Original Research

Bulk queue model with multiphase service and repair with balking and N policy

Binay Kumar[1*]

Received: 7 August 2025 / Accepted: 16 October 2025 / Published online: 22 January 2026

*** Corresponding Author Email, bkmathasr@gmail.com**

1- Department of Mathematics, Magadh Mahila College, Patna University, Patna, India

Abstract
This paper investigates a single-server queueing system in which customers arrive in batches at a constant rate. Each customer undergoes a mandatory first-stage service, followed by an optional second-stage service. The server is subject to breakdowns that can occur at any point during either service stage. A phased repair mechanism, consisting of both essential and optional repairs, is incorporated into the model. Upon completion of the essential repair, the server proceeds to the second repair phase with probability q_1. Likewise, after the $(j-1)th$ phase $(j = 2,3,..k)$, it enters the jth phase with probability q_{j-1}; otherwise, it leaves the repair system to resume service. In this manner, the server may undergo up to k repair phases, including the initial essential phase. Further, it is assumed that customers are impatient in nature may balk from the system if server is not available on their arrival. To analyze the system's steady-state behavior, the study utilizes probabilistic reasoning alongside the supplementary variable technique. Key performance metrics are derived using the generating function method, and their validity is demonstrated through numerical examples.
.

Keywords - Bulk, Phase repair, Balking, N-policy, Phase service

INTRODUCTION

Stochastic queueing models, developed under a variety of assumptions, are extensively employed in domains such as cloud computing, logistics, automated warehousing, and data networks to analyze system behavior and evaluate performance measures. A major challenge in these systems is the unpredictable failure of service units-such as overloaded servers in cloud platforms, malfunctioning robots in warehouses, or faulty routers in communication networks—that can disrupt operations and impede the flow of tasks or data. In practice, continuous service availability cannot be guaranteed, as breakdowns frequently arise from hardware malfunctions, software errors, or environmental factors. Recovery is further complicated when repair personnel or replacement components are not immediately accessible, resulting in prolonged downtime and diminished efficiency.

Classical queueing models generally assume that once a server enters the repair state, restoration is completed in a single uninterrupted phase. However, real-world repair processes are often more complex, requiring multiple optional phases, where subsequent stages are undertaken depending on the server's condition or operational requirements. For instance, in industrial maintenance, an initial diagnostic and basic repair may be followed, if necessary, by component replacement, calibration, or system testing, constituting an additional repair phase. Incorporating such multi-phase repair structures into queueing models

provides a more realistic framework for analyzing system dynamics, particularly in environments where the complexity and variability of repair tasks significantly affect overall performance.

In both service-oriented and industrial queueing environments, customer impatience, typically characterized by behaviors such as reneging or balking, plays a critical role in determining key system performance measures, including average queue length, customer waiting time, and overall system throughput. Neglecting such behavior may lead to overly optimistic and unrealistic assessments of service quality, resource utilization, and system efficiency. For instance, in a telecommunications call center, callers who are subjected to prolonged delays often abandon the call before being served, thereby contributing to service loss, reduced revenue, and diminished customer satisfaction. To more accurately capture system dynamics, it is essential to incorporate customer impatience while developing queueing models. Incorporating such mechanisms allows for a more comprehensive evaluation of the complex trade-offs among staffing levels, service capacity, and customer retention. Consequently, the inclusion of impatience as a fundamental modeling parameter is essential for producing analytically robust and practically relevant insights in the design, analysis, and optimization of modern service systems.

Review of Literature

In contemporary communication and information processing systems, service interruptions constitute a critical operational challenge, often resulting in significant economic losses and reduced customer satisfaction. To address these issues, a substantial body of research has focused on the development of queueing models that explicitly account for service interruptions arising from server failures under diverse operating assumptions.

Early contributions in this area include the works of Sengupta [1], Takine and Sengupta [2], Madan [3], Ke [4], Pearn et al. [5], Ke and Lin [6], Ke [7], Kumar and Arumuganathan [8], and Jain and Agarwal [9]. Building upon these foundations, Choudhury and Tadj [10] analyzed a bulk queueing model with an unreliable server operating under an N-policy, where customers arrive in batches of random size, service is delivered in two sequential phases, and server vacations follow a Bernoulli schedule. Wu and Lian [11] extended this line of research by studying a single-arrival queueing system with both positive and negative arrivals, priority queues, and a retrial mechanism under a Bernoulli vacation schedule, where negative arrivals induce server breakdowns. Singh et al. [12] introduced a bulk-service retrial model with server unreliability, incorporating Bernoulli-type vacation policies and multiple optional services supplementing the essential service. Rajadurai et al. [13] considered a retrial queue with negative arrivals and working vacations, wherein the server initiates a working vacation with probability p when idle, though the vacation may be interrupted by the arrival of a negative customer.

Further developments include the work of Ayyappan and Karpagam [14], who examined an unreliable N-policy-based queueing system with standby service, customer loss, and feedback, where multiple Bernoulli-type vacations are permitted and a standby server is activated during failures of the main server. Ayyappan et al. [15] studied a bulk-service model with two-phase service and standby redundancy, where the standby server is deployed during repair following startup failures. Kumar and Jain [16] applied the matrix-geometric method to analyze an unreliable Markovian queueing model with two-stage service and a hybrid vacation policy. Li and Liu [17] also employed the matrix-geometric approach to study a general system with multi-phase service and working vacations under Bernoulli interruptions. Bharathi and Nandhini [18] used the supplementary variable technique to investigate an unreliable queueing system with compulsory and optional services, where the server takes a vacation with probability p once the orbit becomes empty. Most recently, Ayyappan and Gurulakshmi [19] provided a comprehensive analysis of a Markovian arrival system with multiple vacation types, N-policy-based interruptions, optional service phases, breakdown–repair mechanisms, setup delays, and customer discouragement. In a related study, Kumar [20] examined an unreliable model with delayed repair and a single vacation policy, in which the server delivers service through multiple optional phases.

The impatient behavior of customers represents a critical factor influencing the performance of queueing systems. Recognizing its significance, numerous researchers have incorporated customer impatience into the formulation of queueing models under varying operational assumptions. Notable early contributions in this domain include the works of Wang and Ke [21], Jain et al. [22], Movaghar [23], Xiong and Altiok [24], Chakravarthy [25], and Arrar et al. [26]. In related work, Singh et al. [27] examined a finite-capacity Markovian queueing system with impatient customers operating under two randomly changing environments, each characterized by distinct service rates. Extending this line of research, Yang and Wu [28] analyzed a finite-capacity Markovian queue with impatient customers, introducing the concept of working breakdowns, whereby the server may fail and undergo repair during the service of a customer at a relatively low rate. Morozov et al. [29] employed the matrix-analytic method to study a multiclass retrial queue with impatient customers, where arriving customers either join a corresponding orbit or balk from the system. More recently, Bharathi and Nandhini [30] investigated a single-server model

with a modified Bernoulli vacation schedule and customer balking, under the assumption that the server provides two categories of service: essential and non-essential.

PRACTICAL JUSTIFICATION OF MODEL

In a cloud-based data processing environment, user jobs arrive according to a renewal process and are stored in a common buffer until the scheduler initiates service under an N-policy. Each job undergoes a multiphase service structure: an essential computation phase involving tasks such as data ingestion, preprocessing, and distributed execution, followed by the user's discretion, with a probabilistic optional phase that may include aggregation, visualization, or report generation.

The computing facility is assumed to be unreliable, and failures may occur during any service phase. To address such interruptions, the repair mechanism is designed as a multistage process: initially, a rapid soft-recovery attempt (e.g., restart of the virtual machine or container) is performed; if unsuccessful, an optional migration phase is invoked to reallocate the task to a different host; and, if instability persists, a more comprehensive repair stage (such as image rebuilding or hardware reassignment) is executed before service resumes from the point of interruption.

The activation of the server is governed by the N-policy: service commences only when the number of queued jobs reaches the threshold N, thereby reducing overhead associated with frequent activation. However, this control mechanism introduces additional waiting time, during which customers exhibit impatience. Each job is characterized by a finite patience time, often modelled as a random variable, and any job whose waiting time exceeds its patience threshold abandons the system prior to service initiation.

This integrated framework thus captures the interaction between batch activation (via N-policy), multiphase service with optional stages, multistage repair under unreliability, and customer impatience, making it well-suited for the performance evaluation of modern cloud computing platforms.

The wide applicability of phase-type service and repair mechanisms, together with bulk arrivals and impatient customer behavior, motivates the present study of a queueing model with balking, which incorporates bulk arrivals, an optional service component in addition to the essential one, a multi-phase repair process, and a control mechanism governed by the N-policy. The remainder of this paper is organized as follows. Section 2 presents a detailed description of the model along with the definitions of various terms employed. In Section 3, the governing equations of the system are formulated. Section 4 is devoted to the mathematical analysis of the model based on these equations. In Section 5, key performance indices are derived. Finally, Section 6 provides numerical illustrations and sensitivity analysis to highlight the applicability and robustness of the proposed model.

MODEL DESCRIPTION

In many industrial scenarios, the main objective of any manufacturing/ production process is to produce quality products with low cost and minimum time. To produce optimal results in various congestion situations, including the field of digital communication systems, manufacturing/production systems, etc., where the service may be rendered in two phases. In the present investigation, we assume that there is a provision to go for the immediate repair on an unpredictable breakdown of the server, and the repair may be done in different phases. The repair may be delayed due to the unavailability of the repairman or any other reasons, but it may affect the smooth functioning of the system. The flow of customers during repairs may be influenced by the server status. Keeping in view the above-mentioned situation, we consider a queueing system under $N-$policy wherein the customers join the system according to a Poisson process in batches of random size X with $a_j = P(X = j)$, first and second factorial moment $E(X)$ and $E(X^{(2)})$ respectively. The customer joins the system with Poisson arrival rate λ. The customer may not join the queue with probability $\bar{\varepsilon} = (1 - \varepsilon)$ if the server is busy or under repair. There is a provision of two stages of services. First one is regular service A_1, follows general probability law with distribution function $\mathcal{A}_1(u)$, Laplace transform $\mathcal{A}_1^*(s)$ and finite moments $a_1^{(j)}, j = 1,2$ respectively, while the second stage service $\mathcal{A}_2$ is optional, having general probability law with distribution function $\mathcal{A}_2(u)$, Laplace transform $\mathcal{A}_2^*(s)$ and finite moments $a_2^{(j)}, j = 1,2$ respectively. After availing of the regular service, the arrived units may join the optional service with probability p p or leave the system with probability $1 - p$. It is assumed that the breakdowns of the server may occur during the first stage regular service or second stage optional service with rates $\mathcal{s}_1$ and $\mathcal{s}_2$ respectively. It is assumed that as the failure occurs, it immediately joins the repair facility. The failed server joins the repair station, wherein k phases of repair are available, in which the first phase of repair is essential, while the remaining are optional. When the essential repair of a server is completed, the server may

join the repair system for the second phase of repair with probability q_1. Similarly, after completing the second phase of optional repair, the repairman immediately starts the subsequent third phase of repair with probability q_2; otherwise, the server may leave the repair system and start to provide the service. On a similar pattern, the server may require a maximum of k phases of repair, including the first phase essential repair, with probabilities $q_{j-1}(j = 2,3,..k)$ for moving from$(j-1)th$ phase to jth phase of repair. The random variable $\mathcal{R}_j^{(i)}$ denotes the j^{th} phase repair time of the server with distribution function $\mathcal{G}_{i,j}(y)$, Laplace transform $\mathcal{G}_{i,j}^*(s)$ and finite moments $g_{ij}^{(t)}, t = 1,2$, respectively, when its fails in i^{th} phase of service.

To analyze the present non-Markovian model, we introduce supplementary variables corresponding to elapsed service time, elapsed delay time, and elapsed repair time. Let $\mathcal{N}_q(t)$ denote the number of units in the system, including one being in service. Let $w_i(t)$ denote the elapsed service time of the customer for $i^{th}(i = 1,2)$ the phase of service at time t, and $\psi_{i,j}(t), j = 1,2,..,k$ denote the elapsed repair time of the server when a server breakdown occurs in $i^{th}(i = 1,2)$ phase of service at time t.

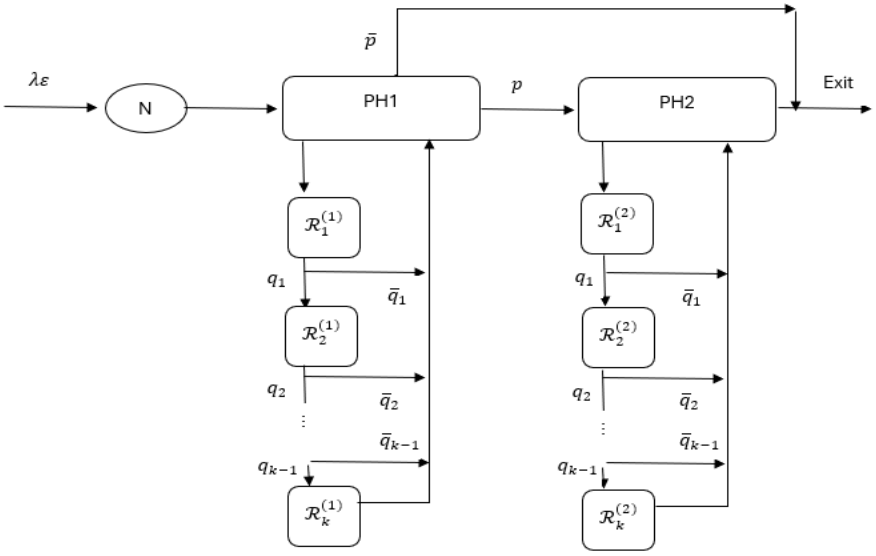

FIGURE 1
SYSTEMATIC REPRESENTATION OF MODEL

To define the server's state, we introduce the random variables $\psi(t)$ defined as

$$\psi(t) = \begin{cases} 0, & \text{if the server is idle at time } t. \\ 1, & \text{if the server is offering first phase service at time } t. \\ 2, & \text{if the server is offering second optional phase service at time } t. \\ 2+j, & \text{if the server is under jth repair when it failed essential service at time } t. \\ 2+k+j, & \text{if the server is under jth phase of repair when it failed in optional service at time } t. \end{cases}$$

For analysis purposes, we consider the bivariate Markov process,$\{\mathcal{N}_q(t), (t)\}$ Xt, where $\mathcal{X}(t)$ takes values $0, w_1(t), w_2(t), \psi_{1,1}(t), \psi_{1,2}(t), \ldots, \psi_{1,k}(t), \psi_{2,1}(t), \psi_{2,2}(t), \ldots, \psi_{2,k}(t)$; if $\zeta(t) = 0,1,2, \ldots, 2+k, 2+k+1, \ldots, 2+2k$ respectively.

The limiting probabilities for system states are defined as

$I_n^{(0)} = \lim_{t\to\infty} \Pr.\{\mathcal{N}_q(t) = 0, \mathcal{X}(t) = 0\}$ for $n = 0,1, \ldots, N-1$

$W_n^{(i)}(u)du = \lim_{t\to\infty} \Pr.\{\mathcal{N}_q(t) = n, \mathcal{X}(t) = w_i(t); u < w_i(t) \le u + du\}; n \ge 1, u > 0, i = 1,2.$

$L_{j,n}^{(i)}(u,v)dv = \lim_{t\to\infty} \Pr.\{\mathcal{N}_q(t) = n, \mathcal{X}(t) = \psi_{i,j}(t); v < \psi_{i,j}(t) \le v + dv \,/\, w_i(t) = u\};$

$n \ge 1, (u,v) > 0, i \in \{1,2\}, j = 1,2, \ldots, k.$

Also, we assume that

$\mathcal{A}_i(0)=0, \mathcal{A}_i(\infty)=1, \mathcal{G}_{i,j}(0)=0, \mathcal{G}_{i,j}(\infty)=1; i=1,2$. We further assume that $\mathcal{A}_i(u)$ is continuous at $u=0$, $\mathcal{G}_{i,j}(v)$ are continuous at $v=0$.

The hazard rate functions for $i^{th}(i=1,2)$ phase service, delay time while failed during i^{th} phase service, and $j^{th}(1\le j\le k)$ phase repair times are defined as follows

$$\varphi_i(u)du=\frac{d\mathcal{A}_i(u)}{[1-\mathcal{A}_i(u)]}; i=1,2.$$

$$\varsigma_{i,j}(v)dv=\frac{d\mathcal{G}_{i,j}(v)}{[1-\mathcal{G}_{i,j}(y)]}; i=1,2 \text{ and } j=1,2,\dots,k.$$

Further, we define the following probability generating functions

$W^{(i)}(u,z)=\sum_{n=1}^{\infty} z^n W_n^{(i)}(u);\ \ W^{(i)}(0,z)=\sum_{n=1}^{\infty} z^n W_n^{(i)}(0)$

$L_j^{(i)}(u,v,z)=\sum_{n=1}^{\infty} z^n L_{j,n}^{(i)}(u,v)$;$L_j^{(i)}(u,0,z)=\sum_{n=1}^{\infty} z^n L_{j,n}^{(i)}(u,0)$ Here $j=1,2,\dots,k$.

$$I_N^{(0)}(z)=\sum_{n=0}^{N-1} z^n I_n^{(0)}$$

Let us further define $\sigma_n(n=0,1,2,\dots,N-1)$ as the probability that a batch of customers finds at least n one customer in the system during the idle period. Thus σ_n satisfies the following recursive relation.

$\sigma_0=1, \sigma_n=\sum_{k=1}^{n} c_k\sigma_{n-k}. 1\le n\le N-1$ (1)

GOVERNING EQUATIONS

In this section, we construct the Kolmogorov steady state equations governing the system states (cf. Cox, 1955; Choudhury et al., 2009) by using the probability reasoning as follows:

$$\frac{d}{du}W_n^{(i)}(u)+[\lambda\varepsilon+\beta_i+\varphi_i(u)]W_n^{(i)}(u)=\sum_{t=1}^{n}\lambda\varepsilon a_t\, W_{n-t}^{(i)}(x)+\int_0^{\infty}\varsigma_{i,k}(v)L_{k,n}^{(i)}(u,v)dv$$

$$+\sum_{j=1}^{k-1}\bar{q}_j\int_0^{\infty}\varsigma_{i,j}(v)L_{j,n}^{(i)}(u,v)dv,\ i=1,2. \quad (2)$$

$$\frac{d}{dv}L_{j,n}^{(i)}(u,v)+\left[\lambda\varepsilon+\varsigma_{i,j}(v)\right]L_{j,n}^{(i)}(u,v)=\sum_{t=1}^{n}\lambda\varepsilon a_t\, L_{j,n-t}^{(i)}(u,v), 1\le j\le k \text{ and } i=1,2. \quad (3)$$

$$\lambda I_0^{(0)}=\int_0^{\infty}\varphi_2(u)\,W_1^{(2)}(u)du+\bar{p}\int_0^{\infty}\varphi_1(u)\,W_1^{(1)}(u)du \quad (4)$$

$$\lambda I_n^{(0)}=\lambda\sum_{j=1}^{n}a_j I_{n-j}^{(0)},\ \ n=0,1,2,\dots,N-1 \quad (5)$$

Equations (2) through (5) are to be solved subject to the boundary conditions specified at the following point u= 0:

$$W_n^{(1)}(0)=\int_0^{\infty}\varphi_2(u)\,W_{n+1}^{(2)}(u)du+\bar{p}\int_0^{\infty}\varphi_1(u)\,W_{n+1}^{(1)}(u)du,\quad 1\le n\le N-1 \quad (6)$$

$$W_n^{(1)}(0)=\int_0^{\infty}\varphi_2(u)\,W_{n+1}^{(2)}(u)du+\bar{p}\int_0^{\infty}\varphi_1(u)\,W_{n+1}^{(1)}(u)du+\lambda\sum_{j=1}^{n}a_{n-j}I_j^{(0)}, n\ge N \quad (7)$$

$$W_n^{(2)}(0)=p\int_0^{\infty}\varphi_1(u)\,W_n^{(1)}(u)du,\quad n\ge 1 \quad (8)$$

And the boundary condition at $v=0$ for fixed value of u (for $i=1,2$), we have

$$L_{1,n}^{(i)}(u,0)=\beta_i W_n^{(i)}(u), u>0, n\ge 1 \quad (9)$$

$$L_{j,n}^{(i)}(u,0)=q_{j-1}\int_0^{\infty}\varsigma_{i,j-1}(v)L_{j-1,n}^{(i)}(u,v)dv,\quad 2\le j\le k, n\ge 1 \quad (10)$$

The normalizing condition is

$$\sum_{n=0}^{N-1}I_n^{(0)}+\sum_{i=1}^{2}\sum_{n=1}^{\infty}\left\{\int_0^{\infty}W_n^{(i)}(u)du+\int_0^{\infty}\int_0^{\infty}\sum_{j=1}^{k}L_{j,n}^{(i)}(u,v)dudv\right\}=1 \quad (11)$$

ANALYSIS

To simplify the analysis and avoid unnecessary complexity, we introduce the following additional notations, which will be used throughout the subsequent discussion.

$\int_0^\infty e^{-sx}(1-M(x))dx = \frac{1-M^*(s)}{s}$, where $M^*(s)$ is LST of $M(x)$. $C(z) = \lambda(1-a(z))$,

$\delta_i(z) = \varepsilon C(z) + \beta_i\left[1 - \left\{\mathcal{G}_{i,1}^*(\varepsilon C(z)) + \sum_{t=2}^{k}\prod_{j=1}^{t-1} q_j \mathcal{G}_{i,j}^*(\varepsilon C(z))\left(\mathcal{G}_{i,t}^*(\varepsilon C(z)) - 1\right)\right\}\right]; i = 1,2.$

$\mathfrak{h}_i^{(1)} = \lambda\varepsilon E(X)\left[1 + \beta_i(\mathcal{g}_{i1}^{(1)} + \sum_{t=2}^{k}(\prod_{j=1}^{t-1} q_t)\mathcal{g}_{it}^{(1)}\right]; i = 1,2$

$$\mathfrak{h}_i^{(2)} = \lambda\varepsilon E(X^{(2)}) + \beta_i\left[2(\lambda\varepsilon E(X))^2\sum_{t=2}^{k}\left(\prod_{j=1}^{t-1} q_t\right)\mathcal{g}_{ij}^{(1)}\mathcal{g}_{it}^{(1)} + [\lambda\varepsilon E(X^{(2)})\mathcal{g}_{i1}^{(1)} + (\lambda\varepsilon E(X))^2\mathcal{g}_{i1}^{(2)}] + \sum_{t=2}^{k}\left(\prod_{j=1}^{t-1} q_t\right)[\lambda\varepsilon E(X^{(2)})\mathcal{g}_{it}^{(1)} + (\lambda\varepsilon E(X))^2\mathcal{g}_{it}^{(2)}]\right]$$

$$\rho_1 = \frac{a_1^{(1)}\mathfrak{h}_1^{(1)} + p\,a_2^{(1)}\mathfrak{h}_2^{(1)}}{\varepsilon}$$

$\rho_2 = a_1^{(1)}\mathfrak{h}_1^{(1)} + p\,a_2^{(1)}\mathfrak{h}_2^{(1)} = \varepsilon\rho_1$ and $\rho = \frac{\rho_1}{1-\rho_2+\rho_1}$

Solving the equations (3) in the usual manner, we have

$$L_j^{(i)}(u,v,z) = L_j^{(i)}(u,0,z)\,exp\{-\varepsilon C(z)v\}[1-\mathcal{G}_{i,j}(v)];\ \ i = 1,2; 1 \le j \le k \tag{12}$$

From equations (9) and (10), we have

$$L_1^{(i)}(u,0,z) = \beta_i W^{(i)}(u,z) \tag{13}$$

$$L_j^{(i)}(u,0,z) = \int_0^\infty \varsigma_{i,j-1}(v)L_{j-1}^{(i)}(u,v,z)dv,\ 2 \le j \le k \tag{14}$$

From equations (12) and (13), we obtain

$$L_j^{(i)}(u,0,z) = q_{j-1}L_{j-1}^{(i)}(u,0,z)\mathcal{G}_{i,j-1}^*(\varepsilon C(z)), 2 \le j \le k \tag{15}$$

On simplification, we have

$$L_j^{(i)}(u,0,z) = L_1^{(i)}(u,0,z)\prod_{t=1}^{j-1} q_t\mathcal{G}_{i,t}^*(\varepsilon C(z)), 2 \le j \le k \tag{16}$$

Solving equations (2) and using (12) and (13), we get

$$W^{(i)}(u,z) = W^{(i)}(0,z)[1-\mathcal{A}_i(u)]\,exp\{-\delta_i(z)u\} \tag{17}$$

By multiplying equations (6) and (7) by appropriate powers of z, summing over all possible values of the variable, and simplifying, we obtain

$$zW^{(1)}(0,z) = -zC(z)I_N^{(0)}(z) + \bar{p}W^{(1)}(0,z)\mathcal{A}_1^*(\delta_1(z)) + W^{(2)}(0,z)\mathcal{A}_1^*(\delta_2(z)) \tag{18}$$

Similarly, from equation (8), we get

$$W^{(2)}(0,z) = pW^{(1)}(0,z)\mathcal{A}_1^*(\delta_1(z)) \tag{19}$$

Using equations (18) and (19), we get

$$W^{(1)}(0,z) = \frac{zC(z)I_N^{(0)}(z)}{[(\bar{p} + p\mathcal{A}_2^*(\delta_2(z))\mathcal{A}_1^*(\delta_1(z)) - z]} \tag{20}$$

From equations (12)-(17), we get

$$L_1^{(i)}(u,v,z) = \beta_i W^{(i)}(0,z)[1-\mathcal{A}_i(u)]\,exp\{-\delta_i(z)u\}[1-\mathcal{G}_{i,j}(v)]\,exp\{-\varepsilon C(z)v\} \quad (21)$$

$$L_j^{(i)}(u,v,z) = \beta_i W^{(i)}(0,z)[1-\mathcal{A}_i(u)]\,exp\{-\delta_i(z)u\}\prod_{t=1}^{j-1} q_t \mathcal{G}_{i,t}^*\big(\varepsilon C(z)\big)\,exp\{-\varepsilon C(z)v\}[1-\mathcal{G}_{i,j}(v)],\ \text{i=1,2 and j=2, 3,…k.} \quad (22)$$

Theorem 1: The joint distribution of the queue length and server state is characterized by the following probability generating functions.

$$I_N^{(0)}(z) = \frac{(1-\rho)\sum_{n=0}^{N-1}\sigma_n z^n}{\sum_{n=0}^{N-1}\sigma_n} \quad (23)$$

$$W^{(1)}(u,z) = \frac{(1-\rho)\sum_{n=0}^{N-1}\sigma_n z^{n+1}C(z)[1-\mathcal{A}_1(u)]e^{-\delta_1(z)u}}{\sum_{n=0}^{N-1}\sigma_n[(\bar{p}+p\mathcal{A}_2^*(\delta_2(z))\mathcal{A}_1^*(\delta_1(z))-z]} \quad (24)$$

$$W^{(2)}(u,z) = p\frac{(1-\rho)\sum_{n=0}^{N-1}\sigma_n z^{n+1}C(z)\mathcal{A}_1^*(\delta_1(z))[1-\mathcal{A}_2(u)]e^{-\delta_2(z)u}}{\sum_{n=0}^{N-1}\sigma_n[(\bar{p}+p\mathcal{A}_2^*(\delta_2(z))\mathcal{A}_1^*(\delta_1(z))-z]} \quad (25)$$

$$L_1^{(1)}(u,v,z) = \frac{\beta_1(1-\rho)\sum_{n=0}^{N-1}\sigma_n z^{n+1}C(z)[1-\mathcal{A}_1(u)]e^{-\delta_1(z)u}[1-\mathcal{G}_{1,1}(v)]e^{-\varepsilon C(z)v}}{\sum_{n=0}^{N-1}\sigma_n[(\bar{p}+p\mathcal{A}_2^*(\delta_2(z))\mathcal{A}_1^*(\delta_1(z))-z]} \quad (26)$$

$$L_j^{(1)}(u,v,z) = \frac{\beta_1(1-\rho)\sum_{n=0}^{N-1}\sigma_n z^{n+1}C(z)[1-\mathcal{A}_1(u)]e^{-\delta_1(z)u}\prod_{t=1}^{j-1} q_t\mathcal{G}_{1,t}^*\big(\varepsilon C(z)\big)\,e^{-\varepsilon C(z)v}\,[1-\mathcal{G}_{1,j}(v)]}{\sum_{n=0}^{N-1}\sigma_n[(\bar{p}+p\mathcal{A}_2^*(\delta_2(z))\mathcal{A}_1^*(\delta_1(z))-z]} \quad (27)$$

$$L_1^{(2)}(u,v,z) = \frac{\beta_2 p(1-\rho)\sum_{n=0}^{N-1}\sigma_n z^{n+1}C(z)\mathcal{A}_1^*(\delta_1(z))[1-\mathcal{A}_2(u)]e^{-\delta_2(z)u}[1-\mathcal{G}_{2,1}(v)]e^{-\varepsilon C(z)v}}{\sum_{n=0}^{N-1}\sigma_n[(\bar{p}+p\mathcal{A}_2^*(\delta_2(z))\mathcal{A}_1^*(\delta_1(z))-z]} \quad (28)$$

$$L_j^{(2)}(u,v,z) = \frac{\beta_2 p(1-\rho)\sum_{n=0}^{N-1}\sigma_n z^{n+1}C(z)\mathcal{A}_1^*(\delta_1(z))[1-\mathcal{A}_2(u)]e^{-\delta_2(z)u}\prod_{t=1}^{j-1} q_t\mathcal{G}_{2,t}^*\big(\varepsilon C(z)\big)\,e^{-\varepsilon C(z)v}\,[1-\mathcal{G}_{2,j}(v)]}{\sum_{n=0}^{N-1}\sigma_n[(\bar{p}+p\mathcal{A}_2^*(\delta_2(z))\mathcal{A}_1^*(\delta_1(z))-z]} \quad (29)$$

$2 \le j \le k$

Proof:

For proof see appendix 1.

Theorem 2: The queue size distribution at different server states is described by the following marginal probability generating functions

$$W^{(1)}(z) = \frac{(1-\rho)\sum_{n=0}^{N-1}\sigma_n z^{n+1}C(z)[1-\mathcal{A}_1^*(\delta_1(z))]}{\sum_{n=0}^{N-1}\sigma_n[(\bar{p}+p\mathcal{A}_2^*(\delta_2(z))\mathcal{A}_1^*(\delta_1(z))-z]\delta_1(z)} \quad (30)$$

$$W^{(2)}(z) = p\frac{(1-\rho)\sum_{n=0}^{N-1}\sigma_n z^{n+1}C(z)\mathcal{A}_1^*(\delta_1(z))[1-\mathcal{A}_2^*(\delta_2(z))]}{\sum_{n=0}^{N-1}\sigma_n[(\bar{p}+p\mathcal{A}_2^*(\delta_2(z))\mathcal{A}_1^*(\delta_1(z))-z]\delta_2(z)} \quad (31)$$

$$L_1^{(1)}(z) = \frac{\beta_1(1-\rho)\sum_{n=0}^{N-1}\sigma_n z^{n+1}[1-\mathcal{A}_1^*(\delta_1(z))][1-\mathcal{G}_{1,1}^*(\varepsilon C(z))]}{\sum_{n=0}^{N-1}\sigma_n[(\bar{p}+p\mathcal{A}_2^*(\delta_2(z))\mathcal{A}_1^*(\delta_1(z))-z]\varepsilon\delta_1(z)} \quad (32)$$

$$L_j^{(1)}(z) = \frac{\beta_1(1-\rho)\sum_{n=0}^{N-1}\sigma_n z^{n+1}[1-\mathcal{A}_1^*(\delta_1(z))]\prod_{t=1}^{j-1} q_t\mathcal{G}_{1,t}^*(\varepsilon C(z))\,\big[1-\mathcal{G}_{1,j}^*(\varepsilon C(z))\big]}{\sum_{n=0}^{N-1}\sigma_n[(\bar{p}+p\mathcal{A}_2^*(\delta_2(z))\mathcal{A}_1^*(\delta_1(z))-z]\varepsilon\delta_1(z)} \quad (33)$$

$$L_1^{(2)}(z) = \frac{\beta_2 p(1-\rho)\sum_{n=0}^{N-1}\sigma_n z^{n+1}\mathcal{A}_1^*(\delta_1(z))[1-\mathcal{A}_2^*(\delta_2(z))][1-\mathcal{G}_{2,1}^*(\varepsilon C(z))]}{\sum_{n=0}^{N-1}\sigma_n[(\bar{p}+p\mathcal{A}_2^*(\delta_2(z))\mathcal{A}_1^*(\delta_1(z))-z]\varepsilon\delta_2(z)} \quad (34)$$

$$L_j^{(2)}(z) = \frac{\beta_2 p(1-\rho)\sum_{n=0}^{N-1}\sigma_n z^{n+1}\mathcal{A}_1^*(\delta_1(z))[1-\mathcal{A}_2^*(\delta_2(z))]\prod_{t=1}^{j-1} q_t\mathcal{G}_{2,t}^*(\varepsilon C(z))\big[1-\mathcal{G}_{2,j}^*(\varepsilon C(z))\big]}{\sum_{n=0}^{N-1}\sigma_n[(\bar{p}+p\mathcal{A}_2^*(\delta_2(z))\mathcal{A}_1^*(\delta_1(z))-z]\varepsilon\delta_2(z)} \quad (35)$$

$2 \le j \le k$

Proof: Integrating equations (24)-(25) with respect to x and equations (26)-(29) with respect to x, y , we get required result.

Theorem 3: Under the stability condition, the probability generating function of the stationary queue length at departure epochs is given by

$$\varsigma(z) = \frac{(1-\varepsilon\rho_1)\sum_{n=0}^{N-1}\sigma_n z^n C(z)\{\bar{p}+p\mathcal{A}_2^*(\delta_2(z))\}\mathcal{A}_1^*(\delta_1(z))}{\lambda E(X)\sum_{n=0}^{N-1}\sigma_n[\{\bar{p}+p\mathcal{A}_2^*(\delta_2(z))\}\mathcal{A}_1^*(\delta_1(z))-z]} \quad (36)$$

Proof: Appendix 2.

PERFORMANCE MEASURES

This section is devoted to the analysis of key performance indicators of the queueing system, achieved by evaluating the probability generating functions of the queue length under suitably chosen parameter configurations.

(a) Long Run Probabilities of the Server States

We derive the expressions for the long run probabilities of the server states by taking limiting values when $z \to 1$of the marginal probability generating function established in Theorem 3.

(i) The probability that the server being busy in rendering the $i^{th}(i = 1,2)$ phase of service is given by

$$P(W_1) = \lim_{z\to 1} W^{(1)}(z) = \frac{\lambda\, a_1^{(1)} E(X)}{1+\rho_1(1-\varepsilon)} \tag{37}$$

(ii) The probability that the server being busy in rendering the second optional phase of service is given by

$$P(W_2) = \lim_{z\to 1} W^{(2)}(z) = \frac{p\lambda\, a_2^{(1)} E(X)}{1+\rho_1(1-\varepsilon)} \tag{38}$$

(iii) The probabilities that the server is under first phase and $j^{th}(j = 2,\dots,k)$ phase repair when failed during the first essential phase service are given by

$$P(L_1^{(1)}) = \lim_{z\to 1} L_1^{(1)}(z) = \frac{\beta_1\lambda\, a_1^{(1)} E(X) g_{11}^{(1)}}{1+\rho_1(1-\varepsilon)} \tag{39}$$

$$P(L_j^{(1)}) = \lim_{z\to 1} L_j^{(1)}(z) = \frac{\beta_1\lambda\, a_1^{(1)} [\prod_{t=1}^{j-1} q_t] g_{1j}^{(1)} E(X)}{1+\rho_1(1-\varepsilon)}; 2 \le j \le k \tag{40}$$

(iv) The probabilities that the server is under under the first phase and $j^{th}(j = 2,\dots,k)$ phase when failed during the optional phase service are given by

$$P(L_1^{(2)}) = \lim_{z\to 1} L_1^{(2)}(z) = \frac{p\beta_2\lambda\, a_2^{(1)} E(X) g_{21}^{(1)}}{1+\rho_1(1-\varepsilon)} \tag{41}$$

$$P(L_j^{(2)}) = \lim_{z\to 1} L_j^{(2)}(z) = \frac{p\beta_2\lambda\, a_2^{(1)} [\prod_{t=1}^{j-1} q_t] g_{2j}^{(1)} E(X)}{1+\rho_1(1-\varepsilon)}; 2 \le j \le k \tag{42}$$

The probability that the server is idle is given by

$$P(I) = 1 - \sum_{i=1}^{2}\left(P(W_i) + \sum_{j=1}^{k} P(L_j^{(i)})\right) = 1 - \rho \tag{43}$$

(b) Mean Queue Length

On differentiating equation (36) and by setting$z = 1$, the expressions for the mean queue length at the departure epoch (L_q) can be obtained as follows:

$$L_q = (1 - \varepsilon\rho_1)\frac{D\prime(1)N\prime\prime(1) - D\prime\prime(1)N\prime(1)}{2(D\prime(1))^2} \tag{44}$$

where

$$N'(1) = -1 \tag{45}$$

$$N''(1) = -\frac{E(X^{(2)})}{E(X)} - 2\left[\frac{\sum_{n=0}^{N-1} n\xi_n}{(\sum_{n=0}^{N-1}\xi_n)} + \varepsilon\rho_1\right] \tag{46}$$

$$D'(1) = (\varepsilon\rho_1 - 1) \tag{47}$$

$$D''(1) = 2pa_1^{(1)}\, a_2^{(1)} \mathfrak{h}_1^{(1)} \mathfrak{h}_2^{(1)} + a_1^{(2)}\{\mathfrak{h}_1^{(1)}\}^2 + a_1^{(1)} \mathfrak{h}_1^{(2)} + pa_2^{(2)}\{\mathfrak{h}_2^{(1)}\}^2 + pa_2^{(1)} \mathfrak{h}_2^{(2)} \tag{48}$$

Also, the Mean waiting time can be obtained as

$$E(W_q) = \frac{L_q}{\lambda_{eff} E(X)} \tag{49}$$

where

$$\lambda_{eff} = \lambda I_N^{(0)}(1) + \lambda\varepsilon \sum_{i=1}^{2} W^{(i)}(1) + \lambda\varepsilon \sum_{i=1}^{2}\left(\sum_{t=1}^{k} L_t^{(i)}(1)\right)$$

(d) Reliability Indices

Let $S_{avl}(t)$ be the system be available at the time t. Then the steady state availability S_{avl}, which is the probability that the server is either busy with rendering service or in an idle state, is obtained using.

$$S_{avl} = \lim_{z\to 1}\{ I_N^{(0)}(1) + W^{(1)}(z) + W^{(2)}(z)\}$$

$$= 1\text{-}\frac{\lambda E(X)a_1^{(1)}\beta_1(g_{11}^{(1)} + \sum_{t=2}^{k}(\prod_{j=1}^{t-1} q_t)g_{1t}^{(1)} + \lambda E(X)a_2^{(1)}p\beta_2(g_{21}^{(1)} + \sum_{t=2}^{k}(\prod_{j=1}^{t-1} q_t)g_{2t}^{(1)}}{1 + \rho_1(1-\varepsilon)} \tag{50}$$

The steady state failure frequency is determined using

$$S_{Ff} = \frac{\lambda E(X)\left[\beta_1 a_1^{(1)} + p\beta_2 a_2^{(1)}\right]}{1+\rho_1(1-\varepsilon)} \tag{51}$$

NUMERICAL ILLUSTRATION

Here we are going to give numerical illustration and sensitivity analysis of the present problems. To facilitate numerical results, we have assume that service distribution follow exponential distribution and having first two moments $a_i^{(1)} = \frac{1}{\mu_i}, a_i^{(2)} = \frac{2}{\mu_i^2}; i = 1,2$ and further we assume the distribution of batch arrival follow geometric distribution with first two moments $E(X) = \frac{b}{a}, E(X^2) = \frac{b(1+b)}{a^2}; b = 1-a.$ The distribution of repair time is also assumed to be exponential g_{ij} and has the first two moments as $g_{ij}^{(1)} = \frac{1}{g_{ij}}, g_{ij}^{(2)} = \frac{2}{g_{ij}^2}; i = 1,2; j = 1,2,\ldots,k.$ To develop a computer program, the coding is done in MATLAB. Now we display the numerical results in Figures 2-5 and Tables I-IV.

For Figures 2-5, we set the default parameters as

$E(X) = 3, \mu_1 = \mu_2 = 8, N = 5, p = 0.6, \beta = 2, \beta_1 = \beta, \beta_2 = 0.8\beta, k = 3$

TABLE I
IMPACT OF ARRIVAL/SERVICE RATE ON MEAN QUEUE LENGTH (L_q) AND WAITING TIME (W_q)

	μ=7.5		μ=8		μ=8.5		μ=9	
λ	L_q	W_q	L_q	W_q	L_q	W_q	L_q	W_q
1.10	27.50	9.66	22.39	7.79	19.09	6.59	16.79	5.75
1.16	33.00	11.08	25.96	8.63	21.61	7.12	18.69	6.11
1.22	40.50	13.02	30.55	9.72	24.73	7.80	20.96	6.55
1.28	51.14	15.79	36.60	11.18	28.64	8.66	23.72	7.11
1.34	66.99	19.90	44.82	13.17	33.66	9.79	27.11	7.82
1.40	92.20	26.41	56.42	15.97	40.26	11.28	31.38	8.71

TABLE II
IMPACT OF ARRIVAL AND FAILURE RATE ON MEAN QUEUE LENGTH (L_q) AND WAITING TIME (W_q)

	p=0.1		p=0.3		p=0.5		p=0.7	
β	S_{avl}	S_{Ff}	S_{avl}	S_{Ff}	S_{avl}	S_{Ff}	S_{avl}	S_{Ff}
1.75	0.934	0.916	0.920	1.113	0.906	1.302	0.893	1.483
2.00	0.926	1.030	0.912	1.224	0.898	1.410	0.885	1.589
2.25	0.917	1.143	0.904	1.334	0.890	1.518	0.878	1.694
2.50	0.909	1.255	0.896	1.444	0.883	1.625	0.870	1.799
2.75	0.901	1.368	0.888	1.554	0.875	1.732	0.863	1.904
3.00	*0.893*	*1.480*	0.880	1.663	0.867	1.839	0.855	2.008

TABLE III
IMPACT OF FAILURE RATE (B) AND PROBABILITY P ON SERVER AVAILABILITY (S_{avl}) AND SERVER FAILURE (S_{Ff})

	B=1.5		B=2		B=2.5		B=3	
Λ	L_q	W_q	L_q	W_q	L_q	W_q	L_q	W_q
1.10	25.15	8.80	27.50	9.66	30.22	10.66	33.41	11.84
1.16	29.71	9.93	33.00	11.08	36.91	12.44	41.63	14.09
1.22	35.77	11.45	40.50	13.02	46.32	14.96	53.60	17.39
1.28	44.07	13.55	51.14	15.79	60.19	18.67	72.09	22.46
1.34	55.92	16.53	66.99	19.90	81.97	24.47	102.99	30.89
1.40	73.70	21.01	92.20	26.41	119.17	34.31	160.83	46.53

TABLE IV
IMPACT OF FAILURE RATE (B) AND BALKING RATE E ON SERVER AVAILABILITY (S_{avl}) AND SERVER FAILURE (S_{Ff})

	ε=0.1		ε =0.3		ε =0.5		ε =0.7	
β	S_{avl}	S_{Ff}	S_{avl}	S_{Ff}	S_{avl}	S_{Ff}	S_{avl}	S_{Ff}
0.1	0.925	1.037	0.918	1.134	0.910	1.251	0.899	1.395
0.3	0.919	1.115	0.911	1.231	0.901	1.374	0.888	1.554
0.5	0.914	1.184	0.905	1.319	0.893	1.487	0.877	1.704
0.7	0.910	1.246	0.899	1.397	0.885	1.591	0.867	1.847
0.9	0.906	1.300	0.894	1.468	0.878	1.687	0.857	1.981

The numerical results highlight several important trends in system performance. An increase in the arrival (λ) rate leads to a monotonic rise in both the average queue length (L_q) and the average waiting time (W_q), whereas higher service rates exert the opposite effect, reducing both measures. When examining the impact of the parameters λ and β, it is observed that increments in either parameter result in higher values of (L_q) and (W_q). However, the sensitivity to changes in β is more pronounced, producing sharper variations compared to those caused by λ.

Further, the joint influence of the failure rate (β)and the opting probability (p) for a second service reveals a clear trade-off between availability (S_{avl}) and failure frequency (S_{Ff}). Specifically, as either β or p increases, server availability decreases while the frequency of failures rises, underscoring the vulnerability of the system under such conditions. A similar deterioration is observed when considering the combined effect of the failure rate and the joining (balking) rate (ε), where higher values of ε lead to reduced server availability and elevated failure frequency. Overall, these findings emphasize that system congestion and unreliability intensify with rising arrival-related and failure-related parameters, while improvements in service rate alleviate these pressures.

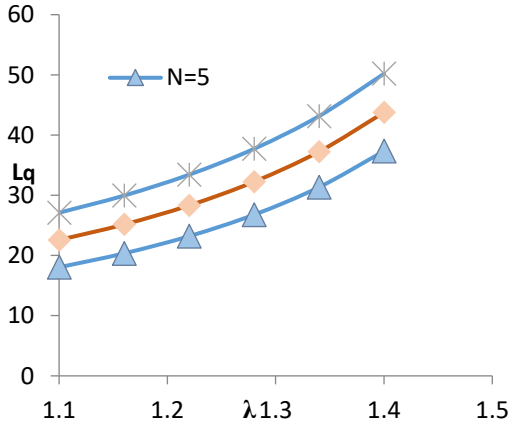

FIGURE 2

L_qVS Λ FOR VARIATION IN N

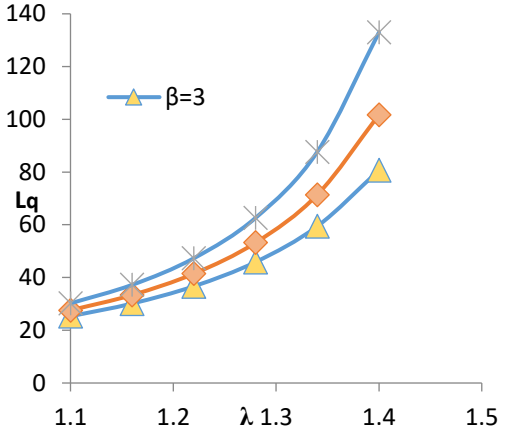

FIGURE 3

L_q VS Λ FOR VARIATION IN FAILURE RATE B

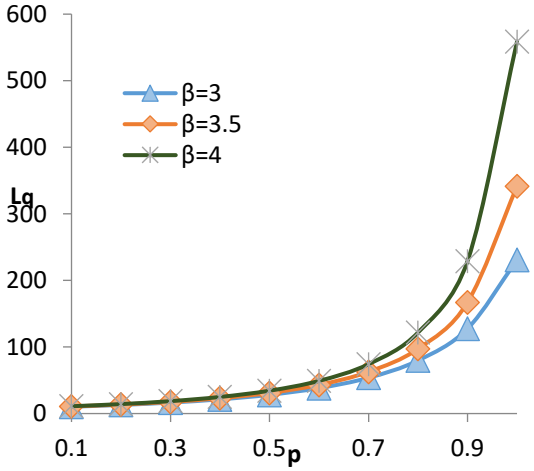

FIGURE.4

L_q VS P FOR VARIATION IN FAILURE RATE B

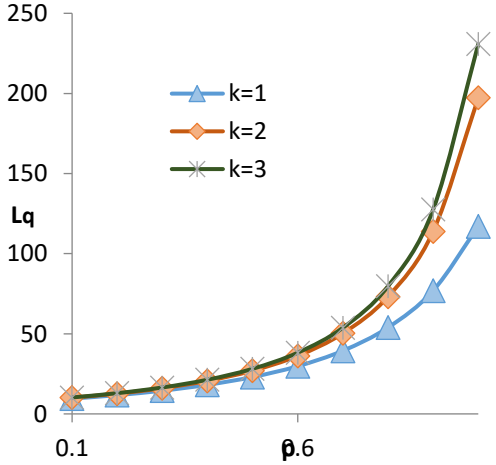

FIGURE 5

L_q VS P FOR VARIATION IN (K)

The graphical results provide several important insights into the behavior of the system. The average queue length (L_q) is observed to increase monotonically with the arrival rate (λ), and this growth becomes convex as λ rises, indicating that congestion intensifies more sharply at higher arrival intensities. A similar trend emerges when incorporating the effect of the failure rate (β); specifically, the curves display greater convexity for higher values of β, suggesting that system congestion is more sensitive to increases in λ under elevated failure conditions.

Further examination of the relationship between (L_q) and the opting probability (p) reveals a highly nonlinear pattern, particularly when the failure rate is large. This demonstrates that simultaneous increases in β and p can substantially magnify instability in the system. The impact of repair-related parameters is also significant: as the number of repair phases increases, the average queue length rises sharply, and this effect becomes more pronounced for larger values of p.

Taken together, the graphical analysis highlights that queue length is strongly influenced by arrival intensity, failure-related parameters, and repair mechanisms. While higher service thresholds and repair phases provide operational flexibility, they also exacerbate congestion when combined with elevated arrival rates, failure probabilities, or opting behavior, underscoring the delicate trade-off between system resilience and efficiency.

REFERENCES

[1] Sengupta, B. (1990) A queue with service interruptions in an alternating random environment, Operations Research, Vol. 38, No. 2, pp.308–318.
[2] Takine, T. and Sengupta, B. (1997) A single server queue with service interruptions, Queueing Systems, Vol. 26, Nos. 3–4, pp.285–300.
[3] Madan, K.C. (2000) An M / G / 1 queue with second optional service, Queueing Systems, Vol. 34, Nos. 1–4, pp.37–46.
[4] Ke, J.C. (2003) The optimal control of an M/G/1 queueing system with server vacations, startup and breakdowns, Computers & Industrial Engineering,Vol. 44(4), Pp. 567-579, https://doi.org/10.1016/S0360-8352(02)00235-8.
[5] Pearn, W.L., Ke, J.C., Chang,Y.C. (2004). Sensitivity analysis of the optimal management policy for a queuing system with a removable and non-reliable server, Computers & Industrial Engineering, Vol. 46(1), Pp. 87-99, https://doi.org/10.1016/j.cie.2003.11.001.
[6] Ke, J.C., Lin, C.H. (2006) Maximum entropy solutions for batch arrival queue with an unreliable server and delaying vacations, Applied Mathematics and Computation, Vol. 183(2), pp. 1328-1340, https://doi.org/10.1016/j.amc.2006.05.174.
[7] Ke, J.C. (2008) Two thresholds of a batch arrival queueing system under modified T-vacation policy with startup and closedown, Mathematical Methods in the Applied Sciences, Vol. 31, No. 2, pp.229–247
[8] Kumar, M.S. and Arumuganathan, R. (2010) An MX / G / 1 retrial queue with two-phase service subject to active server breakdowns and two types of repair, International Journal of Operational Research, Vol. 8, No. 3, pp.261–291
[9] Jain, M. and Agarwal, S. (2010) A discrete-time GeoX / G / 1 retrial queueing system with starting failures and optional service, International Journal of Operational Research, Vol. 8, No. 4, pp.428–457
[10] Choudhury,G., Tadj, L. (2011) The optimal control of an Mx/G/1 unreliable server queue with two phases of service and Bernoulli vacation schedule, Mathematical and Computer Modelling, Vol. 54(1-2), Pp. 673-688, https://doi.org/10.1016/j.mcm.2011.03.010.
[11] Wu, J. and Lian, Z. (2013).A single-server retrial G-queue with priority and unreliable server under Bernoulli vacation schedule, Computers & Industrial Engineering, Vol. 64(1), Pp. 84-93, https://doi.org/10.1016/j.cie.2012.08.015.
[12] Singh CJ, Jain M, Kumar B (2016) MX/G/1 unreliable retrial queue with option of additional service and Bernoulli vacation. Ain Shams Eng J 7(1):415–429
[13] Rajadurai P, Chandrasekaran VM, Saravanarajan MC (2018) Analysis of an unreliable retrial G-queue with working vacations and vacation interruption under Bernoulli schedule. Ain Shams Eng J 9(4):567–580
[14] Ayyappan G, Karpagam S (2019) Analysis of a bulk queue with unreliable server, immediate feedback, N-policy, Bernoulli schedule multiple vacation and stand-by server. Ain Shams Eng J 10(4):873–880
[15] Ayyappan, G., Nirmala, M. and Karpagam, S.(2020) Analysis of Repairable Single Server Bulk Queue with Standby Server, Two Phase Heterogeneous Service, Starting Failure and Multiple Vacation, *Int. J. Appl. Comput. Math* , Vol. **6**, No. 2, 52.
[16] Kumar, A. and Jain, M. (2022) Cost Optimization of an Unreliable server queue with two stage service process under hybrid vacation policy, Mathematics and Computers in Simulation, Vol.204, pp. 259-281
[17] Li, J. J., & Liu, L. W. (2023). The GI/M/1 queue in a multi-phase service environment with working vacations and Bernoulli vacation interruption. *Journal of the Operations Research Society of China, 11*: 627–656. https://doi.org/10.1007/s40305-021-00371-8
[18] Bharathi J, Nandhini S (2024) A single server Non-Markovian with non-compulsory re-service and balking under Modified Bernoulli Vacation. J King Saud Univ Sci, 36(1):103007
[19] Ayyappan, G., & Gurulakshmi, G. A. A. (2024). Analysis of MAP/PH/1 queue with differentiated vacation, vacation interruption under N-policy, optional service, breakdown, repair, setup and discouragement of customers. *International Journal of Mathematics in Operational Research, 27*(4): 415–457. https://doi.org/10.1504/IJMOR.2024.138463
[20] Kumar, B. (2024). Unreliable queue model with multi-phases of services, delay repair, and single vacation. *International Journal of Industrial and Systems Engineering, 48*(1): 100–124. https://doi.org/10.1504/IJISE.2024.140682
[21] Wang, K.H, Jau-Chuan Ke, J.C. (2003). Probabilistic analysis of a repairable system with warm standbys plus balking and reneging,Applied Mathematical Modelling, Vol. 27(4), pp.327-336.
[22] Jain, M, Rakhee, Maheshwari, S. (2004) N-policy for a machine repair system with spares and reneging, Applied Mathematical Modelling, Vol. 28(6), pp: 513-531.
[23] Movaghar, A. (2005) Optimal control of parallel queues with impatient customers, Performance Evaluation, Vol. 60 (1–4), pp. 327-343
[24] Xiong, W, Jagerman, D, Altiok,T. (2008) M/G/1 queue with deterministic reneging times,Performance Evaluation, Vol. 65(3–4), pp.308-316.
[25] Chakravarthy, S.R. (2009) A disaster queue with Markovian arrivals and impatient customers,Applied Mathematics and Computation, Vol. 214(1), pp. 48-59.

[26] Arrar, N.K, Djellab, N.V, Baillon, J.V. (2012) On the asymptotic behaviour of M/G/1 retrial queues with batch arrivals and impatience phenomenon, Mathematical and Computer Modelling, Vol. 55(3–4), pp. 654-665.
[27] Singh CJ, Jain M, Kumar B (2016) Analysis of single server finite queueing model with reneging. Int J Math Oper Res 9(1):15–37
[28] Yang, D.Y., Wu, Y.Y. (2017).Analysis of a finite-capacity system with working breakdowns and retention of impatient customers, Journal of Manufacturing Systems, Vol. 44(1), pp. 207-216.
[29] Morozov, E, Rumyantsev, A., Dey, S., Deepak, T.G.(2019). Performance analysis and stability of multiclass orbit queue with constant retrial rates and balking, Performance Evaluation, Vol. 134,102005.
[30] Bharathi J., Nandhini S.,(2024) A single server non-Markovian with non-compulsory re-service and balking under Modified Bernoulli Vacation, Journal of King Saud University - Science, Vol. 36(1),103007.

Appendix-1

Proof of theorem 1:

Take $z \to 1$ in equations (3.9) ,(3.6) , (3.10) and (3.11) for $i = 1,2$ we get

$$W^{(1)}(0,1) = \frac{\lambda E(X)\sum_{n=0}^{N-1} I_n^{(0)}}{[1-\rho]} \quad (1.1)$$

$$W^{(1)}(u,1) = W^{(1)}(0,1)[1-\mathcal{A}_1(u)] \quad (1.2)$$

$$W^{(2)}(u,1) = pW^{(1)}(0,1)[1-\mathcal{A}_2(u)] \quad (1.3)$$

$$L_1{}^{(1)}(u,v,1) = \beta_1 W^{(1)}(0,1)[1-\mathcal{A}_1(u)]\left[1-\mathcal{G}_{1,1}(v)\right] \quad (1.4)$$

$$L_j{}^{(1)}(u,v,1) = \beta_1 W^{(1)}(0,1)[1-\mathcal{A}_1(u)]\left(\prod_{t=1}^{j-1} q_t\right)\left[1-\mathcal{G}_{1,j}(v)\right],\ 2 \le j \le k \quad (1.5)$$

$$L_1{}^{(2)}(u,v,1) = \beta_2 pW^{(1)}(0,1)[1-\mathcal{A}_2(u)]\left[1-\mathcal{G}_{2,1}(v)\right] \quad (1.6)$$

$$L_j{}^{(2)}(u,v,1) = \beta_2 pW^{(1)}(0,1)[1-\mathcal{A}_2(u)]\left(\prod_{t=1}^{j-1} q_t\right)\left[1-\mathcal{G}_{2,j}(v)\right],\ 2 \le j \le k \quad (1.7)$$

The $I_n^{(0)}, (0 \le n \le N-1)$ satisfy the following relation

$$I_n^{(0)} = C_0\sigma_n,\ n = 0,1,\dots,N-1 \quad (1.8)$$

Where σ_n is given by (1) and C_0 is constant.

Then $I_N^{(0)}(z) = C_0\sum_{n=0}^{N-1}\sigma_n z^n$ (1.9)

Using equations (1.1) - (1.9) in the normalizing condition (11), we have

$$C_0 = \left(1-\frac{\rho_1}{1-\rho_2+\rho_1}\right)\frac{1}{\sum_{n=0}^{N-1}\sigma_n} = \frac{(1-\rho)}{\sum_{n=0}^{N-1}\sigma_n} \quad (1.10)$$

Using the value of (1.10) in (1.9) we get

$$I_N^{(0)}(z) = \frac{(1-\rho)\sum_{n=0}^{N-1}\sigma_n z^n}{\sum_{n=0}^{N-1}\sigma_n} \quad (1.11)$$

where ρ is the utilization factor.

Using the equation (1.11) in equations (17), (19)-(22), for $i = 1,2$ we get the equations (23)- (29).

Appendix-2

Proof of Theorem 3:

To obtain the queue size distribution at the departure epoch, on the line of Choudhury and Tadj (2009) and discussed by Wolff (1982), we have

$$\varsigma_t = C_0\left[\bar{p}\int_0^\infty \varphi_1(u)\, W_{t+1}^{(1)}(u)du + \int_0^\infty \varphi_2(u)\, W_{t+1}^{(2)}(u)du\right] \quad (2.1)$$

where C_0 is the normalizing constant and $\{\varsigma_t; t = 0,1,2,\dots\}$ as the probability that there are t customers in the queue at a departure epoch.

Multiplying equation (2.1) by z^t and using $\varsigma(z) = \sum_{t=0}^{\infty}\varsigma_t z^t$ and after simplification,

We get

$$\varsigma(z) = \frac{C_0 I_N^{(0)}(z)C(z)\{\bar{p}+p\mathcal{A}_2^*(\delta_2(z))\}\mathcal{A}_1^*(\delta_1(z))}{[\{\bar{p}+p\mathcal{A}_2^*(\delta_2(z))\}\mathcal{A}_1^*(\delta_1(z))-z]} \quad (2.2)$$

Utilizing the normalizing condition $\varsigma(1) = 1$, we get

$$C_0 = \frac{1-\varepsilon\rho_1}{\lambda E(X)\sum_{n=0}^{N-1}\sigma_n} \quad (2.3)$$

Putting the value of equation (2.3) in equation (2.2) we get equation (36).