

**MODELING A RETRIAL QUEUE WITH BATCH ARRIVALS AND
IMPATIENT CUSTOMERS UNDER SERVER FAILURES AND BERNOULLI-
CONTROLLED AVAILABILITY**

*** Ratiba MOULAI:**

National Higher School of Statistics and Applied Economics, Algeria
Email : ratiba.moulai@ensta.edu.dz

Isma BENRAIS :

National Higher School of Statistics and Applied Economics, Algeria
Email : benrais.isma@enssea.edu.dz

Mohamed BOUALEM :

*University of Bejaia, Faculty of Technology, Research Unit LaMOS, 06000 Bejaia,
Algeria*
Email : mohammed.boualem@univ-bejaia.dz

Reçu le: 26/04/2025 **Accepté le :** 18/05/2025 **Publication en ligne le:** 03/06/2025

ABSTRACT:

This paper analyzes an $M^X/G/1$ retrial queueing model with batch arrivals and impatient customers, where the server is subject to active breakdowns and follows a Bernoulli vacation schedule that incorporates a reservation period. When primary customers arrive and find the server unavailable—due to service, breakdown, reservation, or vacation—they may either leave the system permanently or enter an orbit, generating future retrial attempts. In case of a server failure during service, the interrupted customer may wait near the server until it is repaired or move to the orbit, which initiates a reservation period. After each service completion, the server may take a vacation with a certain probability or remain active. The supplementary variable technique is used to derive the system's steady-state performance measures. Numerical experiments are conducted through a sensitivity analysis to investigate the effects of key parameters. From a managerial perspective, this model offers valuable insights for the design and control of systems with unreliable servers, limited waiting spaces, and retrial behavior, such as in call centers, air traffic control, and smart service systems, where balancing server availability and customer patience is critical to performance.

Keywords: Retrial queues, batch customer arrivals, server breakdown and repair, Bernoulli-type server vacations, supplementary variable method.

*Ratiba MOULAI

I. INTRODUCTION

Retrial queueing systems have become essential in modeling a wide range of real-world scenarios, notably in telecommunications, call centers, computing, and manufacturing environments. These systems are characterized by the fact that, when the server is temporarily unavailable, arriving customers are redirected to a virtual holding area—known as an orbit—where they attempt to access the server repeatedly after random delays. A foundational contribution in this area was made by Cohen [8], who examined how recurrent call attempts affect traffic in telephony systems, laying the groundwork for improved call management strategies. For comprehensive surveys on retrial queues, readers may consult key references including Falin and Templeton [11], Artalejo [2, 3], Yang and Templeton [24], Artalejo and Gómez-Corral [4], Phung-Duc [21], and Kim and Kim [18].

Over time, various extensions of these models have been proposed, including vacation queueing systems, which simulate situations where the server is unavailable due to maintenance, multitasking across other service areas, or scheduled rest periods. In such cases, customers arriving during the server's absence may have to reattempt service later. This class of models has been widely studied, as seen in the work of Boualem et al. [6, 7], Gao and Zhang [13], Haridass and Arumuganathan [14], Jain and Bhagat [15], and Jose and Beena [16], and Boualem et al. [5]

This study focuses particularly on retrial queueing models that incorporate Bernoulli-type vacations. A detailed overview of research on this topic is provided in Micheal and Indhira [20]. The idea of Bernoulli vacations, introduced by Keilson and Servi [17], involves the server making a probabilistic decision after each service completion: with certain probability, the server enters a vacation period, and with complementary probability, it continues serving. Additionally, when no customers are present post-service, the server takes a vacation by default and awaits future arrivals.

Further contributions to this area include Krishna and Arivudainambi [19], who investigated an $M/G/1$ retrial queue with general retrial times and Bernoulli vacation behavior, using the supplementary variable technique to derive performance metrics and demonstrate the applicability of decomposition laws. Similarly, Arivudainambi and Gowsalya [1] analyzed a variant incorporating two distinct service types and initial server failures.

Another important extension involves modeling unreliable servers that may fail randomly [10]. In these cases, the server might either be repaired immediately or after a stochastic delay. Wu et al. [23] examined a system with active failures during service, where interrupted customers could either remain at the server or move to the orbit, initiating a reservation time. Tian et al. [22] studied breakdowns in $M/M/1$ retrial queues, where idle servers deactivate and resume only upon new arrivals. Gao et al. [12] explored

a similar $M/G/1$ structure, allowing customers to either queue or retry, with repairs triggered upon failure and outcomes depending on the customer's decision at breakdown. In many practical applications, customers or jobs arrive in batches rather than singly—such as grouped manufacturing items or network packets. Zirem et al. [25] investigated such scenarios using an $M^X/G/1$ retrial queue featuring batch arrivals, impatience, and active breakdowns. Their model included the notion of a reservation period, during which an interrupted customer either waits at the server or transitions to the orbit. Performance was analyzed through the supplementary variable approach.

Building upon this body of work, we propose a comprehensive $M^X/G/1$ retrial queueing model integrating batch arrivals, customer impatience, server failures with reservation time, and Bernoulli vacations. When customers arrive to find the server unavailable—due to failure, vacation, or reservation—they may either leave or enter the orbit. During breakdowns, the customer in service chooses whether to remain or retry later, with the server reserved during this interval. After each service, the server may enter a vacation based on a Bernoulli rule, or unconditionally if the system is idle. Using the supplementary variable method, we derive the system's steady-state behavior and assess key performance measures. Numerical sensitivity analyses are performed to examine parameter effects. This model is particularly suited to complex environments such as air traffic control, where retrial behavior, failures, and temporary unavailability are frequent and operationally significant.

The remainder of this paper is organized as follows: Section 2 introduces the model; Section 3 describes its real-world applications; Section 4 establishes the stability condition; Section 5 derives the steady-state distribution; Section 6 presents performance measures; Section 7 contains numerical experiments; Section 8 discusses managerial implications; and Section 9 concludes the study.

II. THE MATHEMATICAL FRAMEWORK :

This study examines a single-server retrial queue characterized by batch arrivals of the $M^X/G/1$ type, incorporating customer impatience, server subject to active failures, and vacations modeled by a Bernoulli process. Customers arrive in groups, or batches, according to a homogeneous Poisson process with a rate of $\delta > 0$. The batch size, denoted by X , is a random variable with the probability distribution $u_i = P(X = i)$ for $i = 1, 2, \dots$, and its generating function is $u(z)$, with the expected value given by $E(X) = u'(1)$. When a group of primary customers arrives and finds the server unavailable (due to being busy, down, reserved, or on vacation), there is a probability ζ that the group will enter an orbit, and a probability $1 - \zeta$ that they will exit the system permanently. When the orbit is unoccupied and the server is free, the customer at the front of the queue begins service immediately, while the remaining customers enter the orbit following the First-Come, First-Served (FCFS)

principle, generating a source of repeated calls. If the orbit is not empty, the customer at the front competes with the primary customer. If the primary customer arrives first, the customer at the head of the orbit has two options: to return to the orbit with probability q , or to exit the system without receiving service with probability $1 - q$. The retrial time is typically described by a distribution function $A(x)$, a density function $a(x)$, and a Laplace-Stieltjes transform $L_A(s)$, with the first two moments denoted as α_k , for $k = 1, 2$. The service time follows a general distribution characterized by the distribution function $F(x)$, the density function $f(x)$, and the Laplace-Stieltjes transform $L_B(s)$, along with its first two moments β_k , for $k = 1, 2$. The server is unreliable and subject to Poisson-distributed failures with a failure rate of ϕ , where failures occur after an exponentially distributed time with a mean of $1/\phi$. When a failure occurs, the server is immediately repaired, and the affected customer may either remain at the server with probability r , or join the service orbit with probability $\bar{r}$, continuing to randomly repeat their service request. The repair time is modeled by a general distribution, defined by the distribution function $H(x)$, the density function $h(x)$, and the Laplace-Stieltjes transform $L_C(s)$. The first two moments of the repair time are denoted as γ_k , where $k = 1, 2$. After the server undergoes repair, the customer remains in the service orbit, unaware of the server's restored functionality. As a result, the server must wait for the customer in orbit to initiate a callback in order to resume the service. During this repair period, while the customer stays in the orbit, the server is considered reserved and cannot accept new customers until the current customer leaves the system. The reservation time is described by a general distribution function $D(x)$, density function $d(x)$, and Laplace-Stieltjes transform $L_D(s)$, with the first two moments given by η_k , for $k = 1, 2$.

This model also incorporates a Bernoulli vacation process. After each service completion, the server has a probability α of taking a vacation, or a probability $1 - \alpha$ of remaining available to serve the next customer. If the orbit is empty after a service completion, the server will automatically enter a vacation period. The vacation time is governed by a general distribution with function $S(x)$, density function $s(x)$, Laplace-Stieltjes transform $L_H(s)$, and its first two moments denoted by θ_k , for $k = 1, 2$. Once the vacation period ends, the server will either wait for a customer in the orbit or for new arrivals.

It is assumed that the inter-arrival times, retrial times, service times, repair times, reservation times, and vacation times are independent of one another.

The conditional completion rates for retrials, service, repair, reservation, and vacation periods are as follows:

$$\alpha(\omega) = a(\omega) / [1 - A(\omega)],$$

$$\beta(x) = f(x) / [1 - F(x)],$$

$$\gamma(y) = h(y) / [1 - H(y)],$$

$$\eta(\tau) = d(\tau) / [1 - D(\tau)],$$

$$\theta(v) = s(v) / [1 - S(v)].$$

SYSTEM STATES

Let J_t denote the state of the server at time t . The server can be in one of five states: it is idle when $J_t = 0$, busy when $J_t = 1$, under repair when $J_t = 2$, reserved when $J_t = 3$, and on vacation when $J_t = 4$. Let J_t^* represent the state of the customer immediately after a server failure. If $J_t^* = 0$, the customer remains in front of the server; if $J_t^* = 0$, the customer enters the service orbit. Additionally, let N_t denote the number of customers in the retrial queue at time t . Given that the service time follows a general distribution, the process $\{J_t, J_t^*, N_t\}_{t \geq 0}$ is non-Markovian. To convert this into a Markov process, we introduce the following auxiliary variables:

-If $J_t = 0$ and $N_t > 0$, ξ_{0t} represents the elapsed retrial time of a customer in the retrial queue.

-If $J_t = 1, 2$ or 3 , ξ_{1t} denotes the elapsed service time.

-If $J_t = 2$ and $J_t^* = 0$ or 1 , ξ_{2t} represents the elapsed repair time.

-If $J_t = 3$, ξ_{3t} represents the elapsed reserved time.

-If $J_t = 4$, ξ_{4t} represents the elapsed vacation time.

Hence, the augmented stochastic process $\{Y_t\}_{t \geq 0} = \{J_t, J_t^*, N_t, \xi_{0t}, \xi_{1t}, \xi_{2t}, \xi_{3t}, \xi_{4t}\}$ becomes a Markov process.

III. PRACTICAL APPLICATIONS OF THE MODEL:

The queuing model we have developed can be applied to the management of runways in international airports. Aircraft often arrive in clusters during brief time windows, particularly during peak hours. If the runway is unavailable due to maintenance, adverse weather, or specific flight reservations, the arriving aircraft may either be diverted to another airport with probability $1 - \zeta$, or enter a holding pattern with probability ζ . This holding pattern, akin to an orbit, represents a scenario where aircraft circle the airport while awaiting clearance for landing.

When the holding zone is not empty, a competition occurs between incoming aircraft and those already waiting in the orbit. If a new group of aircraft arrives before the waiting ones are granted permission to land, the waiting aircraft face two options: they can either remain in the orbit and attempt landing again later with probability q , or they may choose to leave the system permanently with probability $1 - q$.

Runway failures can also disrupt the landing process, such as due to technical

malfunctions or unforeseen weather changes. In such cases, aircraft already in the landing process may either stay close to the runway with probability r , waiting for the runway to be restored, or be directed to a special holding zone (orbit service) with probability $\bar{r}$. Within this orbit, the aircraft must wait until the runway is restored, with the runway being reserved solely for that aircraft. No other aircraft can use the runway until the interrupted landing is completed.

After each landing (i.e., after each service completion), the runway may enter a temporary vacation period with probability α , to allow for safety inspections, maintenance, or cleaning operations. With probability $1 - \alpha$, the runway remains available for the next aircraft. This model reflects the complexities of air traffic management, encompassing vital operational aspects such as aircraft impatience (diverted planes), queuing and holding patterns, competition for runway access, failures during service, exclusive runway reservations, and vacation periods post-service.

Consequently, the proposed queuing model proves to be a useful analytical tool for optimizing runway management strategies in airports, especially under conditions of high traffic and uncertainty.

IV. STABILITY CONDITION :

This section aims to determine the exact condition that ensures the stability of the system. Let $\{t_n; n \in N\}$ represent the sequence of time moments when a vacation ends or a service is completed. The sequence of random variables $Z_n = (J(t_n+), N(t_n+))$ forms a Markov chain, which acts as the embedded Markov chain for our queueing system.

Theorem 1: *The embedded Markov chain $\{Z_n, n \in N\}$ is ergodic if and only if the following condition is satisfied:*

$$\begin{aligned} \zeta \delta u'(1) \left((1 + \phi(\gamma_1 + \bar{r} \eta_1)) \beta_1 + \alpha \theta_1 \right) \\ < (1 - q + q L_A(\delta)) + (1 - L_A(\delta))(1 - u'(1)). \end{aligned}$$

Proof. The proof is derived by directly applying Foster's drift criterion for the sufficiency condition and Kaplan's condition for the necessity. Since customers arrive following a Poisson process, Burke's theorem [9] indicates that the steady-state probabilities of the Markov process $\{Y_t\}_{t \geq 0}$ exist and are positive if and only if the above condition is met.

$$\begin{aligned} \zeta \delta u'(1) \left((1 + \phi(\gamma_1 + \bar{r} \eta_1)) \beta_1 + \alpha \theta_1 \right) \\ < (1 - q + q L_A(\delta)) + (1 - L_A(\delta))(1 - u'(1)). \end{aligned}$$

V. STEADY-STATE DISTRIBUTION :

This section focuses on deriving the joint distribution of both the server's state and the orbit size. To accomplish this, we first define the probability that the server is idle and

that the system contains no customers within the process $\{Y_t\}_{t \geq 0}$, along with the corresponding probability densities, as outlined below:

$$\begin{aligned}
 p_{00}(t) &= P(J_t = 0, N_t = 0), \\
 p_{0i}(t, \omega) d\omega &= P(J_t = 0, N_t = i, \omega < \xi_{0t} < \omega + d\omega), \omega \geq 0, i \geq 1, \\
 p_{1i}(t, x) dx &= P(J_t = 1, N_t = i, x < \xi_{1t} < x + dx), x \geq 0, i \geq 0, \\
 p_{20i}(t, x, y) dx dy &= P(J_t = 2, J_t^* = 0, N_t = i, x < \xi_{1t} < x + dx, y < \xi_{2t} < y + dy), x \geq 0, y \geq 0, i \geq 0, \\
 p_{21i}(t, x, y) dx dy &= P(J_t = 2, J_t^* = 1, N_t = i, x < \xi_{1t} < x + dx, y < \xi_{2t} < y + dy), x \geq 0, y \geq 0, i \geq 0, \\
 p_{3i}(t, x, \tau) dx d\tau &= P(J_t = 3, N_t = i, x < \xi_{1t} < x + dx, \tau < \xi_{3t} < \tau + d\tau), x \geq 0, \tau \geq 0, i \geq 0, \\
 p_{4i}(t, v) dv &= P(J_t = 4, N_t = i, v < \xi_{4t} < v + dv), v \geq 0, i \geq 0.
 \end{aligned}$$

We assume that the stability condition holds:

$$\begin{aligned}
 \zeta \delta u'(1) &\left((1 + \phi(\gamma_1 + \bar{r} \eta_1)) \beta_1 + \alpha \theta_1 \right) \\
 &< (1 - q + q L_A(\delta)) + (1 - L_A(\delta))(1 - u'(1)).
 \end{aligned}$$

This leads to:

$$\begin{aligned}
 \lim_{t \rightarrow \infty} p_{00}(t) &= p_{00}, \quad \lim_{t \rightarrow \infty} p_{0i}(t, \omega) = p_{0i}(\omega), \quad \lim_{t \rightarrow \infty} p_{1i}(t, x) = p_{1i}(x), \\
 \lim_{t \rightarrow \infty} p_{20i}(t, x, y) &= p_{20i}(x, y), \quad \lim_{t \rightarrow \infty} p_{21i}(t, x, y) = p_{21i}(x, y), \quad \lim_{t \rightarrow \infty} p_{3i}(t, x, \tau) \\
 &= p_{3i}(x, \tau), \\
 \lim_{t \rightarrow \infty} p_{4i}(t, v) &= p_{4i}(v).
 \end{aligned}$$

By using the technique of supplementary variables, we derive the balance equations that characterize the system's dynamic behavior.

$$\begin{aligned}
 \delta p_{00} \\
 = \int_0^{+\infty} p_{40}(v) \theta(v) dv,
 \end{aligned} \tag{1}$$

$$\begin{aligned}
 \left(\frac{\partial}{\partial \omega} + \alpha(\omega) + \delta \right) p_{0i}(\omega) &= 0, \quad i \\
 &\geq 1,
 \end{aligned} \tag{2}$$

$$\begin{aligned} \left(\frac{\partial}{\partial x} + \phi + \beta(x) + \zeta\delta \right) p_{1i}(x) &= \zeta\delta(1 - \delta_{i0}) \sum_{j=1}^i u_j p_{1i-j}(x) \\ &+ \int_0^{+\infty} p_{3i}(x, \tau) \eta(\tau) d\tau + \int_0^{+\infty} p_{20i}(x, y) \gamma(y) dy, \end{aligned} \quad (3)$$

$$\begin{aligned} \left(\frac{\partial}{\partial y} + \gamma(y) + \zeta\delta \right) p_{20i}(x, y) \\ = \zeta\delta(1 - \delta_{i0}) \sum_{j=1}^i u_j p_{20i-j}(x, y), \end{aligned} \quad (4)$$

$$\begin{aligned} \left(\frac{\partial}{\partial y} + \gamma(y) + \zeta\delta \right) p_{21i}(x, y) \\ = \zeta\delta(1 - \delta_{i0}) \sum_{j=1}^i u_j p_{21i-j}(x, y), \end{aligned} \quad (5)$$

$$\begin{aligned} \left(\frac{\partial}{\partial \tau} + \eta(\tau) + \zeta\delta \right) p_{3i}(x, \tau) \\ = \zeta\delta(1 - \delta_{i0}) \sum_{j=1}^i u_j p_{3i-j}(x, \tau), \end{aligned} \quad (6)$$

$$\begin{aligned} \left(\frac{\partial}{\partial v} + \theta(v) + \zeta\delta \right) p_{4i}(v) \\ = \zeta\delta(1 - \delta_{i0}) \sum_{j=1}^i u_j p_{4i-j}(v), \end{aligned} \quad (7)$$

$$\begin{aligned} p_{0i}(0) &= (1 - \alpha) \int_0^{\infty} p_{1i}(x) \beta(x) dx \\ &+ \int_0^{\infty} p_{4i}(v) \theta(v) dv, \end{aligned} \quad (8)$$

$$\begin{aligned} p_{1i}(0) &= \delta u_{i+1} p_{00} + q\delta(1 - \delta_{i0}) \sum_{j=1}^i u_j \int_0^{\infty} p_{0,i-j+1}(\omega) d\omega \\ &+ (1 - q) \delta \sum_{j=1}^{i+1} u_j \int_0^{\infty} p_{0,i-j+2}(\omega) d\omega + \int_0^{\infty} p_{0,i+1}(\omega) \alpha(\omega) d\omega, \end{aligned} \quad (9)$$

$$\begin{aligned} p_{20i}(x, 0) \\ = r\phi p_{1i}(x), \end{aligned} \quad (10)$$

$$\begin{aligned} p_{21i}(x, 0) \\ = \bar{r}\phi p_{1i}(x), \end{aligned} \quad (11)$$

$$\begin{aligned} p_{3i}(x, 0) \\ = \int_0^{+\infty} p_{21i}(x, y)\gamma(y) dy, \end{aligned} \quad (12)$$

$$\begin{aligned} p_{40}(0) \\ = \int_0^{+\infty} p_{10}(x)\beta(x) dx, \end{aligned} \quad (13)$$

$$\begin{aligned} p_{4i}(0) &= \alpha \int_0^{+\infty} p_{1i}(x)\beta(x) dx, \quad i \\ &\geq 1, \end{aligned} \quad (14)$$

$$\begin{aligned} \sum_{i=0}^{+\infty} \left(\int_0^{+\infty} p_{1i}(x) dx + \int_0^{+\infty} \int_0^{+\infty} p_{20i}(x, y) dx dy + \int_0^{+\infty} \int_0^{+\infty} p_{21i}(x, y) dx dy \right. \\ \left. + \int_0^{+\infty} \int_0^{+\infty} p_{3i}(x, \tau) dx d\tau + \int_0^{+\infty} p_{4i}(v) dv \right) + p_{00} \\ + \sum_{i=1}^{+\infty} \int_0^{+\infty} p_{0i}(\omega) d\omega \\ = 1, \end{aligned} \quad (15)$$

where, δ_{i0} is the Kronecker delta.

Theorem 2. When the stability condition is satisfied, the joint distribution of the server's state and orbit size is described by the following generating functions:

$$\begin{aligned} p_0(z, \omega) \\ = \frac{\delta z(zM(z) - u(z)K(z)N(z))e^{-\delta\omega}(1 - A(\omega))p_{00}}{E(z)}, \end{aligned} \quad (16)$$

$$\begin{aligned} p_1(z, x) \\ = \frac{\delta p_{00}J(z)e^{-G(\zeta\delta(1-u(z)))x}(1 - F(x))}{E(z)}, \end{aligned} \quad (17)$$

$$p_{20}(z, x, y) = \frac{r\phi \delta p_{00}J(z)}{E(z)} \times e^{-G(\zeta\delta(1-u(z)))x-\zeta\delta(1-u(z))y}(1-F(x))(1-H(y)), \quad (18)$$

$$p_{21}(z, x, y) = \frac{\bar{r}\phi \delta p_{00}J(z)}{E(z)} \times e^{-G(\zeta\delta(1-u(z)))x-\zeta\delta(1-u(z))y}(1-F(x))(1-H(y)), \quad (19)$$

$$p_3(z, x, \tau) = \frac{\bar{r}\phi \delta p_{00}J(z)L_C(\zeta\delta(1-u(z)))}{E(z)} \times e^{-G(\zeta\delta(1-u(z)))x-\zeta\delta(1-u(z))\tau}(1-F(x))(1-D(\tau)), \quad (20)$$

$$p_4(z, v) = \frac{Q(z)}{E(z)} \times \frac{e^{-\zeta\delta(1-u(z))v}(1-S(v)) \delta p_{00}}{L_H(\zeta\delta)}, \quad (21)$$

$$p_{00} = \frac{(1-u'(1)-q)(1-L_A(\delta)) + (1-K'(1)-N'(1))}{R}, \quad (22)$$

where,

$$E(z) = u(z)K(z)N(z)(1-z)(1-q+qL_A(\delta)) - z(z-u(z)K(z)N(z)) + K(z)N(z)L_A(\delta)(z-u(z)),$$

$$J(z) = M(z)(u(z)(1-z)(1-q+qL_A(\delta)) + L_A(\delta)(z-u(z))) - zu(z)(1-M(z)),$$

$$Q(z) = K(z)((1-\alpha) + \alpha L_H(\zeta\delta))(u(z)(1-z)(1-q+qL_A(\delta)) + L_A(\delta)(z-u(z))) - z(1-\alpha)(z-u(z)K(z)),$$

$$R = \left(1-q + (q-\zeta u'(1))L_A(\delta)\right) \left(1 + \frac{K'(1)+N'(1)-M'(1)}{\zeta u'(1)}\right) - M'(1) \left(1 + \frac{1-u'(1)}{\zeta u'(1)}\right) + L_A(\delta) \left(\zeta u'(1) - (K'(1) + N'(1) - M'(1)) \left(\frac{1}{\zeta u'(1)} - \frac{1}{p}\right)\right),$$

$$G(\zeta\delta(1-u(z))) = \zeta\delta(1-u(z)) + \phi - \phi L_C(\zeta\delta(1-u(z)))[r + \bar{r} L_D(\zeta\delta(1-u(z)))],$$

$$K(z) = L_B\left(G(\zeta\delta(1-u(z)))\right),$$

$$M(z) = 1 + \frac{(1 - \alpha)}{L_H(\zeta\delta)}(1 - L_H(\zeta\delta(1 - u(z)))) ,$$

$$N(z) = (1 - \alpha) + \alpha L_H(\zeta\delta(1 - u(z))) .$$

Proof. By solving the equations (1)–(15) and applying the definition of probability generating functions for $|z| \leq 1$, we derive the expressions presented in the theorem. The generating functions are given by the following:

$$\begin{aligned} p_0(z, \omega) &= \sum_{i=1}^{\infty} p_{0i}(\omega) z^i, \quad p_1(z, x) \\ &= \sum_{i=0}^{\infty} p_{1i}(x) z^i, \end{aligned} \quad (23)$$

$$\begin{aligned} p_{20}(z, x, y) &= \sum_{i=0}^{\infty} p_{20i}(x, y) z^i, \quad p_{21}(z, x, y) \\ &= \sum_{i=0}^{\infty} p_{21i}(x, y) z^i, \end{aligned} \quad (24)$$

$$\begin{aligned} p_3(z, x, \tau) &= \sum_{i=0}^{\infty} p_{3i}(x, \tau) z^i, \quad p_4(z, v) \\ &= \sum_{i=0}^{\infty} p_{4i}(v) z^i. \end{aligned} \quad (25)$$

VI. PERFORMANCE MEASURES:

In this section, we utilize the previous theorem to derive key performance metrics that enable the analysis of the steady-state behavior of our queuing system.

1. System state probabilities

Corollary 2.1. *If the system is stable, then:*

The probability that the system is non-empty and the server is idle is:

$$p_0 = \frac{-(1 - u'(1) - K'(1) + M'(1) - N'(1)) \cdot (1 - L_A(\delta))}{R}. \quad (26)$$

The probability that the server is occupied is:

$$p_1 = \frac{\delta \beta_1 ((1 - q + q L_A(\delta)) - L_A(\delta)(1 - u'(1)) - M'(1))}{R}. \quad (27)$$

The probability that the server is on vacation is :

$$p_4 = \frac{\delta \theta_1 \left(((1 - \alpha) + \alpha L_H(\zeta \delta)) \left((1 - q + q L_A(\delta)) - L_A(\delta)(1 - u'(1)) \right) + (1 - \alpha)(1 - u'(1) - K'(1)) \right)}{L_H(\zeta \delta) R}. \quad (28)$$

Proof. We know that:

$$p_0(z) = \int_0^\infty p_0(z, \omega) d\omega, \quad p_1(z) = \int_0^\infty p_1(z, x) dx, \quad p_4(z) = \int_0^\infty p_4(z, v) dv.$$

The integrands are derived based on the results established in Theorem 2. After determining the expressions for $p_0(z)$, $p_1(z)$, and $p_4(z)$, we evaluate their limits as $z \rightarrow 1$ to compute the corresponding probabilities mentioned above.

2. The expected number of customers in the orbit

Corollary 2.2. Let EN_q denote the expected number of customers in the orbit. Provided that the stability condition holds, the following relation is satisfied:

$$EN_q = \frac{C_q'''(1) \times D''(1) - C_q''(1) D'''(1)}{3(D''(1))^2} \cdot \frac{p_{00}}{\zeta L_H(p \delta)}, \quad (29)$$

where,

$$C_q(z) = S_q(z) + T_q(z) - U_q(z),$$

$$S_q(z) = (u(z)(1 - z)(1 - q + q L_A(\delta)) + L_A(\delta)(z - u(z)) \\ [K(z)(N(z)L_H(\zeta \delta)\zeta(1 - u(z) + ((1 - \alpha) + \alpha L_H(\zeta \delta))(1 - L_H(\zeta \delta(1 - u(z)))) - \\ M(z)L_H(\zeta \delta)) + M(z)L_H(\zeta \delta)]],$$

$$T_q(z) = zu(z)K(z)(N(z)L_H(\zeta\delta)\zeta(1-u(z))L_A(\delta) + (1-\alpha)(1-L_H(\zeta\delta(1-u(z)))) + (1-M(z))L_H(\zeta\delta)),$$

$$U_q(z) = z(u(z)(1-M(z))L_H(\zeta\delta) + zL_H(\zeta\delta)\zeta(1-u(z))(1-M(z)(1-L_A(\delta))) + z(1-\alpha)(1-L_H(\zeta\delta(1-u(z))))),$$

$$D(z) = E(z)(1-u(z)).$$

Proof. We know that:

$$EN_q = \lim_{z \rightarrow 1} \frac{\partial p_q(z)}{\partial z},$$

where the generating function $p_q(z)$ is defined as:

$$p_q(z) = \mathbb{E}[z^{N_q}] = p_{00} + p_0(z) + p_1(z) + p_{20}(z) + p_{21}(z) + p_3(z) + p_4(z).$$

By applying l'Hôpital's rule, we obtain equation (29).

VII. NUMERICAL APPLICATIONS:

This section presents numerical examples to investigate the impact of parameter variations on EN_q and p_{00} . The computations are carried out using the R software. We consider a retrial queuing system with an arrival rate δ , where the batch size follows a geometric distribution with parameter g . All parameter values chosen in this section ensure the stability condition of the system is met. The service, repair, reservation, and vacation times are modeled by exponential distributions with rates β, γ, η , and θ , respectively. Additionally, retrial times are examined under different distributions, including exponential, Erlang, and hyperexponential.

Interpretation of the results:

- Figures 1 and 2 show the influence of the arrival rate δ on both the average orbit size EN_q and the probability p_{00} that the system is idle and empty. The following parameter values are used: $\zeta = 0.5, q = 0.5, r = 0.5, g = 0.5, \alpha = 0.5, \phi = 0.5, \beta = 3, \gamma = 6, \eta = 6, \theta = 6$, and $\omega = 2$. As δ increases, EN_q grows exponentially because more customers enter the system. This trend is observed across all retrial time distributions (exponential, hyper-exponential, and Erlang), with the Erlang distribution exhibiting the steepest increase. Conversely, p_{00} decreases as δ rises, as the system is less likely to be empty when more customers

are arriving.

- Figures 3 and 4 illustrate the effect of the average service time $\beta_1 = 1/\beta$ on EN_q and pp_{00} , with the following parameters: $\zeta = 0.5, q = 0.5, r = 0.5, g = 0.5, \delta = 1, \alpha = 0.5, \phi = 0.5, \gamma = 6, \eta = 6, \theta = 6$, and $\omega = 2$. As β_1 increases, EN_q rises because longer service times delay the server's availability, resulting in a larger orbit. Consequently, p_{00} decreases as the server remains occupied for a longer period, reducing the chances of the system being empty. Among the different distributions, the Erlang distribution consistently produces higher EN_q and lower p_{00} values compared to the exponential and hyper-exponential distributions.
- Figure 5 presents the impact of γ_1 ($\gamma_1 = 1/\gamma$) on the mean orbit size EN_q . The parameters used are $\zeta = 0.5, q = 0.5, r = 0.5, g = 0.5, \delta = 0.9, \alpha = 0.5, \phi = 0.5, \beta = 3, \eta = 6$, and $\theta = 6$. The results show a linear increase in EN_q as γ_1 grows, due to longer repair times causing the server to remain idle longer, allowing more customers to accumulate in the orbit. Once again, the Erlang distribution shows significantly higher EN_q values than the exponential and hyper-exponential distributions.
- Figure 6 shows the effect of θ_1 ($\theta_1 = 1/\theta$) on EN_q , using the parameters $\zeta = 0.5, q = 0.5, r = 0.5, g = 0.5, \delta = 0.9, \alpha = 0.5, \phi = 0.5, \beta = 3, \gamma = 6, \eta = 6$, and $\omega = 2$. As θ_1 increases, the mean orbit size grows because longer server vacations extend periods of unavailability, causing more customers to wait. The Erlang distribution continues to dominate, consistently producing the highest EN_q values among the three distributions.

Figure 01: Effect of δ on EN_q

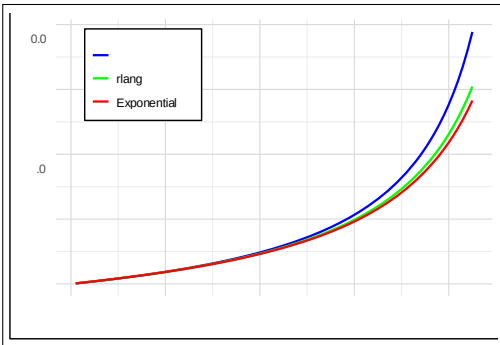


Figure 2: Effect of δ on p_{00}

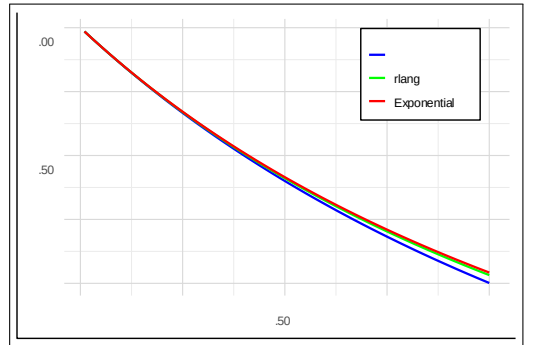


Figure 3: Effect of β_1 on ENq .

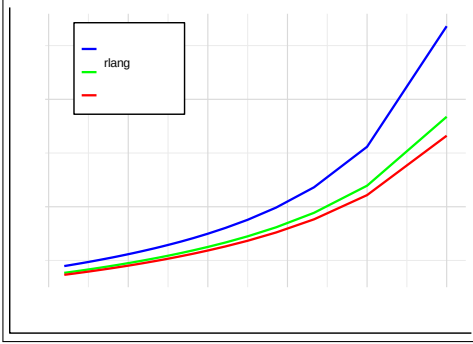


Figure 4: Effect of β_1 on p_{00} .

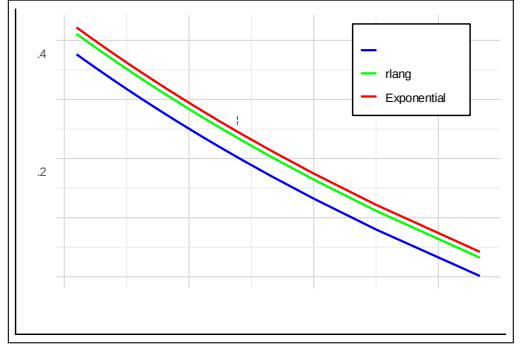


Figure 5: Effect of γ_1 on ENq .

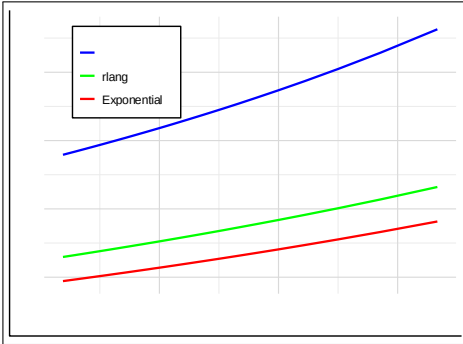
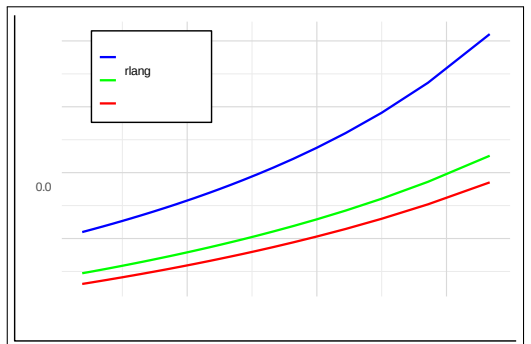


Figure 6: Effect of θ_1 on ENq .



VIII. MANAGERIAL IMPLICATIONS:

The proposed queueing model offers several actionable insights for managing service systems where customer retrials, impatience, and server unreliability are critical considerations.

Managing Customer Impatience: The probabilities ζ (orbit entrance) and $1 - \zeta$ (abandonment) model customer tolerance when the server is unavailable. Managers can influence these by improving estimated waiting time information, offering callback options, or deploying virtual queues to reduce abandonment rates.

Optimizing Callback and Retrial Policies: The use of an orbit and the probability q for retry vs. dropout reflects customer behavior after initial deferral. Adjusting these mechanisms allows managers to manage peak load efficiently without additional physical resources, particularly in automated systems like call centers or ticketing platforms.

Enhancing System Availability through Failure Management: Server failures occur during service (active failures), and customers interrupted during service may re-enter the orbit. This highlights the importance of robust server monitoring and preventive maintenance. Designing quick repair protocols and informing customers about expected downtimes can reduce system congestion and dissatisfaction.

Service Continuity with Reservation Periods: The concept of a reserved server—unavailable during post-repair until the orbiting customer recalls—emphasizes the need for intelligent service resumption protocols. Systems should facilitate automated customer re-engagement to minimize server idle time and improve throughput.

Balancing Utilization and Rest via Bernoulli Vacations: The probabilistic vacation mechanism (α) allows controlled server rest. In high-traffic environments, dynamically adjusting this vacation probability based on system state helps reduce resource exhaustion while maintaining acceptable service levels.

Designing Adaptive Queuing Policies: The state-dependent decisions embedded in the model offer a framework for dynamic policy implementation. Managers can use real-time performance metrics to adapt thresholds for balking, reneging, or server vacations, improving both efficiency and customer experience.

IX. CONCLUSION:

In this study, we analyzed an $M^X/G/1$ retrial queueing system with impatient customers and an unreliable server subject to active breakdowns. The model includes a reservation time after each repair and incorporates a Bernoulli vacation mechanism following service completions. General distributions were assumed for retrial, service, repair, reservation, and vacation times. By applying the supplementary variable method, we derived the generating functions related to the number of customers in the orbit and computed several key performance measures, including the expected orbit size and the probabilities of server idleness, occupation, and vacation. A key modeling assumption was that the server takes a vacation automatically when no customers are present in the orbit. The proposed model is particularly well-suited to real-world systems characterized by service interruptions and retrial behaviors, such as air traffic control during landing operations or resource allocation in telecommunication infrastructures. By simulating

various scenarios and adjusting input parameters, the model provides a robust tool for optimizing operational performance.

From a managerial standpoint, the model assists decision-makers in understanding the trade-offs between customer satisfaction and system reliability. It offers a means to evaluate how service policies and resource configurations influence system efficiency and customer experience. For example, the Bernoulli vacation policy can help balance service availability with planned downtimes, while the introduction of a reservation period reflects the need for prioritized resource recovery after failure.

As a direction for future research, this work could be extended by removing the assumption that the server always takes a vacation when the orbit is empty, or by considering state-dependent retrial rates. Moreover, incorporating statistical inference techniques to estimate model parameters based on real data would significantly enhance its practical applicability. This could include the development of maximum likelihood estimators, Bayesian inference approaches, or simulation-based calibration methods, enabling the model to be adapted and validated using empirical observations. Such statistical integration would support the design of data-driven decision tools and improve predictive accuracy in dynamic environments.

X. REFERENCES :

- [1] D. Arivudainambi and M. Gowsalya. Analysis of an $M/G/1$ retrial queue with Bernoulli vacation, two types of service and starting failure. *International Journal of Artificial Intelligence and Soft Computing*, 6(3):222–249, 2017.
- [2] J. R. Artalejo. Accessible bibliography on retrial queues. *Mathematical and Computer Modelling*, 30(3-4):1–6, 1999.
- [3] J. R. Artalejo. Accessible bibliography on retrial queues: progress in 2000–2009. *Mathematical and Computer Modelling*, 51(9-10):1071–1081, 2010.
- [4] J. R. Artalejo and A. Gómez-Corral. *Retrial Queueing Systems: A Computational Approach*. Springer, 2008.
- [5] M. Boualem, A. A. Bouchentouf, A. Bareche, and M. Cherfaui. Stochastic interpretation for a single server retrial queue with Bernoulli feedback and negative customers. *Applied Mathematics Journal of Chinese Universities*, 40:1–19, 2025
- [6] M. Boualem, N. Djellab, and D. Aïssani. Stochastic inequalities for $M/G/1$ retrial queues with vacations and constant retrial policy. *Mathematical and Computer Modelling*, 50(1-2):207–212, 2009.
- [7] M. Boualem, N. Djellab, and D. Aïssani. An $M/G/1$ retrial queue with exhaustive service and server vacations. *Journal of Communication and Computer*, 8(9):720–

726, 2011.

- [8] J. W. Cohen. Basic problems of telephone traffic theory and the influence of repeated calls. *Philips Telecommunication Rev.*, 18(2):49, 1957.
- [9] R. B. Cooper. Introduction to queueing theory. *Elsevier North Holland Inc.*, 1981.
- [10] N. Dehamnia, M. Boualem, and D. Aïssani. Performance and economic analysis of an unreliable single-server queue with general retrial times and varied customer patience levels. *Hacettepe Journal of Mathematics and Statistics*, pages 1–28, In press 2025.
- [11] G. I. Falin and J.G.C. Templeton. *Retrial Queues*. Chapman and Hall/CRC Press, 1997.
- [12] S. Gao, J. Wang, and T. Van Do. A repairable retrial queue under Bernoulli schedule and general retrial policy. *Annals of Operations Research*, 247(1):169–192, 2016.
- [13] S. Gao and D. Zhang. Performance and sensitivity analysis of an $M/G/1$ queue with retrial customers due to server vacation. *Ain Shams Engineering Journal*, 11(3):795–803, 2020.
- [14] M. Haridass and R. Arumuganathan. Analysis of a single server batch arrival retrial queueing system with modified vacations and N-policy. *RAIRO-Operations Research*, 49(2):279–296, 2015.
- [15] M. Jain and A. Bhagat. Embedded Markov chain approach to retrial queue with vacation, phase repair and multioptional services. *Opsearch*, 52:782–809, 2015.
- [16] K. P. Jose and P. Beena. On a retrial production inventory system with vacation and multiple servers. *International Journal of Applied and Computational Mathematics*, 6(4):108, 2020.
- [17] J. Keilson and L. D. Servi. Oscillating random walk models for $GI/G/1$ vacation systems with Bernoulli schedules. *Journal of Applied Probability*, 23(3):790–802, 1986.
- [18] J. Kim and B. Kim. A survey of retrial queueing systems. *Annals of Operations Research*, 247:3–36, 2016.
- [19] K. B. Krishna, D. Arivudainambi, and A. Vijayakumar. The $M/G/1$ retrial queue with Bernoulli schedules and general retrial times. *Computers and Mathematics with Applications*, 43(1-2):15–30, 2002.
- [20] M. N. G. Micheal and K. Indhira. A literature review on retrial queueing system with Bernoulli vacation. *Yugoslav Journal of Operations Research*, 34, 2024.
- [21] T. Phung-Duc. Retrial queueing models: A survey on theory and applications. *arXiv*

preprint arXiv:1906.09560, 2019.

- [22] R. Tian, X. Wu, L. He, and Y. Han. Strategic analysis of retrial queues with setup times, breakdown and repairs. *Discrete Dynamics in Nature and Society*, 2023(1):4930414, 2023.
- [23] X. Wu, P. Brill, M. Hlynka, and J. Wang. An M/G/1 retrial queue with balking and retrials during service. *International Journal of Operational Research*, 1(1-2):30–51, 2005.
- [24] T. Yang and J. G. C. Templeton. A survey on retrial queues. *Queueing Systems*, 2:201–233, 1987.
- [25] D. Zirem, M. Boualem, K. Adel-Aissanou, and D. Aïssani. Analysis of a single server batch arrival unreliable queue with balking and general retrial time. *Quality Technology & Quantitative Management*, 16(6):672–695, 2019.