. In our study, we implemented an admission control F-policy to regulate customer arrivals in the proposed study. The F-policy serves as an effective mechanism to control customer admissions when the system reaches full capacity (say K). According to the F-policy, once the system capacity is reached, further admission of customers is temporarily suspended until the number of customers served decreases to a predefined threshold F (where $0 \leq F \leq K-1$). For a comprehensive understanding of the F-policy, researchers can refer to key studies such as Gupta [4] and Jain and Sanga [5]. Jain and Sanga [6] examined the single-server Markovian queueing model under admission control F-policy and other queueing characteristics. Using a recursive method, they derived steady-state queue size distributions, which were used to establish various performance measures to predict the system's behavior. Sanga and Charan [7] proposed an $M/M/1/K$ queueing model under admission control F-policy. Their studies established the steady-state probabilities for different states using a recursive technique, which has been further used to develop the performance indices. Wu *et al.* [8] analyzed the $G1/M/1$ queueing model under an F-policy with alternating service mechanisms. They derived the steady-state solution of the model and further extended their analysis by developing key performance measures. They employed the Non-dominated sorting genetic algorithm-II (NSGA-II) to optimize the system's performance.

Customer discouragement represents a significant issue in real-world service scenarios. In some instances, customers may become discouraged and opt not to join a queue due to the presence of a large number of customers already in the system. This phenomenon is commonly referred to as "balking". Customer balking occurs when arriving customers decide not to join the queue and leave the premises without receiving service. Customer balking is observed in various service settings, including banks, post offices, and other service facilities with queueing systems. we can refer to Haight [9] and Jain *et al.* [10] for a detailed description of customer's balking. Consider a popular movie theater that regularly showcases blockbuster movies. During weekends or special occasions, the theater encounters a surge in ticket demand, often leading to long queues at the ticket counter when the seating capacity is reached. In such situations, movie attendees may realize that obtaining a ticket immediately is unlikely. Consequently, some individuals may decide to balk, choosing not to wait in the queue due to the overcrowding at the ticket counter in the theater. Jain and Sanga [11] paid attention to customer balking behavior in a finite capacity Markovian queueing system under the admission control F-policy. They formulated the ChapmanKolmogorov equations and solved them using recursive techniques to establish the steady-state queue size probability distribution. In the recent year, Sanga and Antala [12] investigated a state-dependent finite capacity queueing model with a single unreliable server, focusing on the admission control F-policy and balking. The authors utilized the matrix analytic method to compute the probability of various system states.

In queueing models, a notable scenario occurs when the system becomes empty and the server takes a vacation. During this vacation period, arriving customers are served slower than usual, commonly called "working vacation". Extensive research on the concept of working vacation in queueing theory has been conducted by Chandrasekaran and Indhira [13] and Servi and Finn [14]. These studies give insight into the theoretical foundations, analytical methodologies, and practical implications of working vacation models in queueing systems. Consider a practical example in a hospital setting where no patients are present, prompting the attending doctor to take a break. If patients arrive seeking treatment during the doctor's absence, nurses and other healthcare workers assume responsibility by providing interim care, which may include administering medications, albeit at a reduced rate. This scenario serves as an illustration of the application of a working vacation model within healthcare settings. Yang *et al.* [15] explored the F-policy $M/M/1/K$ queueing system with a working vacation and exponential startup time. The authors employed the matrix-analytic method to generate steady-state probabilities and then derived several system performance metrics. Jain and Dhibar [10] investigated the Markovian queue, which includes working vacation, balking, and other queueing features. They computed system queue size probabilities, which were then used to construct many performance indices in closed form. Bouchentouf *et al.* [16] paid attention to working vacations and customers' impatience behavior in finite capacity Markovian queueing systems. The authors derived steady-state probabilities and closed-form expressions for several system characteristics. In a recent study, Sanga and Antala [17] examined vacation and the admission control F-policy in the $M/M/1/K$ queueing model. They constructed the steady-state probability distribution for various system states and derived various performance measures to predict the system behavior. Chahal and Kumar [18] have investigated the vacation policy in a machining system and derived the steady-state solution for the model. Agarwal *et al.* [19] examined a queueing model that integrates the concept of working vacations. Their study focused on the operational dynamics during periods when the server operates at a reduced service rate instead of completely ceasing service.

The main or leading service that customers are seeking when they join the queue in the queueing system is called the primary service (normal service). Normal service is the main objective for the customers as they are waiting in the queue for service. The system's primary service plays a crucial role in determining customer satisfaction and overall system performance. For example, a bank's primary service could be cash withdrawal or account inquiries. In a hospital, the primary service might be medical consultations or treatment. The service that is provided to unsatisfied customers during primary service or service provided during a working vacation is known as secondary service (additional service). Customers who are unsatisfied during primary service have a chance of entering the queue for secondary service with some probability. For example, if we consider the above example used in working vacation, when the doctor takes a short break, the patients are served by nurses and other healthcare workers. In such situations, patients are not satisfied with their service, so the doctor provides service to them. This is known as secondary service.

Queueing system parameters, particularly service rates, are crucial in optimizing service systems. When service is organized effectively to minimize customer waiting times and system costs, the service system operates smoothly and enhances customer satisfaction. Optimizing service rates and other queueing parameters is fundamental to achieving efficient and effective service delivery. By strategically managing these parameters, service providers can minimize customer wait times, reduce system costs, and improve overall system performance. Cost optimization not only enhances operational efficiency but also contributes to higher levels of customer satisfaction by ensuring timely and cost-effective service delivery. The cost function can be inherently complex due to the involvement of multiple cost elements across different stages and the probabilistic nature of various system states. Analyzing and optimizing such a cost function poses a significant challenge, necessitating the use of iterative methods for effective solutions. Iterative/metaheuristic methods such as the quasi-Newton (cf. [20,21]) and metaheuristic-based genetic algorithm (GA) (cf. [22,23]) are valuable tools for addressing complex cost functions in queueing systems.

After conducting an extensive survey of the queueing literature, a notable research gap has been identified concerning the lack of focus on admission control of customers (F-policy) in finite queues with working vacations. Specifically, limited research explores integrating admission control strategies in the context of finite queueing systems that incorporate working vacation policies. Furthermore, incorporating retrial queues and dual-phase service into the customer's processes adds complexity and uniqueness to this study. The intersection of these features presents an opportunity to investigate how admission control F-policy can be explored to enhance system performance and customer satisfaction in $M/M/1/K/WV$ retrial queue and other advanced service mechanisms. The innovative aspects of the proposed study are as follows:

- We develop an $M/M/1/K/WV$ retrial queueing model with various queueing features, including dual-service and admission control, using the F-policy.
- One innovative aspect of our model is incorporating the "working vacation" concept for the server. This feature allows the server to continue providing service to customers even while on vacation, albeit at a reduced rate compared to normal operations. This approach ensures continuity of service in the queueing system.
- The implementation of dual-phase service to enhance customer satisfaction. This dual-phase service approach ensures that customers receive comprehensive and effective service through multiple stages. By dividing the service process into distinct phases, we optimize the customer experience, improving service quality and overall system performance.
- The integration of the F-policy in the $M/M/1/K/WV$ retrial queueing model makes the model distinct from others in its ability to effectively control arriving customers during periods of congestion. The F-policy restricts further customer admissions when the system reaches its capacity. This innovative feature enhances system efficiency and customer satisfaction by managing service demand in real time.
- The incorporation of customer balking allows customers to decide whether to join or leave the system based on observed queue length. Our model enhances overall system adaptability and customer satisfaction by accommodating customer preferences through balking behavior.
- The matrix analytical method is employed to solve the steady-state governing equations and establish probability distributions for different system states.
- Our model involves the development of various formulae for several system metrics, including system length, waiting time, and throughput. These formulae allow us to explore and analyze the system's performance by observing how these metrics vary under different conditions and scenarios.
- The AI-based adaptive neuro-fuzzy inference system (ANFIS) technique is incorporated to validate the accuracy of the numerical results obtained from the analytical method. ANFIS approach leverages advanced computational algorithms to assess the consistency and reliability of the analytical findings in the $M/M/1/K/WV$ retrial queueing model.
- A non-linear cost function with dual decision variables is formulated and minimized using the quasi-Newton method (QNM) and GA.
- The model's applicability is demonstrated at the automotive service facility (ASF), where malfunctioning bikes (MBs) are brought in for service.

The list of notations and abbreviations used in the paper is presented in Tables 1 and 2, respectively. The paper is structured into nine sections, with the current introductory section forming part of the ongoing discussion. Section 2 provides an overview of the QNM, GA, and ANFIS preliminaries. Section 3 details the model description and its limitations. Practical validation of the model is demonstrated in Section 4. The mathematical analysis of the model using the matrix analytic method is conducted in Section 5. Various system metrics and the cost function are developed in Section 6. Section 7 presents the numerical results and ANFIS computations. Cost optimization using QNM and GA methods to determine the optimal service rates is performed in Section 8. Finally, the study concludes with Section 9, which presents concluding remarks and discusses the future scope of the study.

TABLE 1. List of notations.

Notation	Description
K	System capacity
F	Threshold value
μ	Primary service rate during normal mode
μ_a	Secondary service rate during normal mode
λ	Arrival rate of customers
θ	Vacation time
γ	Retrial rate
η	Primary service rate during working vacation
η_a	Secondary service rate during working vacation
ξ	Setup rate during normal mode
ξ_v	Setup rate during WV mode
q	Joining probability of the customer into the system
$\bar{q}$	Balking probability of a customer
p	Leaving probability of a customer from the system
$\bar{p}$	Joining probability of customer for secondary service
t	Time
$H(t)$	Number of customers at time t
$X(t) = i$	Server's status
$P_{(i,n)}$	Probability of n customer present in the system at ith level
Π	Steady-state probability
L_S	Expected number of customers in the system
$L_S(N)$	Expected number of customers in the system during normal service mode
$L_S(\mathrm{WV})$	Expected number of customers in the system during WV
W_T	Expected waiting time
$W_T(N)$	Expected waiting time during normal service
$W_T(\mathrm{WV})$	Expected waiting time during the WV
$\lambda_{\mathrm{eff}}(N)$	Effective arrival rate of customers during normal service
$\lambda_{\mathrm{eff}}(\mathrm{WV})$	Effective arrival rate of customers during the WV
P_{LN}	Loss probability during normal service
P_{LWV}	Loss probability during the WV
P_{BN}	Balking probability during normal service
P_{BWV}	Balking probability during the WV
ρ_N	Traffic intensity of normal service
ρ_{WV}	Traffic intensity of WV
P_B	Probability of the server when it is in busy mode
P_F	Probability of the server when it is free
P_{WV}	Probability of the server when it is in the WV mode
$\mathrm{TC}(\mu, \mu_a)$	Cost function
$\mathrm{TC}(\mu^*, \mu_a^*)$	Optimal cost
$s(z)$	Actual values obtained by analytical formulae for a particular data set z
$\bar{s}(z)$	Mean of actual data for a particular data set z
$r(z)$	Predicted values by ANFIS for a particular data set z
M	Overall count of the data set used or prediction
R^2	Coefficient of determination

TABLE 2. List of abbreviations.

Abbreviation	Expansion
ANFIS	Adaptive neuro-fuzzy inference system
ANNs	Artificial neural networks
APE	Absolute percentage error
ASF	Automotive service fecility
Exp-D	Exponential distribution
FCFS	First-come-first-served
GA	Genetic algorithm
MAE	Mean absolute error
MAM	Matrix analytical method
MBs	Malfunctioning bikes
QNM	Quasi newton method
RMSE	Root mean square error
RO	Retrial orbit
RQs	Retrial queues
TC	Total cost
TP	Throughput
WV	Working vacation

2. Preliminaries of QNM, GA and ANFIS

This section introduces the foundational concepts of the QNM and GA, which play crucial roles in optimizing the system cost of the proposed model. Furthermore, the ANFIS approach will be employed to validate the numerical results obtained from the optimization methods.

2.1. Quasi-Newton method (QNM)

The QNM is a class of optimization algorithms used to solve optimization problems. It belongs to the family of iterative numerical optimization techniques and is particularly useful for solving unconstrained optimization problems where the objective function is smooth and continuous. For a detailed description of the QNM, one can refer to Rao [24].

In providing an overview of QNM, we introduce the vector $\vec{A}$, summarising the decision variables relating to service rates μ and μ_a. To facilitate this process, the vectorization aids in the deriving of a cost gradient $\vec{\nabla}\mathrm{TC}(\vec{A}) = \left[\frac{\partial \mathrm{TC}}{\partial \mu}, \frac{\partial \mathrm{TC}}{\partial \mu_a}\right]^T$ that contributes to the process. Let $(\mu^*,\ \mu_a^*)$ be the optimal value of the decision variable (μ, μ_a). The process of the QNM involves the following steps:

(1) Set the initial value $\vec{A_o} = [\mu, \mu_a]$ and the tolerance 10^{-6}.
(2) Determine the value of the objective function (cost function) at $\vec{A_o}$ *i.e.*, $\mathrm{TC}(\vec{A_o})$.
(3) Subsequently, at the initial vector $\vec{A_o}$, determine the cost gradient and Hessian matrix for the cost function as $\vec{\nabla}\mathrm{TC}(\vec{A}) = \left[\frac{\partial \mathrm{TC}}{\partial \mu}\, \frac{\partial \mathrm{TC}}{\partial \mu_a}\right]^T\Big|_{\vec{A_o}}$ and $H(\vec{A_j}) = \begin{bmatrix} \frac{\partial^2 \mathrm{TC}}{\partial \mu^2} & \frac{\partial^2 \mathrm{TC}}{\partial \mu \partial \mu_a} \\ \frac{\partial^2 \mathrm{TC}}{\partial \mu \partial \mu_a} & \frac{\partial^2 \mathrm{TC}}{\partial \mu_a^2} \end{bmatrix}$
(4) For the new solution evaluate $\vec{A}_{j+1} = \vec{A_j} - [H(\vec{A_j})]^{-1}\vec{\nabla}\mathrm{TC}(\vec{A_j}), j = 0, 1, ...$
(5) Repeat steps (3) to (4) until the stopping criterion is met *i.e.*, $\max\left(\left|\frac{\partial \mathrm{TC}}{\partial \mu}\right|, \left|\frac{\partial \mathrm{TC}}{\partial \mu_a}\right|\right) < 10^{-6}$.
(6) Compute the optimal value (μ^*, μ_a^*) and minimum total cost $\mathrm{TC}(\mu^*, \mu_a^*)$.

2.2. Genetic algorithm (GA)

The genetic algorithm is a metaheuristic approach recognized as one of the most powerful optimization techniques. Inspired by the principles of natural selection and genetics observed in biological evolution, this method efficiently tackles optimization challenges across linear and non-linear problem domains. The GA comprises three fundamental elements in each iteration: (i) Selection, (ii) Crossover, and (iii) Mutation. The selection process prioritizes chromosomes based on their fitness values, favoring those representing better solutions to the optimization problem. During crossover, pairs of chromosomes exchange genetic information to generate new offspring with potentially improved characteristics. Mutation introduces random changes to the offspring's genes, enhancing diversity in the population. These steps are iteratively executed until a stopping criterion is met or a predefined number of generations is reached, aiming to converge toward optimal solutions for the cost function of the proposed model. The GA employs the following steps to solve an optimization problem effectively:

(1) *Population initialization*: an initial population of chromosomes is randomly generated, and genes are chosen based on the quality of the chromosomes. The problem's complexity and constraints determine the size of the population.
(2) *Selection*: chromosomes are selected from the population to form a new generation. Chromosomes are selected with a probability proportionate to their fitness, and the chromosomes with greater fitness are more likely to be chosen due to their higher chances of selection.
(3) *Crossover*: the selected chromosomes undergo a pairwise crossover operation, which involves exchanging genetic information to create new offspring chromosomes. The crossover mimics genetic recombination in natural reproduction. Crossover is evolved when the newly generated chromosomes provide improved solutions compared to the original ones.
(4) *Mutation*: the generation of new offspring occurs once the crossover process is completed and the mutation operator is applied to these offspring using a specified mutation rate. This operator can modify one or more chromosome genes based on the mutation rate. This mutation process can occur multiple times for each offspring. The gene's value changes if the randomly generated number exceeds the mutation probability.
(5) *Fitness value*: to determine the minimum fitness value, the fitness function (cost function $\mathrm{TC}(\mu, \mu_a)$) is evaluated. In this context, μ and μ_a represent decision variables, and the chromosome fitness value is generated through $\mathrm{TC}(\mu, \mu_a)$.
(6) *Termination*: the algorithm continues to select, crossover, and mutate chromosomes over several generations until a termination condition is fulfilled. Common termination criteria involve reaching a predetermined maximum number of generations or attaining a desired satisfactory fitness level.

2.3. Overview of ANFIS

The ANFIS is a well-established mathematical approach that combines the capabilities of artificial neural networks (ANNs) and fuzzy logic systems, which address complex relationships and patterns within the data. ANFIS introduces a fusion of two distinct yet complementary computational techniques of the intuitive linguistic reasoning of fuzzy logic with the data-driven adaptability of neural networks. ANFIS operates by integrating fuzzy logic concepts and neural network learning algorithms. ANFIS architecture consists of five distinct layers: (i) Input layer, (ii) Fuzzification layer, (iii) Rule layer, (iv) Defuzzification layer, and (v) Output layer. Each distinct layer has a specific role in transforming input data into desired outputs, involving membership functions, rule generation, and overall inference. It initiates the transformation of input data into fuzzy linguistic terms using membership functions, which assign degrees of belonging to different categories. Subsequently, the fuzzified inputs traverse the rule layer to generate rule-firing strengths, quantifying each fuzzy rule's relevance. Then, the defuzzification layer combines these strengths to yield the final output. The neural network learning process is a part of ANFIS that adjusts the model's parameters to minimize the error between predicted and actual outputs using techniques like gradient descent that enhance the accuracy and reliability of the model. The primary objective of ANFIS is to offer a versatile solution for uncovering intricate patterns, making accurate predictions, and offering valuable insights across a broad spectrum of research and application domains. Its core purpose is to bridge the gap between human-like interpretability and data-driven adaptability, empowering researchers and practitioners to navigate intricate data landscapes more effectively and derive meaningful conclusions. For a detailed description of ANFIS, one can refer to Jang [25], Dounis *et al.* [26], Sanga and Jain [27] and Divya and Indhira [28].

3. Model description

In formulating the mathematical model for an $M/M/1/K/WV$ queue with a retrial, customer balking, and WV under the F-policy, several assumptions are made, outlined as follows:

- A single server is present in the system to assist customers, and it operates based on the first-come, first-served (FCFS) discipline.
- Arriving customers have the option to be discouraged by system congestion, choosing not to join the queue and balking indefinitely. It is assumed that if arriving customers decide to join the queue, they do so with a probability q, and balk with a probability $(1-q)$.
- Customer arrivals in the system are modeled to follow a Poisson distribution with a rate of λ.
- Customers who find the server occupied join the retrial orbit, where they make a re-attempt for service at a rate γ following an exponential distribution.
- A two-phase service process is strategically designed to enhance customer satisfaction. The service process unfolds in two distinct phases, each contributing to the overall customer experience.
- In phase 1, customers are served at a rate μ according to an exponential distribution (Exp-D). If customers are not satisfied with the service provided in this phase, they choose to opt for the second phase of service with a probability denoted as $\bar{p} = (1-p)$. In the second phase, the server renders services at a rate μ_a following Exp-D.
- When the server becomes idle, it transitions into a working vacation mode with rate θ according to Exp-D. In this mode, the server undergoes maintenance activities and other essential tasks. During this period, the server provides services at a slower rate η than its standard operational pace (*i.e.*, $\eta < \mu$). Similar to the standard operational process, dissatisfied customers opt for the second-phase service with a probability denoted by $\bar{p} = (1-p)$. In this phase, the server facilitates them at a rate η_a following an Exp-D.
- When the system attains its capacity (K), a setup time, following an Exp-D with a rate ξ (ξ_v during working vacation), restricts the entry of customers. Additional customer admission is only permitted when the system size decreases to a predefined threshold F $(0 \leq F \leq K)$.

The block diagram of the proposed model is illustrated in Figure 1.

Limitations: in addition to highlighting the strengths of the model, this study also identifies a few limitations that are outlined as:

- In this study, we assume the server is 100 percent reliable, *i.e.*, the server will never experience failures.
- This study does not consider the concept of customer reneging, *i.e.*, once customers join the queue, they cannot leave the queue without being served.

Consider the bi-variate Markov process denoted by $\{(H(t), X(t)), t \geq 0\}$, where $H(t)$ and $X(t)$ represent, respectively, the number of customers and the status of the server at time t. The server's status, denoted by $X(t)$, is defined as follows:

$$X(t) = i = \begin{cases} 0 & \text{the server provides service during admission control } F\text{-policy},\\ 1(2) & \text{the server provides primary (secondary) service during normal busy state},\\ 3 & \text{the customers are waiting in the retrial orbit},\\ 4(5) & \text{the server provides the secondary (primary) service during the WV period},\\ 6 & \text{the server provides the service during admission control } F\text{-policy in WV period}. \end{cases}$$

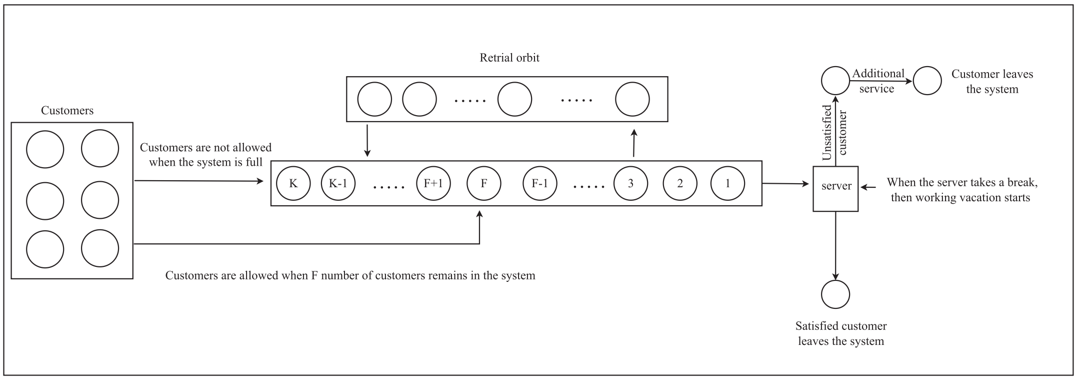


FIGURE 1. Block diagram of $M/M/1/K/WV$ retrial queueing model.

The probability of the system state at time t for the node (i, j) is expressed as $P_{(i,n)}(t) = Prob\{X(t) = i, H(t) = n\}$. Where $X(t)$ and $H(t)$ represent the state of the system and the number of customers present at time t, respectively. The analysis of the model is conducted in a steady-state environment, denoted as $P_{(i,n)} = \lim_{t\to\infty} P_{(i,n)}(t)$, where $i = 0, 1, \ldots, 6$.

4. Motivating example: Automotive Service Facility (ASF)

Automobiles, including cars, bikes, trucks, buses, and similar vehicles, represent self-propelled wheeled conveyances designed for transporting passengers or cargo. These vehicles commonly necessitate maintenance and services, including brake system adjustments, engine diagnostics, battery servicing, and routine upkeep, to uphold operational safety and efficiency standards. The proposed $M/M/1/K/WV$ retrial queueing model holds relevance across diverse domains, including the automobile sector, manufacturing systems, banking institutions, healthcare facilities, etc. Here, we establish the contextual framework in the automobile sector to demonstrate the applicability of the proposed model. This sector exemplifies a pertinent domain where the theoretical constructs of model and analytical methodologies can be effectively employed.

Consider an ASF specializing in repairs for MBs. This establishment operates under a finite capacity constraint, denoted as K, which represents the maximum number of MBs that can be accommodated on its premises at any given time. Once this capacity threshold is reached, the ASF ceases to accept additional MBs for service until the number of MBs decreases to the threshold F, thereby ensuring operational efficiency and minimizing waiting times. In this situation, the ASF is staffed with one mechanic who serves as the primary server for handling MBs, which serve as customers. Additionally, there is one helper personnel who assists the mechanic and operates as a secondary server during the working vacation mode, which occurs when the main mechanic is on vacation. When a bike experiences a malfunction, the owner brings it to the nearest ASF for servicing. Upon arrival at the ASF, MBs seek service from the mechanic. If the mechanic is available and not occupied with another repair, the MB receives immediate attention and is repaired. However, if the mechanic is busy attending to another MB, the newly arrived MB enters a virtual waiting area and awaits service. In some instances, if the waiting time becomes excessively long and the arriving MB perceives a delay in receiving service, the caretaker of the MB may decide to forego joining the queue at the ASF altogether, opting instead to seek service elsewhere or abandon the repair entirely. To ensure effective servicing of MBs, the ASF operates in two distinct phases: a primary and a secondary phase. Initially, the mechanic repairs each MB. Upon completion of the repair, there is a probability p that the MB is successfully repaired and leaves the ASF. However, if the initial repair is unsuccessful, the MB transitions to the secondary phase for further repair attempts, which occurs with a probability $\bar{p}$. Notably, due to the expertise of mechanic, the majority of MBs are successfully repaired during the primary phase, with only a small fraction requiring additional attention in the secondary phase for

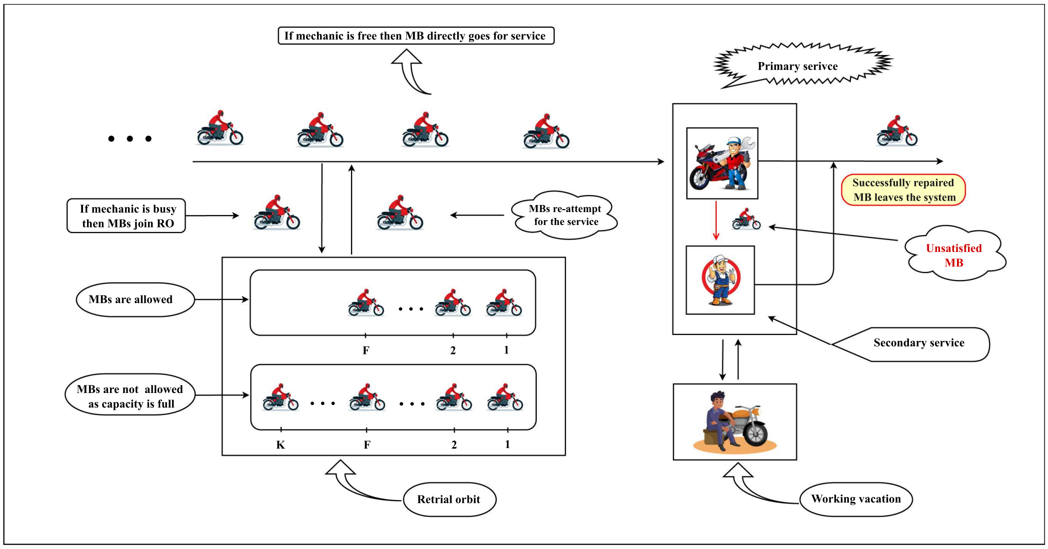


FIGURE 2. Pictorial representation of ASF.

re-repair. In the scenario where no MBs are available for service, the mechanic at the ASF takes a vacation. During this vacation period, if any MBs arrive for service, the mechanic's helpers provide service at a reduced rate compared to the mechanic. This working vacation mode allows the ASF to maintain service availability even when the primary mechanic is on leave, albeit at a slower service rate provided by the auxiliary personnel. Figure 2 shows a pictorial view of ASF.

5. Mathematical analysis of the model

The steady-state difference equations for the various system states ($i = 0, 1, 2, \ldots, 6$) are as follows:

$$(\mu + \lambda q)P_{(1,0)} = \theta P_{(5,0)} + \gamma\, P_{(3,1)} \tag{1}$$

$$(\lambda q + p\mu + \bar{p}\mu)\, P_{(1,n)} = \lambda q\, P_{(1,n-1)} + (n+1)\gamma\, P_{(3,n+1)} + \lambda\, P_{(3,n)} + \theta\, P_{(5,n)}, \qquad n \in [1, K-1], n \neq F-1 \tag{2}$$

$$(\lambda q + p\mu + \bar{p}\mu)P_{(1,F-1)} = \lambda q P_{(1,F-2)} + \lambda P_{(3,F-1)} + \theta P_{(5,F-1)} + \mu P_{(0,F)} + F\gamma P_{(3,F)} \tag{3}$$

$$(\xi + p\mu + \bar{p}\mu)P_{(1,K)} = \lambda q\, P_{(1,K-1)} + \lambda\, P_{(3,K)} + \theta P_{(5,K)} \tag{4}$$

$$\mu_a P_{(2,n)} = \bar{p}\mu P_{(1,n)}, \qquad n \in [1, 2, 3, \ldots, K] \tag{5}$$

$$\lambda q P_{(3,0)} = \mu P_{(1,0)} + p\eta P_{(5,0)} + \bar{p}\eta_a P_{(4,0)} \tag{6}$$

$$(n\gamma + \lambda)P_{(3,n)} = \mu_a P_{(2,n)} + p\mu P_{(1,n)} + p\eta P_{(5,n)} + \eta_a P_{(4,n)}, \qquad n \in [1, 2, 3, \ldots, K] \tag{7}$$

$$\eta_a P_{(4,n)} = \eta \bar{p} P_{(5,n)}, \qquad n \in [0, 1, 2, \ldots, K] \tag{8}$$

$$(\lambda q + p\eta + \bar{p}\eta + \theta)P_{(5,0)} = \lambda q P_{(3,0)} \tag{9}$$

$$(\lambda q + p\eta + \bar{p}\eta + \theta)P_{(5,n)} = \lambda q P_{(5,n-1)}, \qquad n \in [1, 2, \ldots, F-2, F, F+1, \ldots, K-1] \tag{10}$$

$$(\lambda q + p\eta + \bar{p}\eta + \theta)P_{(5,F-1)} = \lambda q P_{(5,F-2)} + \eta P_{(6,F)} \tag{11}$$

$$(\xi_v + p\eta + \bar{p}\eta + \theta)P_{(5,K)} = \lambda q P_{(5,K-1)} \tag{12}$$

$$\eta P_{(6,n)} = \eta P_{(6,n+1)}, \qquad n \in [F, F+1, F+2, \ldots, K-1] \tag{13}$$

$$\eta P_{(6,K)} = \xi_v P_{(5,K)} \tag{14}$$

$$\mu P_{(0,n)} = \mu P_{(0,n+1)}, \qquad n \in [F, F+1, F+2, \dots, K-1] \tag{15}$$

$$\mu P_{(0,K)} = \xi P_{(1,K)}. \tag{16}$$

5.1. Matrix analytic method (MAM)

The matrix analytical method holds significant importance in addressing the steady-state equations of queueing models. It serves as a potent and valuable tool for analyzing and solving complex problems in the domain of queueing theory. This method is frequently employed to derive closed-form solutions from the steady-state equations. For an in-depth understanding of the MAM, one may consult Neuts [29] and Sanga and Antala [12].

We have developed the steady state probabilities $P_{(0,n)}(F \le n \le K), P_{(1,n)}(0 \le n \le K), P_{(2,n)}(1 \le n \le K), P_{(3,n)}(0 \le n \le K), P_{(4,n)}(0 \le n \le K), P_{(5,n)}(0 \le n \le K), P_{(6,n)}(F \le n \le K)$ of this model by this method. The transition rate matrix $\mathbf{J}$ that corresponds to this Markov chain displays a block-tridiagonal pattern:

$$\mathbf{J} = \begin{bmatrix} \mathbf{A}_0 & \mathbf{B}_0 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{C}_1 & \mathbf{A}_1 & \mathbf{B}_1 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_2 & \mathbf{A}_2 & \mathbf{B}_2 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{C}_3 & \mathbf{B}_3 & \mathbf{A}_3 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{C}_4 & \mathbf{A}_4 & \mathbf{B}_4 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{C}_5 & \mathbf{A}_5 & \mathbf{B}_5 \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{C}_6 & \mathbf{A}_6 \end{bmatrix}.$$

The transition rate matrix $\mathbf{J}$ of our state process with working vacation and F-policy shares similarities with the quasi-birth and death category. we know that each and every element within the transition rate matrix $\mathbf{J}$ constitutes a square matrix characterized by a size of $(K+1)$, that follows as:

$$\mathbf{A}_j = \begin{bmatrix} P_{j,0} & Q_{j,0} & 0 & 0 & \cdots & 0 & 0 & 0 \\ R_{j,1} & P_{j,1} & Q_{j,1} & 0 & \cdots & 0 & 0 & 0 \\ 0 & R_{j,2} & P_{j,2} & Q_{j,2} & \cdots & 0 & 0 & 0 \\ 0 & 0 & R_{j,3} & P_{j,3} & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & P_{j,K-2} & Q_{j,K-2} & 0 \\ 0 & 0 & 0 & 0 & \cdots & R_{j,K-1} & P_{j,K-1} & Q_{j,K-1} \\ 0 & 0 & 0 & 0 & \cdots & 0 & R_{j,K} & P_{j,K} \end{bmatrix} \quad \text{for} \quad j \in [0,6].$$

It is necessary to define an indicator function $\Lambda_{[x,y]}(i)$, before representing the elements of $A_j(i \in [0,K])$.

$$\Lambda_{[x,y]} = \begin{cases} 1, & \text{if } x \le i \le y \\ 0, & \text{otherwise.} \end{cases}$$

Consequently, we can represent $P_{j,i}$, $Q_{j,i}$ and $R_{j,i}$ in the following manner:

$$P_{j,i} = \begin{cases} \phi\Lambda_{[0,F-1]}(i) + \mu\Lambda_{[F,K]}(i), & \text{if } j=0,\ i \in [0,K], \\ (\mu+\lambda q)\Lambda_{[0,0]}(i) + (\bar{p}\mu + p\mu + \lambda q)\Lambda_{[1,K-1]}(i) + (\bar{p}\mu + p\mu + \xi)\Lambda_{[K,K]}(i), & \text{if } j=1,\ i \in [0,K], \\ \phi\Lambda_{[0,0]}(i) + \mu_e\Lambda_{[1,K]}(i), & \text{if } j=2,\ i \in [0,K], \\ \lambda q\Lambda_{[0,0]}(i) + (\lambda + K\gamma)\Lambda_{[i,K]}(i), & \text{if } j=3,\ i \in [0,K], \\ \eta_a\Lambda_{[1,K]}(i), & \text{if } j=4,\ i \in [0,K], \\ (\lambda q + p\eta + \bar{p}\eta + \theta)\Lambda_{[0,K-1]}(i) + (p\eta + \bar{p}\eta + \theta + \xi_v)\Lambda_{[K,K]}, & \text{if } j=5,\ i \in [0,K], \\ \phi\Lambda_{[0,K]}(i) + \eta\Lambda_{[F,K]}(i), & \text{if } j=6,\ i \in [0,K] \end{cases}$$

$$Q_{j,i}=\begin{cases}0, & \text{if } j=0,2,3,4,6, \quad i\in[0,K-1],\\ -\lambda q\Lambda_{[0,K]}(i), & \text{if } j=1,5, \quad i\in[0,K-1]\end{cases}$$

and

$$R_{j,i}=\begin{cases}0, & \text{if } j\in[1,5], \quad i\in[0,K],\\ -\mu\Lambda_{[F+1,K]}(i), & \text{if } j=0, \quad i\in[1,K],\\ -\eta\Lambda_{[F+1,K]}(i), & \text{if } j=6, \quad i\in[1,K]\end{cases}$$

$$\mathbf{B}_j=\begin{bmatrix} b_{j,0} & 0 & 0 & \cdots & 0 & 0 & 0\\ d_{j,1} & b_{j,1} & 0 & \cdots & 0 & 0 & 0\\ 0 & d_{j,2} & b_{j,2} & \cdots & 0 & 0 & 0\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ 0 & 0 & 0 & \cdots & d_{j,K-1} & b_{j,K-1} & 0\\ 0 & 0 & 0 & \cdots & 0 & d_{j,K} & b_{j,K}\end{bmatrix}, \quad j\in[0,5]$$

are $(K+1)$ square matrices, for $i\in[0,K]$.

$$b_{j,i}=\begin{cases}\phi\Lambda_{[0,K]}(i), & \text{if } j=0,2\\ \phi\Lambda_{[0,K-1]}(i)-\xi\Lambda_{[K,K]}(i), & \text{if } j=1,\\ \phi\Lambda_{[0,0]}(i)-\lambda\Lambda_{[1,K]}(i), & \text{if } j=3,\\ -\eta_e\Lambda_{[1,K]}(i), & \text{if } j=4,\\ -(p\eta+\bar{p}\eta+\theta)\Lambda_{[1,K]}(i), & \text{if } j=5\end{cases}$$

$$d_{j,i}=\begin{cases}-(i)\gamma, & \text{if } j=3, \quad i\in[1,K],\\ 0, & \text{otherwise}\end{cases}$$

$$\mathbf{C}_j=\begin{bmatrix} c_{j,0} & 0 & 0 & \cdots & 0 & 0\\ 0 & c_{j,1} & 0 & \cdots & 0 & 0\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\ 0 & 0 & 0 & \cdots & c_{j,K-1} & 0\\ 0 & 0 & 0 & \cdots & 0 & c_{j,K}\end{bmatrix} \quad j\in[1,6]$$

are $(K+1)$ square matrices and

$$c_{j,i}=\begin{cases}-\mu\Lambda_{[0,0]}(i)-(p\mu+\bar{p}\mu)\Lambda_{[1,K]}(i), & \text{if } j=1,\\ \phi\Lambda_{[0,0]}(i)-\mu_a\Lambda_{[1,K]}(i), & \text{if } j=2,\\ -\lambda q\Lambda_{[0,0]}(i)\phi\Lambda[1,K](i), & \text{if } j=3,\\ \phi\Lambda_{[0,K]}(i), & \text{if } j=4,\\ \phi\Lambda_{[0,K-1]}(i)-(\xi_v)\Lambda_{[K,K]}(i), & \text{if } j=5,\\ \phi\Lambda_{[0,K]}(i), & \text{if } j=6.\end{cases}$$

Note: ϕ is an element that belongs to the set of non-zero real numbers within the given expressions.

Let us consider Π to represent a steady state probability with the corresponding vector of $\mathbf{j}$. By representing the vector Π as $\Pi=(\Pi_0,\Pi_1,\Pi_2,\Pi_3,\Pi_4,\Pi_5,\Pi_6)$ where $\Pi_j=(P_{j,0},P_{j,1},P_{j,2},\ldots,P_{j,k-1},P_{j,k})$ forms a row vector of size $1\times(K+1)$ for values of $j=0,1,2,3,4,5,6$. Here, $P_{(j,n)}$ when $0\le n\le K$ and $(F\le n\le K)$ represent the steady state probabilities from our model, while the probabilities $P_{2(,0)}$, $P_{(0,n)}$ and $P_{(6,n)}$ are set to zero. Upon solving the steady-state equations $\Pi J = 0$, we obtain the following results,

$$\Pi_0 A_0 + \Pi_1 C_1 = 0 \tag{17}$$
$$\Pi_0 B_0 + \Pi_1 A_1 + \Pi_2 C_2 = 0 \tag{18}$$
$$\Pi_1 B_1 + \Pi_2 A_2 + \Pi_3 C_3 = 0 \tag{19}$$
$$\Pi_2 B_2 + \Pi_3 A_3 + \Pi_4 C_4 = 0 \tag{20}$$
$$\Pi_3 B_3 + \Pi_4 A_4 + \Pi_5 C_5 = 0 \tag{21}$$
$$\Pi_4 B_4 + \Pi_5 A_5 + \Pi_6 C_6 = 0 \tag{22}$$
$$\Pi_5 B_5 + \Pi_6 A_6 = 0. \tag{23}$$

Now on simplification, we get

$$\Pi_6 = -\Pi_5 B_5 {A_6}^{-1} \tag{24}$$
$$\Pi_5 = -\Pi_4 B_4 [A_5 - B_5 A_6^{-1} C_6]^{-1} \tag{25}$$
$$\Pi_4 = -\Pi_3 B_3 [A_4 - B_4 (A_5 - B_5 A_6^{-1}) C_5]^{-1} \tag{26}$$
$$\Pi_3 = -\Pi_2 B_2 [A_3 - B_3 (A_4 - B_4 (A_5 - B_5 A_6^{-1} C_5)^{-1}) C_4]^{-1} \tag{27}$$
$$\Pi_2 = -\Pi_1 B_1 [A_2 - B_2 [A_3 - B_3 (A_4 - B_4 (A_5 - B_5 A_6^{-1} C_5)^{-1}) C_4]^{-1} C_3]^{-1} \tag{28}$$
$$\Pi_1 = -\Pi_0 B_0 [A_1 - B_1 [A_2 - B_2 [A_3 - B_3 (A_4 - B_4 (A_5 - B_5 A_6^{-1} C_5)^{-1}) C_4]^{-1} C_3]^{-1} C_2]^{-1} \tag{29}$$
$$\Pi_0 (A_1 + \Psi_1 C_1) = 0. \tag{30}$$

By setting a value $\Psi_j = -B_{j-1} \times \, [A_j + \Psi_{j+1} C_{j+1}]^{-1}$ $(j = 1, 2, 3, 4, \ldots)$, we can simplify $\Pi_1, \Pi_2, \Pi_3, \Pi_4$ and Π_5 to $\Pi_1 = \Pi_0 \Psi_1, \Pi_2 = \Pi_1 \Psi_2, \Pi_3 = \Pi_2 \Psi_3, \Pi_4 = \Pi_3 \Psi_4, \Pi_5 = \Pi_4 \Psi_5$ respectively. Additionally, we observed that $\Psi_6 = -B_5 A_6^{-1}$. Equation (30) determines Π_0 up to a multiplicative constant, while equations (24)–(29) determines $\Pi_1, \Pi_2, \Pi_3, \Pi_4, \Pi_5$ and Π_6 up to the same constant. The constant can be uniquely determined by the following normalization equation

$$\sum_{j=0}^{6} (\Pi_j) e = 1. \tag{31}$$

An optimized computer program was developed to solve for Π_j and $P_{j,n} (0 \leq j \leq 6)$, utilizing the column vector e in which every element is set to one.

6. System metrics

In examining the proposed model, various system metrics have been formulated. These metrics are developed to assist system organizers in the decision-making process. The important system metrics include the expected number of customers in the system (L_S), the anticipated waiting time ((W_T)) for customers before the commencement of their service, throughput (TP), and the long-run probabilities associated with different system states.

(1) Expected number of customers in the system:

$$L_S = L_S(N) + L_S(\mathrm{WV}) \tag{32}$$

where

– $L_S(N) = \sum_{n=1}^{K} n P_{(3,n)} + \sum_{n=1}^{K} n P_{(2,n)} + \sum_{n=1}^{K} n P_{(1,n)} + \sum_{n=F}^{K} n P_{(0,n)};$

– $L_S(\mathrm{WV}) = \sum_{n=F}^{K} nP_{(6,n)} + \sum_{n=1}^{K} nP_{(5,n)} + \sum_{n=1}^{K} nP_{(4,n)}$.

(2) The system's total expected waiting is indicated by W_T.

$$W_T = W_T(N) + W_T(\mathrm{WV}) \tag{33}$$

where
– Waiting time of system during normal service $W_T(N) = \frac{L_S(N)}{\lambda_{\text{eff}}(N)*(1-P_{\text{LN}})}$;
– Waiting service of system during working vacation $W_T(\mathrm{WV}) = \frac{L_S(\mathrm{WV})}{\lambda_{\text{eff}}(\mathrm{WV})*(1-P_{\text{LWV}})}$;
– Effective arrival rate during normal service

$$\lambda_{\text{eff}}(N) = \left(\lambda q \sum_{n=0}^{K-1} P_{(1,n)} + \lambda \sum_{n=1}^{K} P_{(3,n)}\right)(1-P_{\text{BN}})(1-P_{\text{LN}})(1-q)(1-p);$$

– Effective arrival during working vacation

$$\lambda_{\text{eff}}(\mathrm{WV}) = \left(\lambda q \sum_{n=0}^{K-1} P_{(5,n)}\right)(1-P_{\text{BWV}})(1-P_{\text{LWV}})(1-q)(1-p).$$

(3) The system's throughput is indicated by TP.

$$\mathrm{TP} = \eta \sum_{n=F}^{K} P_{(6,n)} + (p\eta + \bar{p}\eta) \sum_{n=0}^{K} P_{(5,n)} + \eta_a \sum_{n=0}^{K} P_{(4,n)} + \mu_a \sum_{n=0}^{K} P_{(2,n)} + (p\mu + \bar{p}\mu) \sum_{n=0}^{K} P_{(1,n)} + \mu \sum_{n=F}^{K} P_{(0,n)}. \tag{34}$$

(4) Long run probabilities:
– Loss probability during normal service $P_{\text{LN}} = \frac{\rho_N^K(1-\rho_N)}{1-\rho_N^{K+1}}$;
– Loss probability of working vacation $P_{\text{LWV}} = \frac{\rho_{\text{WV}}^K(1-\rho_{\text{WV}})}{1-\rho_{\text{WV}}^{K+1}}$;
– Balking probability during normal service $P_{\text{BN}} = \frac{1-\rho_N^K}{1-\rho_N^{K+1}}$;
– Balking probability during working vacation $P_{\text{BWV}} = \frac{1-\rho_{\text{WV}}^K}{1-\rho_{\text{WV}}^{K+1}}$;
– The traffic intensity during normal service is $\rho_N = \frac{\lambda \cdot q}{\mu \cdot p \cdot K}$;
– The traffic intensity of working vacation is $\rho_{\text{WV}} = \frac{\lambda \cdot q}{\eta \cdot p \cdot K}$;
– Busy state probability of the server is indicated by P_B.

$$P_B = \sum_{n=F}^{K} P_{(6,n)} + \sum_{n=1}^{K} P_{(5,n)} + \sum_{n=1}^{K} P_{(4,n)} + \sum_{n=1}^{K} P_{(2,n)} + \sum_{n=1}^{K} P_{(1,n)} + \sum_{n=F}^{K} P_{(0,n)}. \tag{35}$$

– Free state probability of the server is indicated by P_F.

$$P_F = 1 - P_B. \tag{36}$$

– Working vacation state probability of the server is indicated by P_{WV}.

$$P_{\text{WV}} = \sum_{n=F}^{K} P_{(6,n)} + \sum_{n=0}^{K} P_{(5,n)} + \sum_{n=0}^{K} P_{(4,n)} + P_{(3,0)}. \tag{37}$$

6.1. Cost function

The cost associated with the proposed queueing model constitutes a crucial consideration, exerting a significant impact on the overall revenue of the company. The ability to maintain lower costs while optimizing services serves as an attractant for an increased influx of customers into the system. To minimize the cost of the system and maximize its profitability, we formulate a cost function per unit of time that correlates with the service rates μ and μ_a as decision variables.

$$\mathrm{TC}(\mu,\mu_a) = C_1 L_S + C_2 p\mu + C_3 \bar{p}\mu_a + C_4 p\eta + C_5 \bar{p}\eta_a + C_6\theta + C_F\mu + C_{(\mathrm{WV})}\eta \tag{38}$$

$$\min\ \mathrm{TC}(\mu,\mu_a) = \mathrm{TC}(\mu^*,\mu_a^*) \tag{39}$$

where cost components per unit of time are as follows:

C_1	The holding cost for each customer waiting in the system.
C_2	Cost incurred during normal service in the system.
C_3	Cost incurred when the server provides secondary service during normal operation.
C_4	Cost incurred by customers during primary service in the WV period.
C_5	Cost incurred when the server provides secondary service during the WV period.
C_6	Cost incurred when the server is free during normal state.
C_F	Cost incurred during admission control under the F-policy in normal buys state.
C_{WV}	Cost incurred during admission control under the F-policy in WV period.

7. Numerical results

To evaluate the effectiveness of the proposed model, we present a numerical example illustrating the impact of arrival rate (λ), service rate (μ), retrial rate (γ), threshold parameter (F), and system capacity (K) on various system metrics. Additionally, employing the ANFIS offers a valuable means of validating the numerical results. For computational purposes, we consider the following input values: $K = 5$, $F = 3$, $\lambda = 3$, $\mu = 5$, $\mu_a = 3$, $\eta = 3$, $\eta_a = 2$, $\xi = 1$, $\xi_v = 1$, $p = 0.7$, $\bar{p} = 0.3$, $q = 0.9$, $\theta = 2$.

- **Effect of service rate μ:** the service rate is an essential factor in addressing system congestion. A higher service rate allows customers to be processed more quickly, resulting in shorter queue lengths under constant customer arrival rates. we systematically varied the system capacity (K), threshold parameter (F), and service rate (μ) as outlined in Table 3. Increasing the service rate while keeping other parameters constant led to a reduced system size and waiting time for customers, accompanied by increased throughput. This indicates that a higher service rate enhances system efficiency by reducing customer waiting times and increasing the overall throughput of the queueing system. Furthermore, the probabilities associated with server status exhibit distinct trends: the probability of the server being on a WV and the probability of the server being free both increased with higher service rates, while the probability of the server being busy showed an inverse relationship, decreasing as the service rate increases. We also investigated the impact of system capacity (K) and threshold parameter F on the developed $M/M/1/K/WV$ retrial queueing model performance, as reported in Table 3. We observe that increasing the system capacity and admission control threshold value (F) resulted in higher numbers of customers (L_S), consistent with increased waiting times (W_T). This trend suggests that larger system capacities and higher threshold values may lead to greater congestion and longer waiting times for customers within the queueing system. The graphical representation of numerical results for different values of μ is illustrated in Figure 4. This figure demonstrates the impact of customer joining probability (q) and the probability of customers entering the first-phase service (p) on system length (L_S) and waiting time (W_T). By varying μ, we analyze how these probabilities influence queueing system performance metrics, providing insights into the relationship between customer behavior and system dynamics. Figure 4 serves as a visual tool to understand the effects of key parameters on queueing system operations.

TABLE 3. Performance measures by changing the parameter μ.

(K,F)	μ	L_S	W_T	TP	P_{WV}	P_B	P_F
(5, 3)	3	2.793	0.097	2.328	0.066	0.740	0.260
	5	1.868	0.044	2.688	0.195	0.553	0.447
	7	1.372	0.019	2.759	0.305	0.442	0.558
(10, 7)	3	7.229	0.346	2.389	5.56e−16	0.796	0.204
	5	5.942	0.284	2.794	4.73e−15	0.643	0.357
	7	5.203	0.219	2.890	6.13e−16	0.539	0.461
(15, 12)	3	12.229	0.623	2.389	1.64e−15	0.796	0.204
	5	10.942	0.585	2.794	2.95e−13	0.643	0.357
	7	10.204	0.513	2.889	6.02e−16	0.539	0.461

– **Effect of arrival rate λ:** Table 4 illustrates the influence of customer arrival rates on various system metrics, including L_S, W_T, TP, and system state probabilities. The corresponding graphical representation of these effects is depicted in Figure 3. When the arrival rate of customers (λ) increases, the system experiences congestion, resulting in higher waiting times and increased server busyness, as indicated in Table 4. System capacity (K) and threshold value (F) also significantly impact these system metrics. Figure 3 further demonstrates the impact of λ on L_S and W_T across different values of p and q. The graphical representation in Figure 3 reveals that higher customer joining probabilities (q) contribute to increased system congestion. In contrast, a higher proportion of satisfied customers in the first-phase service (p) leads to reduced system length and waiting time. These analyses provide valuable insights into how changes in customer arrival rates impact queueing system performance, enabling a deeper understanding of system behavior under varying workload conditions.
– **Effect of retrial rate γ:** the impact of the retrial rate of customers is demonstrated through numerical results presented in both tabular form (Tab. 5) and graphical representation (Fig. 5). When customers reattempt for service at a higher rate, the system appears less congested, and waiting times decrease. Conversely, increasing system capacity and threshold values while keeping the retrial rate constant results in inverse behavior of the metrics, as reported in Table 5. Minor changes are observed in throughput with varying retrial rates, and a slight variation is also noted in system state probabilities. Figure 5 illustrates the variation of the γ on L_S and W_T for different values of p (0.5, 0.7, 0.9) and q (0.5, 0.7, 0.9). It is evident from the figure that the system length increases with higher values of q and decreases with higher values of p.

7.1. ANFIS computing

This section highlights the application of ANFIS in the proposed model, aimed at validating the numerical results derived from the analytical method discussed earlier. ANFIS leverages the combined capabilities of neural networks and fuzzy logic to model complex systems effectively. A detailed overview of ANFIS is provided in Section 2.3, illustrating its role in enhancing the accuracy and reliability of the analytical findings presented in this study. To obtain results *via* ANFIS (predicted results), we use one input parameter and three membership functions corresponding to linguistic variables: small, medium, and large. The *evalfis* function in MATLAB software is employed for this purpose. To assess the accuracy of the results, we use several metrics, including root mean square error (RMSE), mean absolute error (MAE), coefficient of determination (R^2), and absolute percentage error (APE), with their mathematical expressions given in equations (40)–(44).

$$\mathrm{MAE} = \frac{1}{M}\sum_{z=1}^{M}|s(z) - r(z)| \tag{40}$$

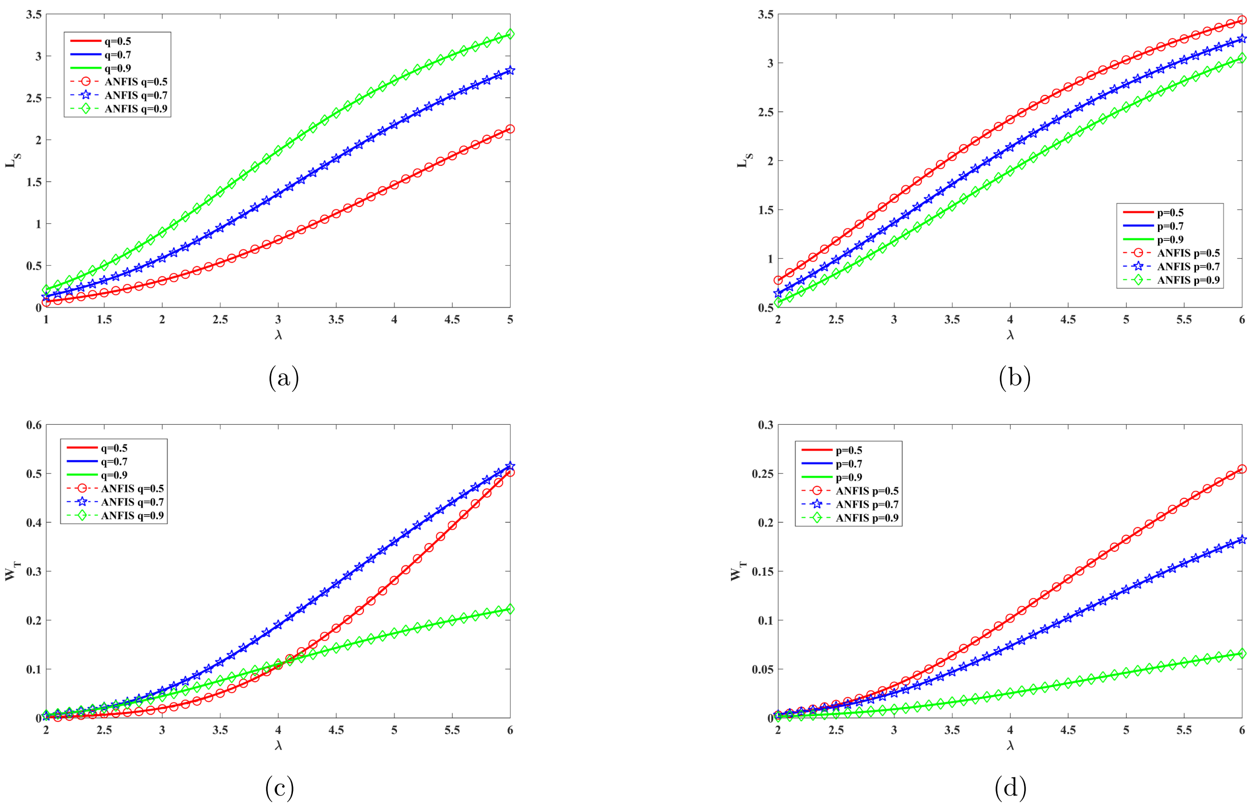


FIGURE 3. Impact of varying λ on L_S and W_T for different combinations of p and q. (a) L_S *vs.* λ for various values of q. (b) L_S *vs.* λ for various values of p. (c) W_T *vs.* λ for various values of q. (d) W_T *vs.* λ for various values of p.

$$\text{RMSE} = \sqrt{\frac{1}{M}\sum_{z=1}^{M}(s(z) - r(z))^2} \tag{41}$$

$$R^2 = \frac{\displaystyle\sum_{x=1}^{N}(r(z) - s(\bar{z}))^2}{\displaystyle\sum_{x=1}^{M}(s(z) - s(\bar{z}))^2} \times 100 \tag{42}$$

$$\text{APE} = \frac{|s(z) - r(z)|}{s(z)} \times 100 \tag{43}$$

$$\text{Accuracy } (\%) = 100 - \frac{\text{APE}}{M} \tag{44}$$

where $s(z)$ is the actual value obtained by analytical formulae, $r(z)$ is the predicted value using ANFIS for a particular data set z, and M represents the overall count of data sets used for prediction and $s(\bar{z})$ is the mean of actual data.

By using input parameters such as (i) μ, (ii) λ, and (iii) γ, along with data from Tables 3 to 5, we derive results for L_S and W_T for $p = 0.5, 0.7, 0.9$ and $q = 0.5, 0.7, 0.9$ combinations using the *evalfis* function in

TABLE 4. Impact of λ on system metrics.

(K,F)	λ	L_S	W_T	TP	P_{WV}	P_B	P_F
$(5,3)$	1	0.215	0.0001	1.011	0.797	0.092	0.908
	3	1.868	0.0445	2.688	0.195	0.553	0.447
	5	3.255	0.1731	3.407	0.028	0.764	0.236
$(10,7)$	1	4.168	0.025	1.163	3.82e−13	0.268	0.732
	3	5.942	0.283	2.794	4.73e−15	0.643	0.357
	5	7.748	0.541	3.486	2.42e−16	0.793	0.207
$(15,12)$	1	9.168	0.115	1.161	2.62e−09	0.268	0.732
	3	10.942	0.585	2.794	2.95e−13	0.643	0.357
	5	12.748	0.946	3.486	2.08e−14	0.793	0.207

TABLE 5. Performance measures by changing the parameter γ.

(K,F)	γ	L_S	W_T	TP	P_{WV}	P_B	P_F
$(5,3)$	0.5	3.075	0.087	2.570	0.025	0.578	0.422
	1.5	2.106	0.054	2.680	0.144	0.567	0.433
	2.5	1.704	0.038	2.690	0.237	0.540	0.460
$(10,7)$	0.5	7.501	0.342	2.625	2.62e−16	0.599	0.401
	1.5	6.247	0.296	2.772	8.95e−16	0.637	0.363
	2.5	5.734	0.275	2.807	3.10e−16	0.646	0.354
$(15,12)$	0.5	12.501	0.638	2.625	1.29e−11	0.600	0.400
	1.5	11.247	0.597	2.772	6.12e−13	0.637	0.363
	2.5	10.734	0.576	2.807	2.08e−12	0.646	0.354

MATLAB. Subsequently, we compute several metrics outlined in equations (40)–(44). The values of these metrics are tabulated in Tables 6–8, and the predicted values (marked with ticks) are presented graphically in Figures 3–5.

From Tables 6 to 8, we observe that the MAE, RMSE, and APE values are nearly zero, indicating a high level of accuracy in the predictions. Further, the coefficient of determination (R^2) values are close to one, suggesting an excellent fit of the model to the data. The accuracy achieved is over 99%, demonstrating the robustness and reliability of the ANFIS model in predicting system metrics. These metrics collectively affirm the effectiveness and precision of the ANFIS approach in this context. In Figures 4 and 5, the continuous line represents the actual results, while the tick-marked symbols ($\circ, \star, \diamond$) depict the values predicted by ANFIS. It is evident from these figures that the actual values closely overlap with the predicted values by ANFIS, indicating a strong alignment between the model's predictions and the observed outcomes.

8. Cost optimization

Customers always seek faster service to minimize their waiting time in queues, which can increase system costs and reduce overall revenue. Therefore, the objective of the system organizer is to optimize service rates to minimize system costs and maximize revenue. The system can achieve a balance that ensures efficient customer service while maximizing profitability by providing service at optimal rates. This objective highlights the critical role of strategic decision-making in effectively managing queueing systems to achieve optimal performance and ensure customer satisfaction. We formulate a nonlinear cost function with the decision variable as the service rate (μ and μ_a), as given in equation (38). This highly nonlinear and complex cost function incorporates various system metrics and probability distributions. To obtain optimal values of the decision variables and minimize

TABLE 6. Error of L_S and W_T generated through ANFIS by varying of μ.

	p	MAE	RMSE	R^2	APE	Accuracy
L_S	0.5	0.0029	0.0033	99.9919	0.1739	99.8262
	0.7	0.0033	0.0038	99.9868	0.2323	99.7677
	0.9	0.0035	0.0041	99.9811	0.2787	99.7213
W_T	0.5	0.00028	0.00032	99.98531	0.98667	99.01333
	0.7	0.00018	0.00021	99.98377	0.74426	99.25577
	0.9	0.00006	0.00007	99.97796	0.72468	99.27532
	q	MAE	RMSE	R^2	APE	Accuracy
L_S	0.5	0.00466	0.00540	99.97456	0.47716	99.52284
	0.7	0.00344	0.00390	99.99230	0.22398	99.77602
	0.9	0.00212	0.00239	99.99713	0.10595	99.89405
W_T	0.5	0.00089	0.00106	99.93940	3.37549	96.62450
	0.7	0.00044	0.00049	99.99272	0.73635	99.26365
	0.9	0.00008	0.00010	99.99836	0.19062	99.80937

TABLE 7. Error of L_S and W_T generated through ANFIS by varying of λ.

	p	MAE	RMSE	R^2	APE	Accuracy
L_S	0.5	0.00253	0.00295	99.99868	0.11467	99.88533
	0.7	0.00185	0.00216	99.99926	0.09667	99.90333
	0.9	0.00141	0.00162	99.99953	0.09257	99.90743
W_T	0.5	0.00029	0.00036	99.99799	1.64991	98.35009
	0.7	0.00021	0.00026	99.99792	1.32433	98.67567
	0.9	0.00008	0.00010	99.99747	1.36418	98.63582
	q	MAE	RMSE	R^2	APE	Accuracy
L_S	0.5	0.00293	0.00341	99.99720	1.05878	98.94122
	0.7	0.00243	0.00292	99.99884	0.50698	99.49302
	0.9	0.00275	0.00315	99.99896	0.30025	99.69975
W_T	0.5	0.00134	0.00150	99.99090	10.97433	89.02567
	0.7	0.00066	0.00079	99.99764	2.25052	97.74948
	0.9	0.00020	0.00025	99.99866	0.76968	99.23032

the system cost, we employ iterative methods such as the QNM and metaheuristic-based GA. These approaches enable us to navigate the complex cost landscape and identify optimal solutions efficiently. For a detailed description of QNM and GA, please refer to Sections 2.1 and 2.2, respectively.

To compute the desired results, we set the cost elements as follows: $C_1 = 50, C_2 = 20, C_3 = 30, C_4 = 25, C_5 = 10, C_6 = 10, C_F = 30, C_{WF} = 30$, and the queueing input parameters as $\lambda = 1, \theta = 2, \eta = 3, \eta_a = 2, \xi = 1, \xi_v = 1, p = 0.7, \bar{p} = 1 - p, q = 0.9$. We also investigate the impact of system capacity, threshold value, and retrial rate on the minimum cost of the system. This analysis aims to understand how variations in these parameters affect the overall cost optimization and performance of the system.

Quasi-Newton method: Refer to Section 2.1 for the algorithmic steps of the QNM. To initiate the QNM, we select the initial values for the service rates as $\mu_0 = 0.5$ and $\mu_{a0} = 0.1$ with a retrial rate $\gamma = 2$, and system parameters $(K, F) = (10, 8)$. We proceed with iterative computations, and after six iterations, we achieve convergence as the stopping criterion is met *i.e.*, $\max\left(\left|\frac{\partial \mathrm{TC}}{\partial \mu}\right|, \left|\frac{\partial \mathrm{TC}}{\partial \mu_a}\right|\right) < 10^{-6}$. The optimal values for (μ) and (μ_a) are determined to be $(\mu^* = 1.75)$ and $(\mu_a^* = 0.40)$, resulting in a minimum system cost $(\mathrm{TC}(\mu^*, \mu_a^*) = 550.5103)$.

TABLE 8. Error of L_S and W_T generated through ANFIS by varying of γ.

	p	MAE	RMSE	R^2	APE	Accuracy
L_S	0.5	0.00347	0.00415	99.95998	0.25440	99.74560
	0.7	0.00341	0.00410	99.95394	0.30026	99.69974
	0.9	0.00324	0.00391	99.94974	0.33512	99.66488
W_T	0.5	0.00012	0.00015	99.94452	0.50392	99.49608
	0.7	0.00012	0.00014	99.93065	0.63034	99.36966
	0.9	0.00004	0.00005	99.91827	0.72991	99.27009
	q	MAE	RMSE	R^2	APE	Accuracy
L_S	0.5	0.00308	0.00375	99.93144	0.49094	99.50906
	0.7	0.00366	0.00439	99.95395	0.33158	99.66842
	0.9	0.00348	0.00413	99.96768	0.21913	99.78087
W_T	0.5	0.00016	0.00019	99.84183	1.10108	98.89892
	0.7	0.00028	0.00035	99.92508	0.72866	99.27134
	0.9	0.00013	0.00016	99.96250	0.40229	99.59771

TABLE 9. Optimal values obtained using QNM and GA.

(K, F)	γ	Method	(μ^*, μ_a^*)	$\mathrm{TC}(\mu^*, \mu_a^*)$
$(10, 8)$	2	QNM	(1.750, 0.400)	550.51
		GA	(1.729, 0.418)	550.49
	3	QNM	(1.700, 0.450)	544.38
		GA	(1.685, 0.437)	544.37
$(15, 12)$	2	QNM	(1.850, 0.450)	754.09
		GA	(1.857, 0.471)	754.09
	3	QNM	(1.800, 0.500)	747.65
		GA	(1.800, 0.487)	747.65
$(20, 16)$	2	QNM	(1.950, 0.550)	956.12
		GA	(1.937, 0.499)	956.12
	3	QNM	(1.850, 0.500)	949.55
		GA	(1.872, 0.513)	949.59

The same process is repeated for $\gamma = 3$ and system configurations $(K, F) = \{(10, 8), (20, 16)\}$. The optimal values of the service rates and corresponding costs are presented in Table 9.

Genetic algorithm: Refer to Section 2.2 for the overview of GA. We use the tournament selection method to choose the fittest chromosome from the population. This method involves randomly selecting a subset of chromosomes and choosing the best individual (with the highest fitness) from that subset as a parent for the next generation. For crossover, we adopt the two-point crossover method, which involves selecting two points along the parents' chromosomes to exchange segments and create new offspring (solutions). This process helps explore different combinations of genetic material. Further, we employ the bit inversion method for mutation, where a random bit in the chromosome is flipped to introduce genetic diversity. We set the number of generations as the stopping criterion, ensuring that the GA iterates sufficiently to converge towards an optimal solution. To obtain the desired optimal solution for service rates (μ^*, μ_a^*) and minimum cost $\mathrm{TC}(\mu^*, \mu_a^*)$, we set the parameters as follows:

– Population size: 20;
– Crossover probability: 1;
– Mutation rate: 0.08;

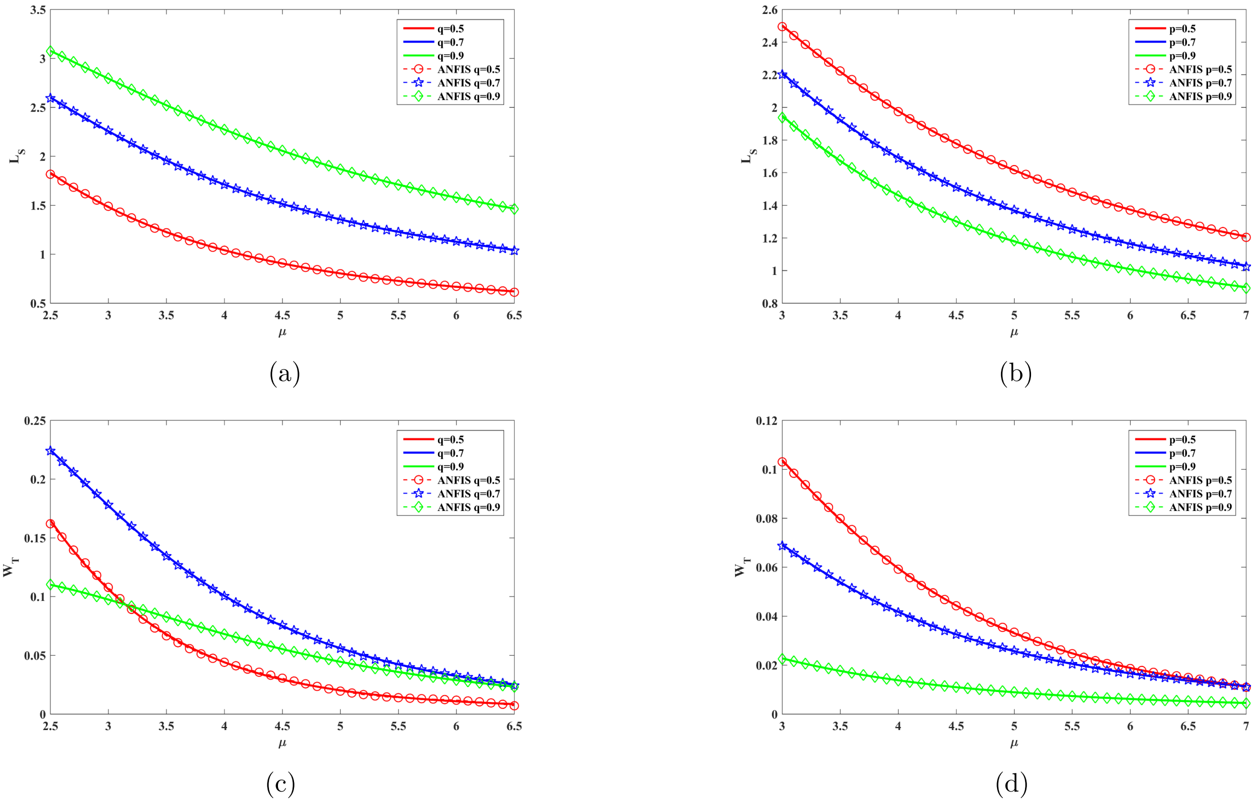


FIGURE 4. Impact of varying μ on L_S and W_T for different combinations of p and q. (a) L_S *vs.* μ for various values of q. (b) L_S *vs.* μ for various values of p. (c) W_T *vs.* μ for various values of q. (d) W_T *vs.* μ for various values of p.

– Stopping criterion (number of generations): 50.

To identify the optimal solution using the GA, we applied the specified GA parameters and queueing input parameters. Specifically, for the system configuration of $(K, F) = (10, 8)$ and $\gamma = 2$, we conducted a search in the range of $\mu = [0.1, 5]$ and $\mu_a = [0.1, 5]$. Following a series of iterations (generations), the GA converges after 50 iterations. The convergence plot of GA is shown in Figure 6a. The resulting optimal solution is recorded and is presented in the second row of Table 9. To assess the impact of system parameters K, F, and γ, we repeat the GA process and recorded the optimal values of μ^*, μ_a^* and $\mathrm{TC}(\mu^*, \mu_a^*)$ in Table 9. The convergence of the GA is illustrated in Figures 6b and 6c for the system configurations $(K, F) = (15, 12)$ and $(20, 16)$, respectively. These figures represent how the genetic algorithm progresses over iterations to converge towards optimal solutions. This process demonstrates how the GA efficiently explores the solution space to determine the μ^* and μ_a^* that minimize the total cost $\mathrm{TC}(\mu^*, \mu_a^*)$ of the system in the specified constraints. The cost function exhibits convexity, as illustrated in Figure 7.

The Table 9 is structured to demonstrate the performance of both QNM and GA across different system configurations and retrial rates. For each combination of (K, F) and γ, the optimal service rates and corresponding total costs are reported, allowing for a comparative analysis of the effectiveness of the optimization methods under varying conditions. As customers retry for service faster, leading to a less congested retrial orbit, the system cost decreases. Conversely, increasing the system's capacity and the admission control threshold parameter causes the system cost to rise due to higher customer numbers.

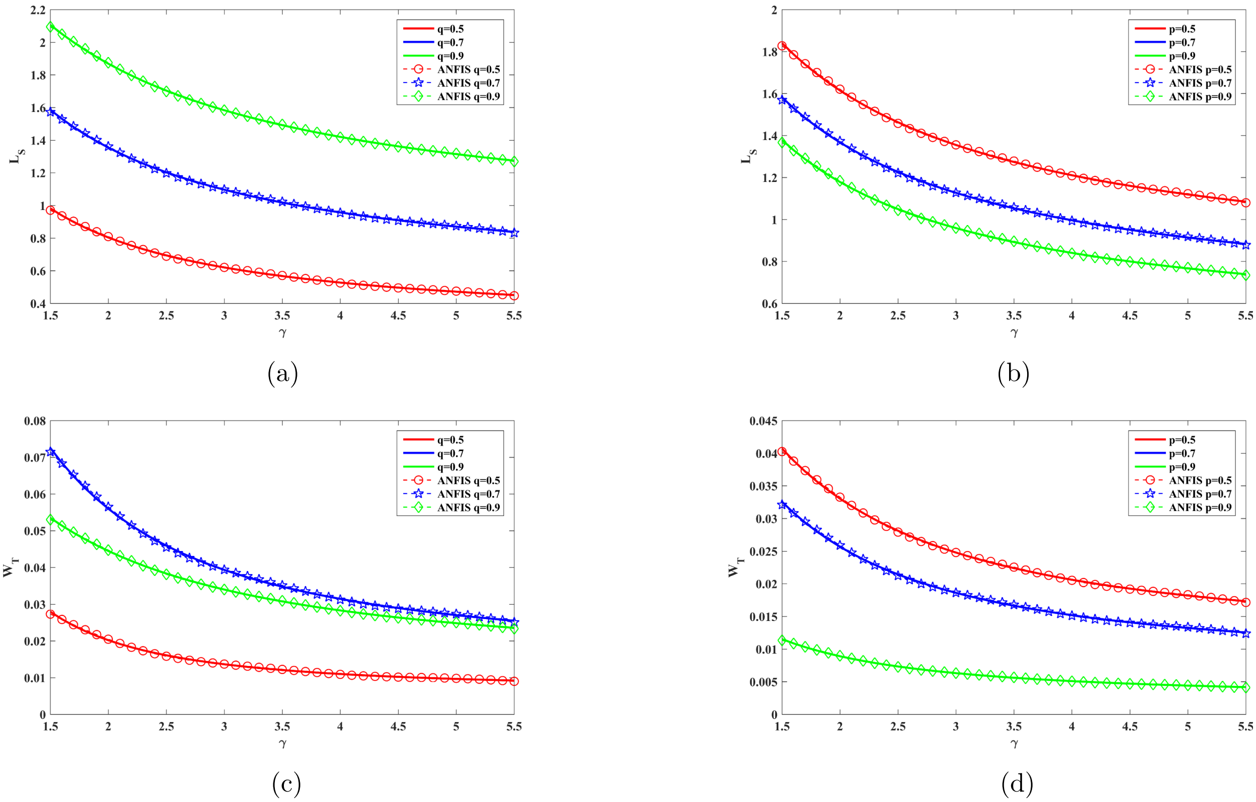
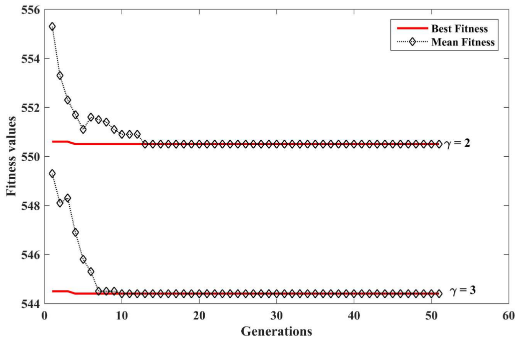


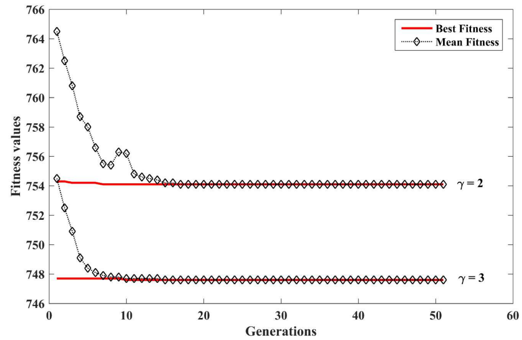
FIGURE 5. Impact of varying γ on L_S and W_T for different combinations of p and q. (a) L_S *vs.* γ for various values of q. (b) L_S *vs.* γ for various values of p. (c) W_T *vs.* γ for various values of q. (d) W_T *vs.* γ for various values of p.

9. Conclusions

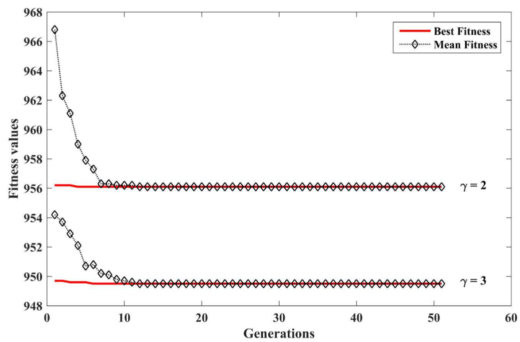
The queueing management of the $M/M/1/K/WV$ retrial queue, incorporating the F-policy along with dual-phase service and working vacation, has been effectively explored. The application of these strategies to customers in the model has contributed to a more realistic and comprehensive development of the queueing system. By implementing the F-policy, working vacation, and dual-phase service, the model reflects practical considerations and operational scenarios encountered in real-world queueing environments, enhancing the relevance and applicability of the proposed queueing model. The mathematical analysis of the model involved formulating steady-state difference equations for various system states and solving them using a matrix analytic method to derive probability distributions. This rigorous approach enabled the development of key system metrics, such as system size, customer waiting time, throughput, long-run probabilities, and other performance indicators. These system metrics offer valuable insights into the behavior and performance of the queueing system. By utilizing probability distributions and system metrics, one can gain a deeper understanding of how the proposed model functions and performs, facilitating informed decision-making. In addition to the mathematical analysis, we implemented the ANFIS technique to validate the numerical results obtained for system length and customer waiting time. By utilizing the ANFIS technique, we validated the accuracy and reliability of our numerical results. The implementation of ANFIS provides a robust validation mechanism of the performance metric of the $M/M/1/K/WV$ queueing model. Through this approach, we enhance the credibility and applicability of our findings, reinforcing the effectiveness of our proposed queueing model in practical scenarios. Moreover, we



(a)



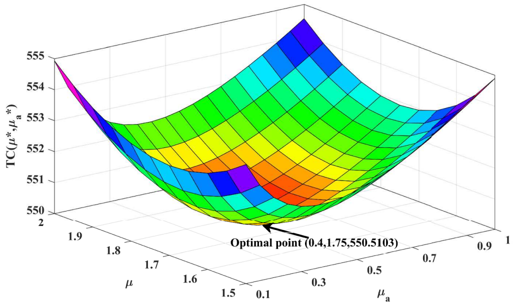
(b)



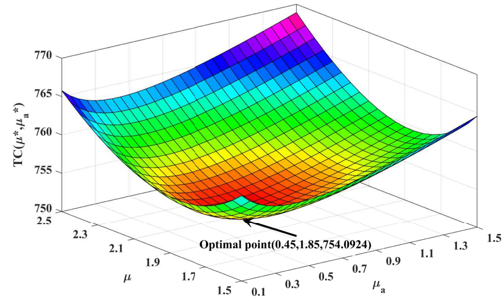
(c)

FIGURE 6. TC *vs.* Generation for evaluation of μ and μ_a by varying the values of γ. (a) $K = 10, F = 8$. (b) $K = 15, F = 12$. (c) $K = 20, F = 16$.

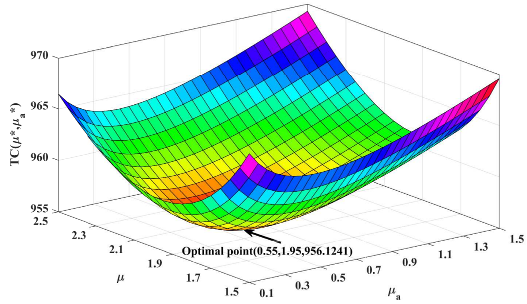
extended our study by formulating a cost function with service rates as decision variables and applied optimization methods such as QNM and GA to minimize the cost function. The investigated queueing model holds significant potential across diverse domains, particularly in ASF. We conducted a detailed examination of how the model applies specifically to ASF settings, where MBs arrive for service. This application highlights several important real-world dynamics. For example, ASF often faces peak demand during certain times, leading to congestion and increased waiting times. The retrial mechanism built into the model captures scenarios where MBs, unable to receive immediate attention, leave the facility but return later for service. During off-peak periods, technicians may reduce their service rate rather than halting work altogether, reflecting situations where minor repairs or inspections are carried out. This aligns with real-world practices in ASF, where less urgent tasks are performed during slow periods to maintain overall efficiency. This detailed application underscores the practical relevance and utility of our queueing model in real-world service environments. Future research of this study could incorporate multiple vacation situations for service providers, addressing customer reneging behaviors in the system, and exploring the impact of multiple servers on system performance. These enhancements would contribute to developing a more comprehensive and practical queueing model that addresses real-world complexities and challenges effectively.



(a)



(b)



(c)

FIGURE 7. Total cost varying with the combination of μ and μ_a. (a) $K = 10, F = 8$. (b) $K = 15, F = 12$. (c) $K = 20, F = 16$.

Acknowledgments

We thank the editors and anonymous referees for the constructive suggestions and feedback that helped us a lot in improving the subject matters as well as the presentation of the paper. The first author, Sudeep Singh Sanga, expresses gratitude for receiving funding from the Science and Engineering Research Board (SERB), DST, Government of India, through the Startup Research Grant scheme (SRG/2023/001642).

Data availability statement

The research data associated with this article are included in the article.

References

[1] J.G. Templeton and G.I. Falin, Retrial Queues. Taylor & Francis (2023).

[2] G. Falin, A survey of retrial queues. *Queue. Syst.* **7** (1990) 127–167.

[3] J.R. Artalejo and A. Gómez-Corral, Retrial queueing systems. *Math. Comput. Modell.* **30** (1999) xiii–xv.

[4] S.M. Gupta, Interrelationship between controlling arrival and service in queueing systems. *Comput. Oper. Res.* **22** (1995) 1005–1014.

[5] M. Jain and S.S. Sanga, State dependent queueing models under admission control F-policy: a survey. *J. Ambient Intell. Human. Comput.* **11** (2020) 3873–3891.

[6] M. Jain and S.S. Sanga, F-policy for $M/M/1/K$ retrial queueing model with state-dependent rates, in Performance Prediction and Analytics of Fuzzy, Reliability and Queuing Models: Theory and Applications. Springer (2019) 127–138.

[7] S.S. Sanga and G.S. Charan, Fuzzy modeling and cost optimization for machine repair problem with retrial under admission control F-policy and feedback. *Math. Comput. Simul.* **211** (2023) 214–240.
[8] C.-H. Wu, D.-Y. Yang and C.-R. Yong, Performance evaluation and bi-objective optimization for F-policy queue with alternating service rates. *J. Ind. Manage. Optim.* **19** (2023) 1–21.
[9] F.A. Haight, Queueing with balking. *Biometrika* **44** (1957) 360–369.
[10] M. Jain, S. Dhibar and S.S. Sanga, Markovian working vacation queue with imperfect service, balking and retrial. *J. Ambient Intell. Human. Comput.* (2022) 1–17.
[11] M. Jain and S.S. Sanga, Optimal control F-policy for $M/M/R/K$ queue with an additional server and balking. *Int. J. Appl. Comput. Math.* **5** (2019) 1–16.
[12] S.S. Sanga and K.S. Antala, Performance analysis and cost investigations for state-dependent single unreliable server finite queue under F-policy using GA and QNM. *J. Comput. Appl. Math.* **441** (2024) 115679.
[13] V. Chandrasekaran, K. Indhira, M. Saravanarajan and P. Rajadurai, A survey on working vacation queueing models. *Int. J. Pure Appl. Math.* **106** (2016) 33–41.
[14] L.D. Servi and S.G. Finn, $M/M/1$ queues with working vacations $M/M/1/WV$. *Perform. Eval.* **50** (2002) 41–52.
[15] D.-Y. Yang, K.-H. Wang and C.-H. Wu, Optimization and sensitivity analysis of controlling arrivals in the queueing system with single working vacation. *J. Comput. Appl. Math.* **234** (2010) 545–556.
[16] A.A. Bouchentouf, M. Boualem, L. Yahiaoui and H. Ahmad, A multi-station unreliable machine model with working vacation policy and customers' impatience. *Qual. Technol. Quant. Manage.* **19** (2022) 766–796.
[17] S.S. Sanga and K.S. Antala, Cost optimization and machine learning based prediction of waiting time for fault-tolerance machining system with general vacation and F-policy. *Appl. Soft Comput.* **154** (2024) 111404.
[18] P.K. Chahal and K. Kumar, A genetic algorithm for cost optimization in queueing models of machining systems with multiple working vacation and generalized triadic policy. *Decis. Anal. J.* **10** (2024) 100395.
[19] D. Agarwal, R. Agarwal and S. Upadhyaya, Detection of optimal working vacation service rate for retrial priority g-queue with immediate bernoulli feedback. *Results Control Optim.* **14** (2024) 100397.
[20] M. Jain and S.S. Sanga, Admission control for a finite capacity queueing model with general retrial times and state-dependent rates. *J. Ind. Manage. Optim.* **16** (2020).
[21] M. Jain and S.S. Sanga, Unreliable single server double orbit retrial queue with balking. *Proc. Nat. Acad. Sci. India Sect. A Phys. Sci.* **91** (2021) 257–268.
[22] J.H. Holland, Genetic algorithms. *Sci. Am.* **267** (1992) 44–50.
[23] S.S. Sanga and M. Jain, $FM/FM/1$ double orbit retrial queue with customers' joining strategy: a parametric nonlinear programing approach. *Appl. Math. Comput.* **362** (2019) 124542.
[24] S.S. Rao, Engineering Optimization: Theory and Practice. John Wiley & Sons (2019).
[25] J.-S. Jang, Adaptive-network-based fuzzy inference system ANFIS. *IEEE Trans. Syst. Man Cybern.* **23** (1993) 665–685.
[26] A.I. Dounis, G. Leftheriotis, S. Stavrinidis and G. Syrrokostas, Electrochromic device modeling using an adaptive neuro-fuzzy inference system: a model-free approach. *Energy Buildings* **110** (2016) 182–194.
[27] S.S. Sanga and M. Jain, Cost optimization and ANFIS computing for admission control of $M/M/1/K$ queue with general retrial times and discouragement. *Appl. Math. Comput.* **363** (2019) 124624.
[28] K. Divya and K. Indhira, Performance analysis and anfis computing of a markovian queuing model with intermittently accessible server under a hybrid vacation policy. *RAIRO: Oper. Res.* **58** (2024) 1257–1279.
[29] M.F. Neuts, Matrix-analytic methods in queuing theory. *Eur. J. Oper. Res.* **15** (1984) 2–12.